International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

A Systematic Review of Deep Learning Approaches for Information Disorder Detection on Social Media: The Context of Misinformation, Disinformation, and Mal-information

[1]Andualem Woldegiorgis, [2]Mohammed Abebe, [3]Durga Prasad Sharma, [4]Worku Jimma

[1,2]Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia
[3]Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia; Expat Professor, Arba Minch University, Ethiopia
[4]Faculty of Computing and Informatics, Jimma University, Ethiopia

[1]andualemwoldegiorgis@bongau.edu.et, ORCID iD: 0000-0001-9851-228X
[2]mohammed.abebe@amu.edu.et, ORCID iD: 0000-0003-0622-4841
[3]dp.shiv08@gmail.com, ORCID iD: 0000-0002-1654-901X
[4]worku.jimma@ju.edu.et, ORCID iD: 0000-0001-7330-4054

**Corresponding author: Andualem Woldegiorgis*

Abstract: Social media platforms have become powerful tools for communication, information exchange, networking, and community building. However, their widespread use has facilitated the proliferation of misinformation, disinformation, and mal-information, posing significant challenges to societal harmony and individual well-being. Attacks such as social pressure, bullying, extortion, manipulation, abuse, prejudice, and violence can deepen divisions and conflicts, demonize minorities, and undermine the integrity of public opinion and democratic elections. This study presents a systematic review with quantitative comparative analysis of machine learning and deep learning approaches for information disorder detection, with a unique emphasis on low-resource multilingual contexts, particularly Amharic and Afaan Oromo, while addressing critical challenges related to multimodality, data scarcity, and contextual linguistic complexity. Using the PRISMA framework, 963 studies were screened, resulting in the selection of 99 high-quality articles for in-depth analysis. This study provides a dataset-level synthesis, comparing benchmark and low-resource datasets across multiple dimensions such as data modalities (text, image, memes, audio) and learning models. The findings indicate an increasing adoption of transformer-based and deep learning models, which demonstrated strong performance, particularly on large benchmark and multimodal datasets. Importantly, the review identifies critical research gaps, including limited multimodal fusion techniques, insufficient multi-class classification approaches, and a lack of localized context-aware models for low-resource languages. Unlike global surveys, this study provides a localized Ethiopian-context analysis that addresses linguistic diversity, resource constraints, and real-world deployment challenges. The study contributes an integrated perspective encompassing multilingual, multimodal, and multiclass dimensions and outlines future directions for developing scalable, context-aware, and inclusive detection models.

Keywords: Social media, Deep learning, Machine learning, Disinformation, Misinformation, Mal-information, Information Disorder Detection.

1 INTRODUCTION

Nowadays, the rapid expansion of social media and the World Wide Web (WWW) has substantially accelerated the spread of false, misleading, and malicious information, which has had an extensive impact on modern societies. Information disorder promotes prejudice and violence while intensifying divisions, conflicts, discrimination, war, and civil unrest, demonizing minorities, threatening peaceful coexistence, and compromising democratic elections [1], [2]. Furthermore, it intentionally promotes political affective polarization, which represents a society's tendency toward strong opposition across social, political, and ideological issues [3]. The emergence of artificial intelligence technology continues to advance, and it provides a potentially useful tool for recognizing, evaluating, stopping, and detecting the spread of information disorder [4]. Artificial intelligence describes systems capable of performing tasks such as learning and decision-making that are associated with human intelligence [5]. Deep learning, a branch of machine learning that focuses on developing models that mimic human information processing, is a subset of artificial intelligence. Natural language processing (e.g., voice translation, machine translation, and semantic understanding), speech recognition, medical applications, computer vision (e.g., object detection, tracking, and measurement), graphics, and intelligent transportation systems are among the fields in which deep learning models have been applied.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

This is achieved through multilayered neural networks that identify patterns and support data-driven decision-making [6], [7]. Additionally, deep learning can be applied to social media platforms for sentiment analysis, image and video recognition, and personalized content recommendation. It is also essential for identifying information disorders and helping filter harmful content. Currently, social media platforms play an important role in daily life by connecting millions of people worldwide for communication and information sharing across various domains, including education, democracy, healthcare, and the medical sector [8]. Additionally, social media platforms serve as an essential means of communication for reporting incidents related to community safety and various crises, such as criminal activity, violence, conflict, civil unrest, and other social emergencies. They provide updates on healthcare emergencies, accidents, natural disasters, displacement, and other crises resulting from factors such as ethnic and religious extremism, political differences, sexual assault, and land disputes [9].

However, the growing use of social media platforms has negatively affected societal cohesion and individual well-being by facilitating the spread of disruptive information content, including hate speech, offensive speech, harassment, propaganda, clickbait, fake news, manipulation of public opinion, and misleading content. Such content leads to confusion, mistrust, and the circulation of inaccurate information, often driven by financial or political interests and, in some cases, linked to suicidal ideation [10]. The remainder of this paper is organized as follows: Section 3 presents the review methodology, Section 4 discusses the results and analysis, Section 5 summarizes comparative findings, and the final sections present future research directions and conclusions.

2 Literature Survey

Disinformation is the deliberate dissemination of false information to achieve particular goals, including fake news [11], intent-based propaganda, conspiracy, hoaxes, clickbait, fabricated content, and deepfakes [12]. In contrast, misinformation refers to the unintentional dissemination of inaccurate information, including rumors, false information, and urban legends. Mal-information refers to the dissemination of harmful truths or falsehoods intended to harm individuals, social groups, organizations, or countries, such as hate speech and harassment [13]-[17].

Table 1. Conceptual Comparison of Misinformation, Disinformation, and Mal-information

Aspect	Misinformation	Disinformation	Mal-information
Definition	False or inaccurate information shared without intent to harm	Deliberately false or manipulated information shared with the intent to deceive	Information (true or false) shared with the intent to cause harm
Intent	Unintentional	Intentional (deceptive)	Intentional (harmful)
Truth Value	False or misleading	False or fabricated	Can be true, partially true, or false
Purpose	Lack of awareness, misunderstanding	Manipulation, propaganda, influencing opinions	Causing harm, harassment, and reputational damage
Typical Forms	Rumors, false information, urban legends	Fake news, propaganda, conspiracy theories, deepfakes	Hate speech, harassment, doxing, offensive content
Source Behavior	Ignorant or misinformed users	Coordinated actors, political groups, bots	Malicious individuals or groups
Example Scenario	Sharing an unverified rumor, believing it is true	Creating fake news to influence elections	Sharing private information to harm an individual
Impact	misinformation spread	Public manipulation, societal division	Psychological harm, social conflict, violence
References	[18]-[20]	[11], [21], [22]	[13]-[16],[23], [24]

As illustrated in Table 1, the comparative analytical framework distinguishes forms of information disorder based on intent, truth value, purpose, behavioral sources, and social impact. Fig. 1 illustrates the relationship among these forms of information disorder (misinformation, disinformation, and mal-information). One form of information disorder is hate speech, which refers to communication in writing, speech, or symbolic form that promotes violence, discrimination, or hatred against an individual or group based on inherent characteristics such as race, ethnicity, religion, gender, sexual orientation, or other protected attributes, thereby fostering an environment of division and intolerance [25]. As demonstrated by the variety of definitions offered by numerous scholars in the field, hate speech still has no universally recognized definition, and this ambiguity is reflected in the academic discourse [1], [26]-[29]. According to the United Nations Strategy and Plan of Action, hate speech refers to any spoken, written, or behavioral communication that targets or discriminates against an individual or group based on identity, including factors such as religion, ethnicity, nationality, race, color, descent, gender, or other characteristics, according to the United Nations' strategy and plan of action [30].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

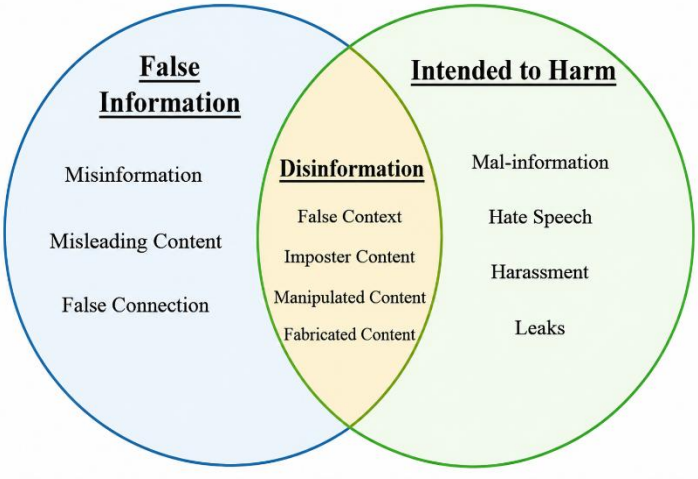

Fig. 1. Concepts of misinformation, disinformation, and mal-information [17]

Ethiopia's rich cultural and ethnic diversity is evident in its multilingual environment, featuring over 80 languages. Amharic, a Semitic language that uses the Ge'ez script, serves as the federal working language and employs a Fidel system comprising 34 base characters and six supplementary ones [31]-[33]. Despite progress in machine learning applications for text analysis and hate speech detection, Amharic remains under-resourced and lacks essential processing tools. Challenges like character redundancy and similar-sounding words complicate text classification efforts. Many studies rely on limited textual data, highlighting the necessity for publicly available datasets, stop-word lists, and annotated multimodal data, including audio, video, and images [33]-[37]. Addressing these gaps is essential, as hate speech can fuel intergroup conflict, exacerbate social divides, lead to mass violence and even genocide, and foster opposition [38].

Similarly, the propagation of disinformation, including fake news, clickbait, propaganda, hoax, and deep fake, presents serious challenges and has become increasingly prevalent in Ethiopia's media landscape, particularly across online social media platforms. Combating mal-information, misinformation, and disinformation is essential for safeguarding the well-being of citizens, protecting lives, and ensuring the stability of the country [39], [40]. Easy access to social media platforms may contribute to excessive use and its associated consequences, including social pressure, bullying, extortion, manipulation, abuse, suicide advocacy, and offensive messaging. Misinformation and disinformation emerged among the most critical global risks in 2024 and ranked among the highest in severity according to the World Economic Forum [41]. Additionally, these information disorders are projected to remain among the top global risks over the next decade, reflecting their persistent impact on societies through prejudice and violence, intensified divisions and conflicts, marginalization of minorities, and threats to electoral integrity [2].

These challenges have been amplified by emerging online social interaction platforms and have become a critical societal concern. A major challenge lies in the rapid evolution and dynamic nature of online language, making manual content moderation increasingly insufficient. Social media companies have faced increasing pressure to address this issue by investing hundreds of millions of euros annually and continue to face criticism for insufficient action [42]. One challenge in addressing these issues is the reliance on traditional methods, such as manual content review, to detect and remove hate speech and offensive material. These approaches are labor-intensive, time-consuming, and ultimately impractical for large-scale deployment [42]. Numerous researchers have explored semantic content analysis techniques based on natural language processing (NLP) and machine learning (ML) to address the growing need for scalable and automated hate speech detection approaches [43]-[45]. However, numerous challenges remain, including resource constraints, limited advanced technology for automatic detection, and limited public dataset availability. Because of these challenges, the Ethiopian language ecosystem lacks sophisticated tools for identifying false, misleading, and harmful content due to financial and technological limitations. Furthermore, various multimodal components, such as text, images, memes, video, and audio, together with proper annotations, were not sufficiently addressed in previous studies.

Additionally, there is growing concern regarding developments in Generative AI, which have increased the potential for creating hyper-realistic deepfake images and videos that may seriously harm individuals, vulnerable populations, and society through the spread of misinformation, disinformation, and propaganda [46]. Deepfake technology is a global concern; however, there remains limited research and a lack of publicly available datasets specifically addressing the Ethiopian context. This gap limits the ability to understand and address the potential impacts of deepfakes within Ethiopia's social, cultural, and political context. Despite the rapid advancement of machine learning and deep learning techniques for detecting misinformation, disinformation, and mal-information on social media, several critical gaps remain in the existing literature.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

First, most existing studies focus primarily on binary classification and unimodal data (mainly textual), with limited exploration of multimodal data integration and fusion methods involving images, memes, videos, and audio. Second, there is a significant lack of large-scale, publicly available, and well-annotated datasets, particularly for low-resource languages such as Amharic and Afaan Oromo. Third, existing approaches often lack context-aware and linguistically adaptive capabilities, making them less effective in capturing cultural, semantic, and multilingual distinctions. In addition, existing systematic reviews largely provide generalized insights without emphasizing the Ethiopian context or the challenges associated with multimodal, multiclass, and multilingual classification tasks. The main objective of this study is to conduct a comprehensive systematic literature review with quantitative comparative analysis of social media information disorders, with a focus on the Ethiopian context. Therefore, this study addresses the following research questions.

The selected studies were organized to address the following three research questions.

RQ1: What are the limitations of existing state-of-the-art approaches for detecting misinformation, disinformation, and mal-information?

RQ2: What multimodal and multi-class deep learning models have demonstrated strong performance in detecting disinformation, misinformation, and mal-information?

RQ3: What are the contextual research gaps for linguistic localization of scalable social media datasets, and deep learning model accuracy for handling futuristic multimodal, multilingual, and multi-class classification tasks?

This study makes the following key contributions:

1. Comprehensive Systematic Review: The study presents a large-scale systematic review of 99 selected studies (from an initial pool of 963 records) using the PRISMA framework to ensure methodological rigor and transparency.
2. Quantitative Comparative Analysis: The study provides a comparative analysis of machine learning, deep learning, and transformer-based models across multiple datasets, modalities, and evaluation metrics.
3. Multimodal and Multiclass Perspective: The study synthesizes existing approaches by emphasizing multimodal (text, image, memes, and audio) and multi-class classification frameworks beyond traditional binary classification.
4. The study identifies context-specific challenges in Ethiopian languages and highlights gaps in linguistic resources, datasets, and model adaptability.
5. The study identifies future research opportunities, including scalable and well-annotated dataset development, multimodal fusion techniques, and context-aware deep learning models for low-resource and multilingual settings.

This study differs from existing systematic reviews in several important ways:

1. Compared to existing reviews, this study provides a context-aware analysis tailored to low-resource and multilingual environments, particularly focusing on Ethiopia.
2. It integrates multimodal, multilingual, and multi-class perspectives into a unified review perspective, which remains largely underexplored in previous studies.
3. The study extends descriptive review by incorporating quantitative analysis of datasets, models, and performance metrics, enabling deeper insight into model effectiveness.
4. The study also discusses emerging challenges related to transformer models, large language models (LLMs), and multimodal AI, which have received limited coverage in earlier surveys.
5. Finally, it connects global research trends and localized implementation challenges, offering insights for practical deployment.

This study discusses practical deployment considerations, policy relevance, and real-world implications. Deep learning-based information disorder detection models may support Ethiopia's government information systems, fact-checking organizations, and social media monitoring platforms in detecting and reducing harmful content such as hate speech and fake news. From a policy standpoint, the results encourage the development of frameworks for digital governance, language-inclusive AI policies for Afaan Oromo, Amharic, and other languages, and strategies to prevent conflicts and ensure election integrity. In real-world applications, implementing such systems requires scalable, multilingual, and multimodal models while considering infrastructure limitations, explainability, and ethical use.

3 Review Methodology and Structured Protocol

This study conducted an extensive literature review on information disorder detection, focusing on the impact of social media influence on peaceful coexistence among people from diverse regions and religions using recent literature. Structured inclusion and exclusion criteria were applied using the PRISMA framework. Overall, 963 records were retrieved from peer-reviewed journals and proceedings relevant to the research scope. A total of 99 studies were thoroughly analyzed to ensure comprehensive coverage and reliability of the findings.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

3.1. A Detailed Review Architecture and PRISMA Framework

As illustrated in Fig. 2, the workflow adopted in this study begins with defining key concepts and research questions through literature review. This is followed by a comprehensive survey of related work, structured screening, and comparative evaluation of datasets and techniques using machine learning, deep learning, and hybrid models across different data modalities, including textual, visual, and audio data. Performance evaluation metrics were incorporated to systematically assess model performance across the respective datasets. Finally, the study identifies research gaps, including limited multilingual and multimodal coverage, and recommends future directions in linguistic localization and context-aware deep learning for scalable and inclusive detection models.

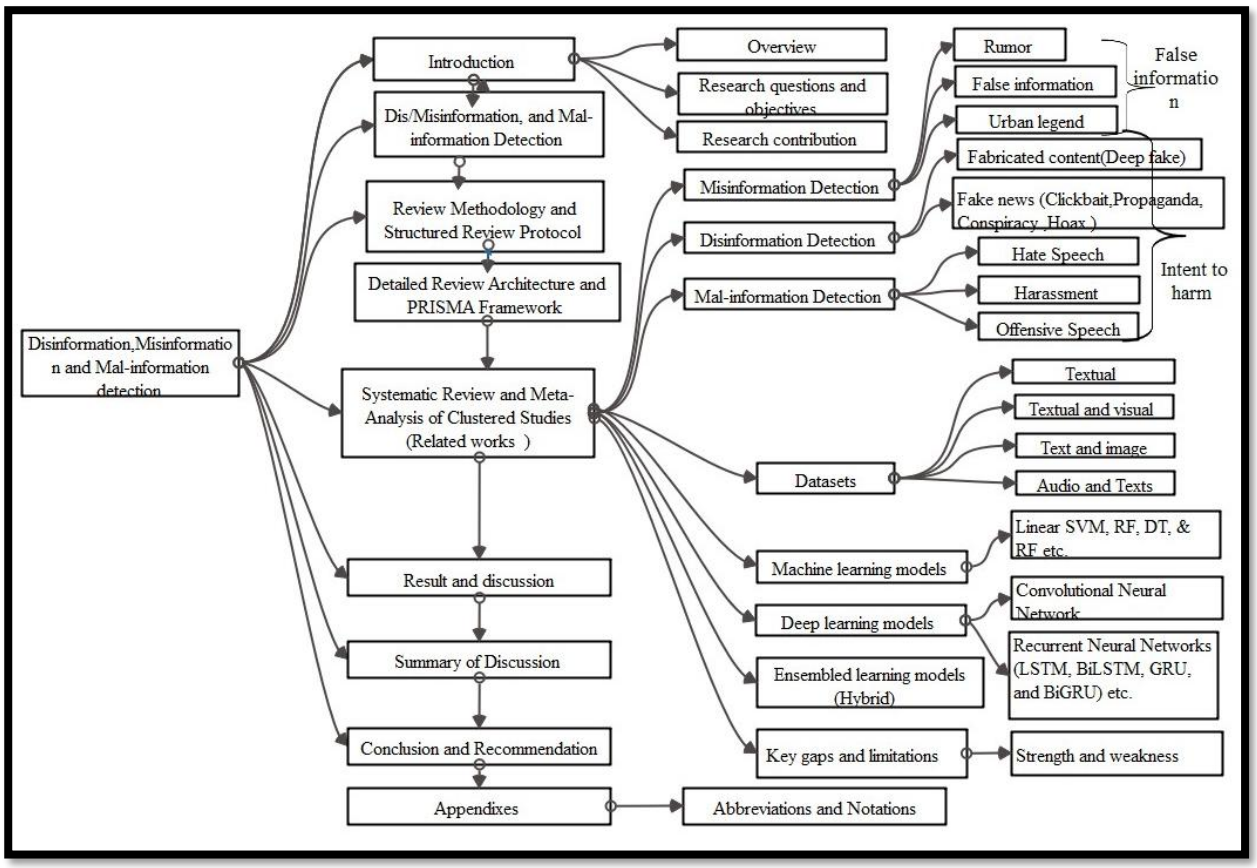

Fig. 2. Architectural Framework of the Systematic Review and Analysis Process

3.2. PRISMA Design and Search Strategy

In this review study, the reporting strategy followed the PRISMA 2020 statement guidelines for systematic reviews [47],[48]. The review process followed the PRISMA 2020 27-item checklist, which covers key aspects such as literature search strategy, study selection, data extraction, synthesis of findings, and assessment of bias. A PRISMA 2020 flow diagram documented the study selection process, including the number of records identified, screened, excluded, and finally included in the analysis. In this study, a structured multi-stage screening process in accordance with PRISMA 2020 guidelines as shown in Fig. 3.

First, during the identification stage, a total of 963 records were retrieved from major academic databases, including Scopus, Springer, Elsevier, IEEE, MDPI, Wiley Online, and others.

- Deduplication stage: Duplicate records (n = 437) were identified and removed through manual screening.
- Title and abstract screening: The remaining 526 records were screened based on titles and abstracts to assess thematic relevance, and 308 records were excluded due to lack of relevance to misinformation, disinformation, and mal-information detection.
- Methodological screening: The remaining 218 studies were evaluated based on methodological criteria, including the use of machine learning and deep learning techniques, and 75 studies were excluded due to the absence of these methods.
- Full-text eligibility assessment: A total of 143 full-text articles were assessed for eligibility, and 44 were excluded due to insufficient relevance to the research scope and lack of empirical validation.
- Final inclusion: A total of 99 studies were included in the systematic review with quantitative comparative analysis.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

To ensure consistency, reproducibility, and data quality, dataset selection in this study followed a structured filtering process. Studies were included if they utilized publicly available datasets related to social media–based information disorder detection, encompassing textual, visual, or multimodal data. Additionally, the datasets were required to support empirical evaluation using standardized performance metrics, with the language scope restricted to English, Amharic, and Afaan Oromo. Conversely, studies were excluded if they lacked dataset transparency, reproducibility, omitted evaluation metrics such as accuracy, precision, recall, and F1-score, or were not based on social media data. Accordingly, priority was given to benchmark datasets such as ISOT, WELFake, and FakeNewsNet, while also allowing limited but critical inclusion of low-resource datasets, including ETH_FAKE and Amharic hate speech datasets. This filtering process ensured that only empirically validated and relevant datasets were included in the analysis.

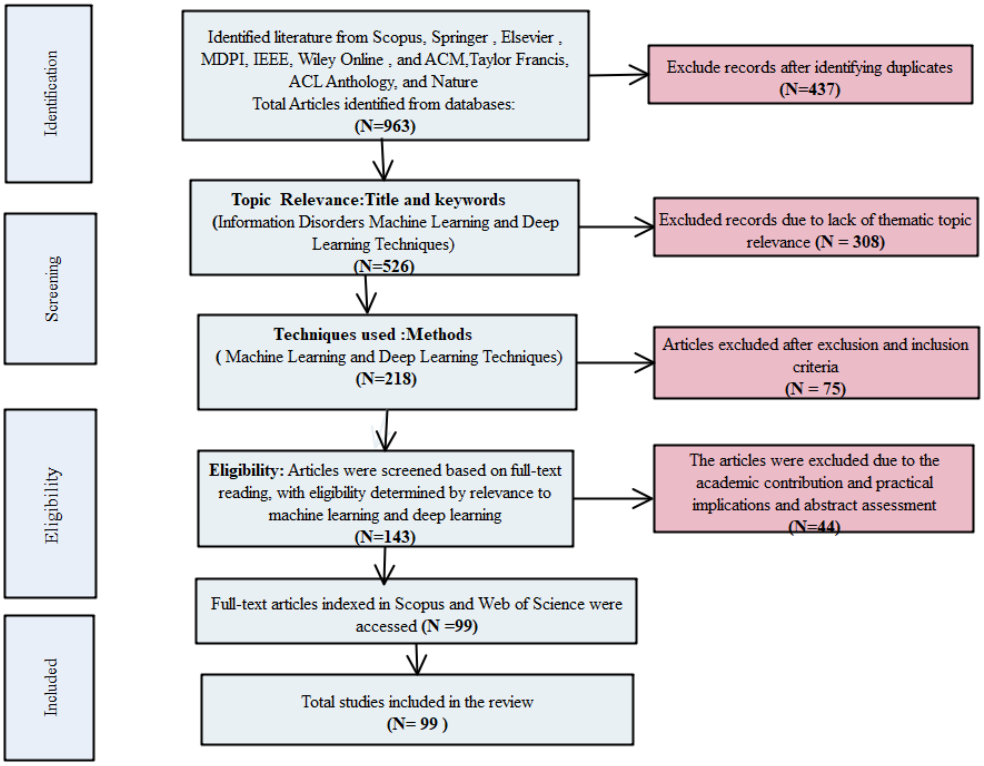

Fig. 3. PRISMA 2020-based Study Selection and Screening Process

To minimize potential bias throughout the review process, a structured bias control strategy was applied for reliability through the inclusion of peer-reviewed studies from diverse sources, including high-impact journals and preprints with clear methods and standard evaluation metrics; dataset selection based on publicly available, well-documented benchmark datasets and low-resource datasets, and consistency was ensured via PRISMA-based multistage screening with inclusion and exclusion criteria. To improve transparency and align with PRISMA 2020, the exclusion process is explicitly summarized in Table 2.

Table 2. Exclusion Justification Counts across Screening Stages in Accordance with PRISMA 2020 Standards

Screening Stage	Records Excluded	Justification
Duplicate Removal	437	Removing redundant entries across databases
Title/Abstract Screening	308	Irrelevant to information disorder or social media context
Methodological Screening	75	No Machine Learning/Deep Learning techniques used
Full-Text Eligibility	44	Lack of empirical validation and insufficient methodological clarity
Total excluded	864	—
Final included Studies	99	Eligible for systematic review

As illustrated in Table 2, the exclusion process provides a detailed account of the number of records excluded at each stage of the screening process, including duplicate removal, title/abstract screening, methodological filtering, and full-text eligibility assessment. Each exclusion stage is accompanied by explicit justification counts, ensuring transparency in study selection and demonstrating how the initial pool of records was systematically refined to the final set of included studies (n = 99).

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

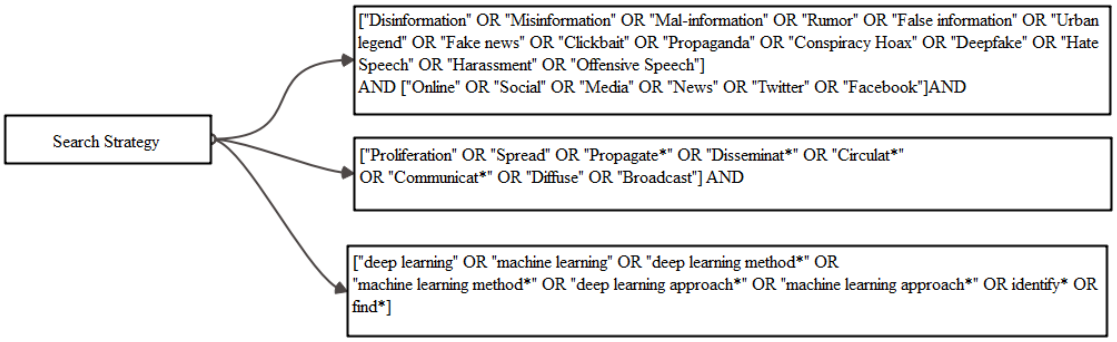

Fig. 4. Structured search strategy for literature retrieval

Fig. 4 illustrates the logical combination of keywords and Boolean operators used to retrieve relevant studies across multiple academic databases. The strategy integrates two key dimensions: (1) type of information disorder, i.e., misinformation, disinformation, and mal-information, (2), social media context, and machine learning and deep learning models. This structured search strategy ensures comprehensive coverage of the research domain.

3.3. Inclusion and Exclusion Criteria

As illustrated in Table 3, the inclusion criteria for this review were designed to ensure that the selected studies were relevant, reproducible, and published in high-quality journals. The following conditions were applied: (1) articles must be published in peer-reviewed journals; (2) each study must focus on applying machine learning or deep learning models to detect information disorder on social media platforms; and (3) articles must be written in English. The scope of the review was limited to studies published between 2018 and 2025. All selected studies had to be fully accessible at the time of review to enable detailed examination of methodologies, findings, and citation impact. To ensure empirical consistency, each study was required to utilize at least one publicly available dataset as part of its experimental framework. Only datasets in English, Amharic, and Afaan Oromo were included to maintain consistency across the reviewed works. These inclusion and exclusion criteria primarily favored studies indexed in major academic databases, including Web of Science and Scopus.

Table 3. Inclusion and Exclusion Criteria Applied for Study Selection

S. No.	Selection Criterion	Inclusion criteria	Exclusion criteria
1.	Topic Relevance	Studies focused on detecting information disorders (misinformation, disinformation, or mal-information) on social media using Machine learning and deep learning techniques.	Articles not addressing information disorders specifically, or those unrelated to social media.
2.	Techniques used	Utilizes machine learning (Linear SVM, RF, DT) and deep learning models (e.g., CNN, RNN, LSTM, Transformer, BERT).	Studies employing natural language processing within statistical methods were excluded.
3.	Publication Type	Peer-reviewed journal articles, papers, and high-quality preprints (e.g., arXiv with citations).	Non-peer-reviewed blogs and posters were excluded.
4.	Publication date	Articles published between January 2018 and 2025	Articles published earlier than 2018 were excluded.
5.	Study Design	Empirical studies encompass experimental research, benchmark evaluations, and comparative analyses of model performance.	Review articles, editorials, opinion pieces, letters, or purely theoretical papers without empirical validation were excluded.

The study follows PRISMA 2020 guidelines by explicitly reporting identification, screening, eligibility, and inclusion stages, along with transparent exclusion justifications and dataset selection criteria to ensure reproducibility and methodological thoroughness.

4 Results and Discussion

The review followed PRISMA guidelines to ensure structured reporting and transparency, and the synthesized findings were used to identify key gaps and formulate directions for future research [47], [48].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

As depicted in Fig. 3, a total of 963 records were identified from major databases, including Scopus, Springer, Elsevier, MDPI, IEEE, Wiley Online, ScienceDirect, Google Scholar, and others such as Taylor & Francis, ACL Anthology, and Nature. The retrieved studies were screened through title and abstract assessment [49], [50], and duplicate records were removed according to the exclusion criteria, as shown in Table 2. Subsequently, a total of 143 studies met the inclusion criteria. Following detailed eligibility assessment criteria based on [51], 99 studies were selected for final review and analysis, as illustrated in Fig. 3. The study selection process and source distribution are summarized in Tables 4 and 5. No studies were selected from AfricArXiv, bioRxiv, or medRxiv. As presented in Fig. 5 and Fig. 6, the analysis illustrates the annual distribution of publications and their corresponding citation counts. The analysis indicates that publication volume increased substantially in 2024 and 2025, while studies published in 2021 and 2024 received comparatively higher citation counts, highlighting their influence in the field.

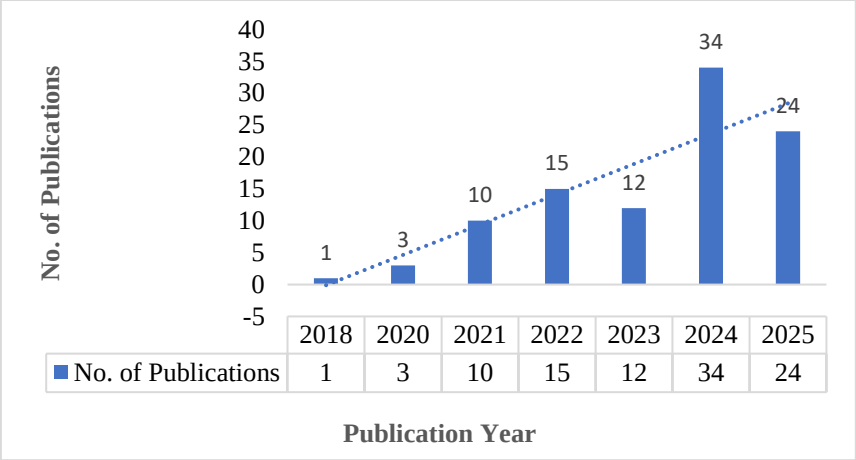

Fig. 5. Temporal Distribution of Publications on Information Disorder Detection

As shown in Fig. 5, annual publication trends reveal a substantial increase in studies after 2021, with a peak in 2024 and 2025. This indicates a growing academic and practical interest in combating information disorder and presenting the rapid evolution of deep learning approaches in this domain.

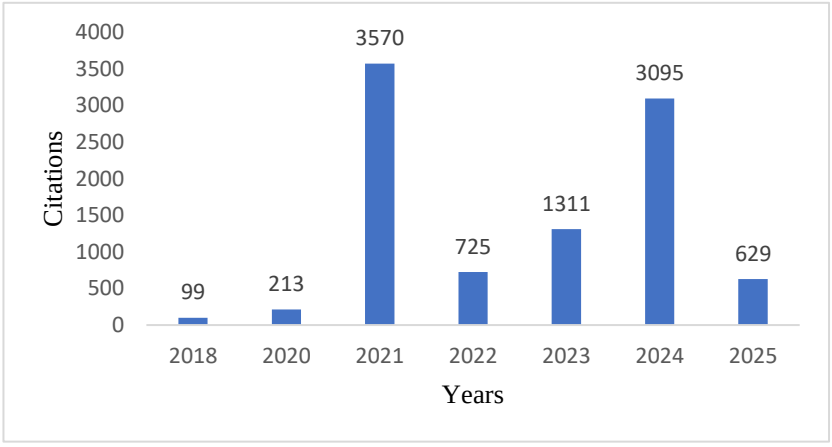

Fig. 6. Citation Trend Analysis of Selected Studies

Fig. 6 presents the citation patterns across publication years, demonstrating that studies published in recent years, particularly in 2021 and 2024, have achieved comparatively higher citation impact. Table 4 summarizes the PRISMA-guided the study selection process. Table 5 presents the source-wise distribution of the 99 studies included in the systematic review, presenting the contribution of major academic databases and publishers. The results indicate that a substantial proportion of studies originated from leading publishers such as Springer (32) and Elsevier (24), reflecting their strong representation in research on information disorder and deep learning. Contributions from IEEE and MDPI further reflect the technological and computational focus of the field, while smaller counts from sources such as ACL Anthology and ACM Digital Library indicate emerging work in natural language processing and computational linguistics. The distribution demonstrates broad coverage of a technology-driven research landscape, ensuring comprehensive coverage of high-quality, peer-reviewed literature.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 4. Summary of the PRISMA-guided Study Selection Process

Processes	Count	Academic databases
Records identified (all databases)	963	Scopus, Springer, Elsevier, MDPI, IEEE, Wiley Online, and ACM, Taylor Francis, ACL Anthology, and Nature
Duplicates removed	437	Duplicate records were identified and removed using deduplication procedures
Records excluded due to lack of thematic relevance	308	Studies lacking thematic relevance to the research focus were excluded.
Records screened based on techniques and methods used	75	Studies lacking machine learning or deep learning techniques were excluded during screening
Full-text articles were assessed for eligibility based on relevance to machine learning and deep learning approaches	44	Articles lacking sufficient academic contribution, methodological clarity, practical relevance, or abstract-level relevance were excluded.
Final included studies were synthesized after eligibility assessment and included studies indexed in major academic databases	99	Eligible studies were peer-reviewed and investigated Amharic misinformation, disinformation, and mal-information detection using machine learning and deep learning approaches, with relevance to low-resource language research

Table 5. Summary of Source Distribution of Included Studies Across Major Academic Databases and Publishers

Databases category	Count
Scopus	6
SPRINGER	32
Elsevier	24
Taylor & Francis	3
MDPI	11
IEEE	11
ACL Anthology	5
ACM Digital Library	2
Wiley Online	3
Research square	1
Other	1
Total	**99**

4.1. Performance Evaluation Metrics

Model evaluation was conducted using the confusion matrix, precision, accuracy, recall, and F1-score as performance metrics [52], [53]. Models were evaluated using standard performance metrics, including confusion matrix, accuracy, recall, precision, and F1-score, calculated using Equations (1)–(4). A concise description of the associated metrics is provided below. Confusion matrix is used for visualizing the performance of the prediction model. Accuracy is the percentage of examples that are correctly classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where, TP is True positive; TN is True negative; FP is False positive, and FN is False negative. Recall is the percentage of examples that are correctly classified.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-score is a single metric that combines recall and precision, providing a balance between the two. It is the harmonic mean of precision and recall.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Precision is the percentage of predicted positive cases that were correctly classified.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Finally, comparative model evaluation was conducted to identify limitations and challenges in detecting and classifying misinformation, disinformation, and mal-information on social media. The following subsections present the analysis and address the stated research questions.

5 Systematic Literature Review with Quantitative Comparative Analysis of Clustered Studies

RQ1: What are the limitations of existing state-of-the-art approaches for detecting misinformation, disinformation, and mal-information?

As shown in Table 6 and Fig. 7, recent studies on disinformation, misinformation, and mal-information detection using various datasets and learning models were compiled, critically reviewed, and evaluated based on their contributions, gaps, limitations, and research outcomes. There is notable variation in the classification of misinformation, disinformation, and mal-information, particularly regarding the categorization of "fake news". Some studies classify fake news as a type of misinformation [52]-[57], [14], others classify it as disinformation [58], [21], [59], some classify it as both misinformation and disinformation [60]-[63] while others distinguish fake news from both categories [64]. This study adopts a classification of information disorder based on intent and content type.

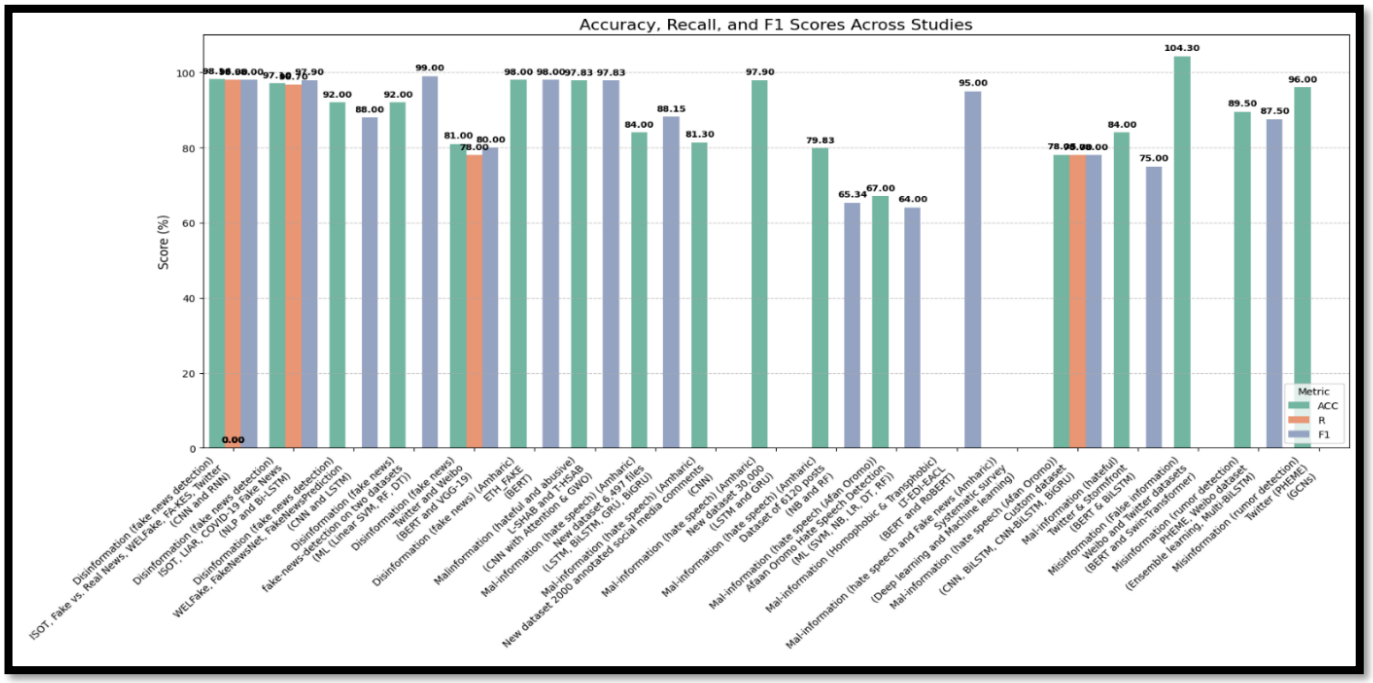

Fig. 7. Summary of the existing state of the art related works (problem addressed, dataset, and techniques used)

The first classification is mal-information, which refers to genuine or manipulated information shared with the intention to cause harm (hate speech, harassment, and offensive speech) [23], [65], [13]. The second classification is disinformation, which refers to deliberately false or manipulated content shared with the explicit intent to mislead and cause harm (fake news [21],[59], clickbait [66], propaganda [67], deep fake [68]-[70]). The third classification is misinformation [71], referring to false or misleading content shared without intent to deceive, such as rumor, false information, and urban legend [67], [72], [14], [73]-[78]. As presented in Fig. 7, the summary of existing state-of-the-art studies and the quantitative comparative analysis indicate a methodological shift from machine learning toward deep learning and transformer-based approaches. Machine learning models such as NB, SVM, and RF achieved moderate performance (65–92%) on small or custom datasets using unimodal data but generally showed limited capability in learning multimodal representations. In contrast, deep learning models, including CNNs, LSTMs, and BiLSTMs, demonstrated stronger performance on benchmark datasets such as ISOT, WELFake, and FakeNewsNet, often achieving accuracy above 90% and benefiting from both unimodal and multimodal representations. Table 6 presents the parametric summary of existing state-of-the-art methods.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 6. Comparative Summary of Existing State-of-the-Art Studies with Remarks on Key Gaps and Limitations

Ref.	Year	Problem addressed	Dataset used	Data types	Techniques used	Evaluation metrics			Key contributions	Research limitations
						ACC	R	F1		
[73]	2025	Disinformation (fake news detection)	ISOT Fake vs. Real News WELFake FA-KES Twitter	Textual and visuals	CNN and RNN	0.98	0.98	0.98	Proposed multi-level fusion-based framework (CDLR). Which integrates CNN and RNN	The study requires effective integration of multimodal data, a more thorough investigation of fusion techniques, and greater exploration of underperforming datasets.
[74]	2025	Disinformation (fake news detection)	ISOT, LIAR, and COVID-19 Fake News	Textual and visuals	NLP and Bi-LSTM	0.97	0.96	0.98	The study presents a novel framework for fake news detection called Multi-Modal Fake News Detection (MM-FND)	The study is overly reliant on a specific dataset, which limits the adaptability of the proposed methods to emerging misinformation techniques in multimodal contexts
[75]	2024	Disinformation (fake news detection)	WELFake, FakeNewsNet, & FakeNewsPrediction	Textual	CNN and LSTM	0.92	-	0.88	Proposed a hybrid approach by combining CNN and LSTM for fake news detection	The study lacks multi-modal integration, lacks extraction of long article sentences, and lacks checks with fact-checks
[76]	2024	Disinformation (fake news)	fake-news-detection on two data sets	Textual	ML (Linear SVM, DT, and RF)	0.92	-	0.99	The study reveals that various ML (SVM), (GB), (RF), and (DT) offer promising solutions for fake news detection.	The study focuses on ML and does not show the possibility of using DL to identify textual fake news.
[77]	2022	Disinformation (fake news)	Twitter and Weibo	Textual and visual	BERT and VGG-19 model	0.81	0.78	0.80	The study introduces two adaptive Multi-modal Compact Bilinear Pooling (MCBP) modules for text and visual feature extraction in fake information detection.	Lack of integration of multiple modalities, primarily focusing on binary classification and the VGG-19 model for visual data, and other deep-learning visual feature extraction models have not been addressed.
[53]	2020	Misinformation (rumor detection)	Twitter (PHEME)	Textual and user profile	GCNs	0.25-0.30 improvement In accuracy			A study reveals the use of graph convolutional networks (GCNs) to detect fake news campaigns by classifying diffusion graphs of news on social media.	The study has a significant challenge with data imbalance, as most comments are non-toxic It identifies toxic content over 50% of the time but struggles with precision and recall

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Ref.	Year	Problem addressed	Dataset used	Data types	Techniques used	Evaluation metrics			Key contributions	Research limitations
						ACC	R	F1		
[79]	2024	Misinformation (rumor detection)	TWITTER, WEIBO (COVID-19)	Text + propagation structure	XLM-R BiGCN	0.873	-	0.86	The study proposed an adversarial contrastive learning framework design to enhance rumor detection in low-resource	Lacks the detailed analysis of how the rumor propagated in different languages and lacks the integration of multimodal
[80]	2021	Misinformation (rumor detection)	Twitter PHEME	Textual	SVM-TK, GRU-RNN, RvNN, Bi-GCN, PLAN	0.89, 0.90 (Twitter 15,16)	-		The study proposed the ClaHi-GAT model, aiming to improve over state-of-the-art baselines in rumor detection	The study lacks a detailed analysis of low-resource language analysis, limited to Twitter, PHEME data, and the English language, and limited to integrating multimodal
[81]	2022	Misinformation (False information)	Weibo and Twitter datasets	Text and image	BERT and Swin-Transformer	Accuracy improvement by 0.043 compared to existing methods			The proposed method utilizes a sophisticated fusion strategy that begins with early fusion, integrating text and image features through a deep autoencoder.	The study lacks detailed analysis results with a comparison between other existing methods
[39]	2022	Mal-information (hate speech and Fake news (Amharic)	Systematic survey	-	Deep learning and Machine learning	-			The review's findings show that combining DL with other traditional ML techniques with a well-rounded dataset can enhance detection accuracy.	The study fails to provide a comprehensive demonstration across different modalities and lacks proper evaluation. It primarily focuses on binary classification and does not adequately address multiclass classification model.
[82] **[83]**	2025 2023	Mal-information (hate speech)	GOAT-Bench:	Text and image	LMM (GPT-4V, LLaVA,and Qwen-VL)	0.72	-	0.70	The study developed a benchmark *GOAT-Bench* dataset and conducted an extensive evaluation of LLMs (GPT-4V, LLaVA, and Qwen-VL)	The study uses a small dataset and linguistic scope and fails to generalize. Lacks a detailed analysis of how the memes' information was extracted and preprocessed, focused on binary classifications, multi-class, multi-label, and multilingual were not explored.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Ref.	Year	Problem addressed	Dataset used	Data types	Techniques used	Evaluation metrics			Key contributions	Research limitations
						ACC	R	F1		
[33]	2021	Mal-information (hate speech (Afaan Oromo)	Afaan Oromo Hate Speech Detection	Textual	ML (SVM, NB, LR, DT, and RF)	0.67		0.64	The study demonstrated that ML models can effectively identify Afaan Oromo hate speech.	The study achieves lower accuracy and F1 score. Lacks analysis and detailed comparison with existing state-of-the-art studies, and lacks integration of multimodal and conducts binary classification
[84] [85]	2022 , 2025	Mal-information (hate speech (Afaan Oromo)	Custom dataset	Textual	CNN, BiLSTM, CNN-BiLSTM, and BiGRU	0.78.	0.78	0.78	The proposed model effectively detects and classifies offensive speech.	The study lacks focus on bilingual hate speech detection in Ethiopian languages, particularly Amharic and Afaan Oromo, and fails to address the complexities of code-mixing in bilingual communication.
[34]	2018	Mal-information (hate speech) (Amharic)	A dataset created by the authors, a total of 6120	Textual	Machine learning (NB, and RF)	0.79		0.65	The study uses Apache Spark and machine learning algorithms (Random Forest and Naïve Bayes) to classify posts as 'hate' or 'not hate.	Lacks analysis with other ML models and a detailed comparison with existing state-of-the-art studies Lacks integration of multimodal and conducts binary classification, and small dataset size
[35]	2020	Mal-information (hate speech) (Amharic)	A new dataset created by the authors, a total of 30,000	Textual	Deep learning (LSTM and GRU)	0.97	-	-	The study developed an LSTM-based model for detecting hate speech in Amharic Facebook posts. The GRU model, though effective, performed lower at 88%.	The study fails to address the consideration of long sentences with and without comparison with BiLSTM. There are a lack of publicly available datasets and a compression of existing deep learning models.
[37]	2022	Mal-information (hate speech) (Amharic)	A new dataset was created total of 2000 annotated social media comments	Textual	Deep learning (CNN)	F1 score of 0.81			The study proposes a multi-channel Convolutional Neural Network, which outperforms traditional single-channel CNNs in feature extraction. While a baseline single-channel CNN scored 78.2% and the (SVM) model 92.5%.	Small dataset sizes are used and fail to generalize. The study only considers the F1 score rather than other performance evaluation metrics. Lacks multimodal integration, conducts binary classification, and lacks a detailed analysis of the proposed model.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Ref.	Year	Problem addressed	Dataset used	Data types	Techniques used	Evaluation metrics			Key contributions	Research limitations
						ACC	R	F1		
[36]	2022	Mal-information (hate speech) (Amharic)	The new dataset created by the authors uses a total of 6,497 files	Audio and Text	DL (LSTM, BiLSTM, GRU, and BiGRU)	The accuracy of LSTM and BiLSTM is 0.84 and 0.8815, respectively			The study uses deep learning to detect hate speech in Amharic	Uses a small dataset size and fails to generalize. Lacks multimodal integration and only performs binary classification. The study did not address any other offensive and abusive speech
[86]	2022	Disinformation (fake news) (Amharic)	(ETH_FAKE)	Textual	BERT	0.98,	-	0.98	The study used deep learning techniques and fast Text word embedding to detect fake news in Amharic.	The study fails to check it with fact-check addresses compression, considering that other abusive information is not included. Lacks multimodal integration and only performs binary classification.
[87]	2023	Mal-information (Harmful Memes)	Harm-C (COVID-19), Harm-P (political memes), FHM (Facebook Hateful Memes)	Text and image)	LLM; cross-attention fusion; T5-encoder–decoder	0.89 (Harm-P)		0.89 (Harm-P)	The study develops a novel generative fame work (MR. HARM) for harmful meme detection, which includes two training stages: that distill reasoning knowledge from LLMs into smaller multimodal models	The study fails to address image captioning, and limited visual background knowledge leads to errors in culturally or historically rich images. It fails to consider gestures and symbols related to hate. It lacks standardized benchmarks to rigorously evaluate multimodal reasoning and explainability.
[69]	2024	Mal-information (hateful and abusive)	L-SHAB and *T-HSAB*	Textual	CNN with an attention mechanism and (GWO)	0.97		0.97	The study proposed a hybrid approach combining a CNN with an attention mechanism and an optimized RF, demonstrating strong performance.	The study fails to address the consideration of long sentences with and without comparison with BiLSTM. Lacks detailed analysis with other deep learning models Performed only on textual data

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Ref.	Year	Problem addressed	Dataset used	Data types	Techniques used	Evaluation metrics			Key contributions	Research limitations
						ACC	R	F1		
[88]	2025	Mal-information (Homophobic & Transphobic)	LT-EDI-EACL	Textual	BERT and RoBERT	-	-	0.95	The study demonstrates a transformer-based model that effectively classifies comments as homophobic, transphobic, or non-anti-LGBT+ in both English and Malayalam.	The study lacks detailed analysis results with a comparison between other existing models.
[89]	2025	Mal-information (hateful)	Twitter & Stormfron	Textual	BERT & BiLSTM	0.84		0.80, 0.75	A study addressed the detection of white supremacy in the English language, distinguishing between white supremacist and non-white supremacist content. Creation of a custom dataset	Lacks the integration of other modalities, limited scope on one language, and does not discuss the long-term effectiveness of the proposed models.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 7. Summary and Comparative Analysis of Recent Survey Studies

Ref.	Year	Problem addressed	Research type applied	Techniques used	Key contributions	Research limitations
[39]	2022	Mal-information (hate speech and Fake news (Amharic)	Systematic survey	Deep learning and Machine learning	The review's findings show that combining DL with other traditional ML techniques with a well-rounded dataset can enhance detection accuracy.	The study fails to provide a comprehensive demonstration across different modalities and lacks proper evaluation. It primarily focuses on binary classification and does not adequately address a multiclass classification model. Further, it does not incorporate and analyze various forms of abusive information, such as Misinformation, mal-information, disinformation, and fusion techniques.
[90]	2024	Misinformation	Survey	Generative AI	The study analyzes 135,838 fact checks (1995–2022) and shows that ~80% of misinformation was media-based, with video dominant, and introduces a new AMMEBA dataset.	The study lacks generalizability due to its reliance on English-language fact-checks, limiting its ability to capture misinformation trends and modalities in other languages. In addition, it lacks robustness due to attrition of harmful claims, incomplete early typologies, and limited cross-lingual coverage, presenting the need for context-aware, multimodal detection
[91]	2024	Misinformation, disinformation, and fake news	Systematic literature review	Quantitative content analysis of 1261 articles	Provides a comprehensive interdisciplinary review; identifies major research themes (detection, dissemination, impact); highlights the rapid growth of the field and key research gaps.	The study is limited to English-language; strong geographical bias; lacks real-world validation of computational detection methods, and insufficient investigation to the impact of misinformation, disinformation, and fake news on social movements and democracy, leaving a gap in understanding the broader societal consequences.
[92]	2024	Misinformation	Systematic literature review	Natural language processing and deep learning	The study findings show that misinformation detection is mainly focused on health (≈55%, especially COVID-19), with machine learning, particularly ensemble models, widely used. Social media (especially Twitter) is the primary data source (~66%), highlighting the need for reliable sources and stronger mitigation strategies.	The study lacks a detailed analysis of research trends, cross-country collaboration issues, and temporal impact. It also lacks focus on practical implementation and user education for effective misinformation verification.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Ref.	Year	Problem addressed	Research type applied	Techniques used	Key contributions	Research limitations
[93]	2024	Disinformation	Bibliometric and systematic review	Quantitative content analysis of 50 articles	The study finds that research on disinformation has expanded rapidly since 2018, largely driven by COVID-19 and vaccine-related topics. It identifies disinformation as a highly prominent area of study (99.819%) and shows that quantitative methods (64%) are used more frequently than qualitative approaches (26%).	The study fails to address emerging technologies with respect to data modalities and does not address methodological analysis.
[94]	2023	Disinformation (Fake news)	Systematic survey	Machine learning approaches	This study synthesizes definitions and theories to explain why fake news is believed and spread, reviews machine learning and deep learning methods for its detection, summarizes key characteristics and datasets, and highlights ongoing challenges such as its rapid evolution and the need for more robust, scalable solutions	The study lacks sufficient empirical analysis and provides limited discussion on real-world implementation. Lacks detailed analysis and generalization of machine learning models with complex datasets, such as multimodal data, as well as for the exploration of novel algorithms specifically tailored to FND tasks.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Furthermore, the quantitative comparative analysis shown in Fig. 7 indicates that research on disinformation, mal-information, and misinformation, particularly in low-resource languages such as Amharic and Afaan Oromo, has employed both machine learning and deep learning techniques. Deep learning models (such as LSTM, CNN, and BERT) generally achieved higher performance than traditional machine learning methods in several reviewed studies, with some reporting accuracy above 97%. Overall, deep learning and transformer-based architectures demonstrated strong performance in complex detection tasks under both unimodal and multimodal settings, particularly when evaluated using domain-specific datasets and advanced optimization techniques. As illustrated in Table 7, the comparative analysis of recent survey studies indicates that this study provides a more context-aware and comprehensive review of misinformation, disinformation, and mal-information and their related concepts. While existing studies were largely English-centric and focused on unimodal binary classification, this study focuses on low-resource languages such as Amharic and Afaan Oromo and explicitly addresses challenges such as data scarcity and linguistic complexity. In addition, this study emphasizes multimodal (text, image, and memes) and multiclass classification settings. Methodologically, this study adopts a PRISMA 2020-based study selection and screening process. Additionally, this manuscript integrates transformers, LLMs, and multimodal data and links them to practical and localized deployment. Overall, this study contributes to ongoing research in multilingual, multimodal, and context-aware information disorder detection.

RQ2: What multimodal and multi-class deep learning models have demonstrated strong performance in detecting disinformation, misinformation, and mal-information?

As shown in Tables 8 and 9, a comparative analysis was conducted on recent deep learning and machine learning models for information disorder detection across various datasets. Table 9 shows that hybrid deep learning and transformer-based models frequently achieved strong performance in information disorder detection, with reported accuracies and F1-scores reaching approximately 97–98% on textual and multimodal datasets. Larger datasets such as ISOT and WELFake generally produced more stable and reliable evaluation outcomes, while smaller datasets like FA-KES may produce unstable results. In addition, well-annotated datasets support more consistent labeling, whereas custom datasets may introduce annotation bias. Binary classification tasks were generally simpler and often achieved higher accuracy than multiclass tasks. Differences in evaluation metrics and validation methods also limit comparability. Therefore, reported performance may reflect dataset characteristics rather than inherent model superiority, and fair comparison requires the same dataset.

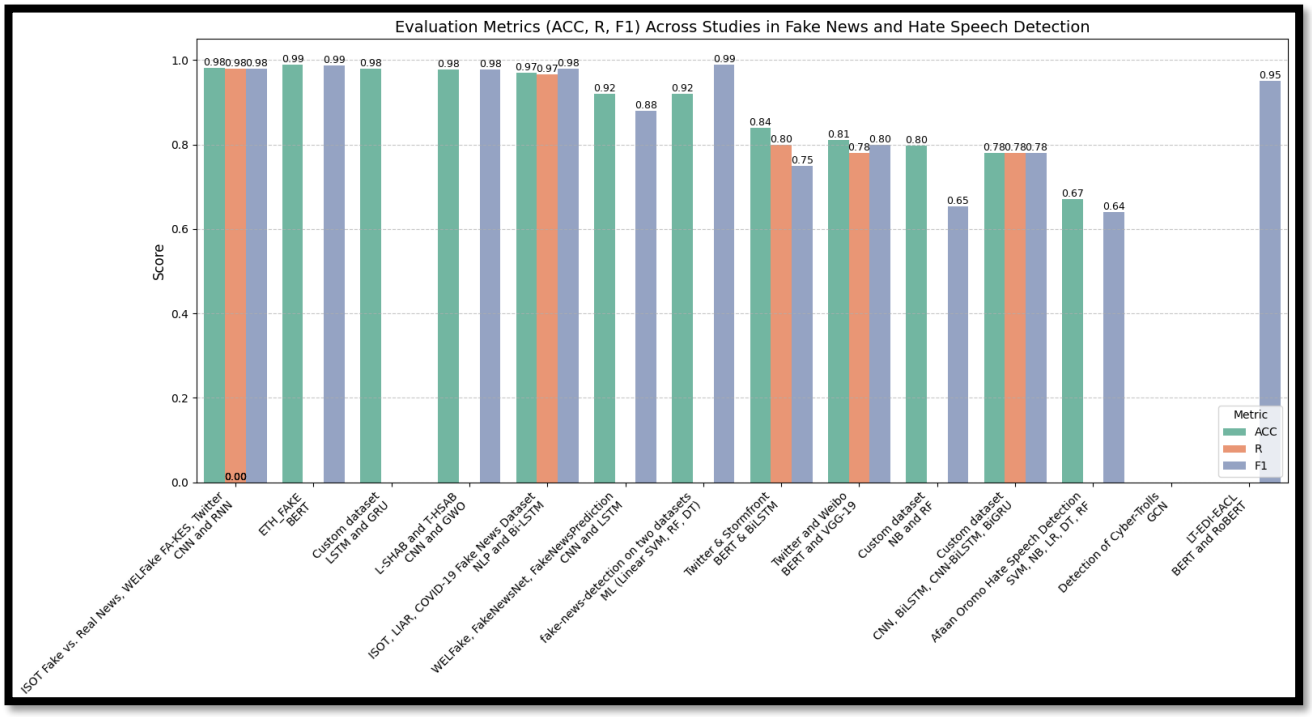

Fig. 8. Summary of the Dataset and Techniques used

As presented in Fig. 8, the analysis indicates that datasets such as ISOT, WELFake, FA-KES, Twitter, ETH_FAKE, L-SHAB, and T-HSAB demonstrated strong performance when evaluated using advanced deep learning and transformer-based models.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 8. Comparative Analysis of Multimodal and Multi-Class Deep Learning Models for Information Disorder Detection

Ref.	Problem addressed	Dataset used	Techniques used	Evaluation metrics		
				ACC	R	F1
[75]	Fake news detection	WELFake, FakeNewsNet, and FakeNewsPrediction	CNN and LSTM	0.92	-	0.88
[95]	Toxic content detection	Detection of Cyber-Trolls	GCN			
[76]	Fake news	Fake-news-detection on two data sets	ML (Linear SVM, RF, DT, and RF	0.92	-	0.99
[77]	Fake news	Twitter and Weibo	BERT and VGG-19 model	0.81	0.78	0.80
[87]	Mal-information (Harmful Memes)	Harm-C (COVID-19), Harm-P (Political memes), FHM (Facebook Hateful Memes	LLM; cross-attention fusion; T5-encoder–decoder	0.8958 (Harm-P)	-	0.8957 (Harm-P)
[82]	Mal-information (Hate speech)	*GOAT-Bench:*	LMM (GPT-4V, LLaVA, and Qwen-VL)	0.7217	-	0.7029
[53]	Rumor detection	Twitter (PHEME)	GCNs		0.25-0.30	
[79]	Rumor detection	TWITTER, WEIBO (COVID-19)	XLM-R BiGCN	0.873		0.861
[96]	Rumor detection	PHEME, Weibo dataset	Ensemble learning, Multi-BiLSTM	0.042	-	0.12 0.079
[81]	False information	Weibo and Twitter datasets	BERT and Swin-Transformer		0.043	
[39]	Hate speech and Fake news (Amharic)	Systematic survey	Deep learning and Machine learning		-	
[33]	Hate Speech Detection, Afaan Oromo	Afaan Oromo Hate Speech Detection	ML (SVM, NB, LR, DT, and RF)	0.67		0.64
[34]	Hate speech, Amharic	A dataset created by the authors, a total of 6120	Machine learning (NB, and RF)	0.7983		0.653
[35]	Hate speech, Amharic	A new dataset created by the authors, a total of 30,000	Deep learning (LSTM and GRU)	0.979	-	-
[37]	Hate speech, Amharic	A new dataset created a total of 2000 comments	Deep learning (CNN)			0.813
[36]	Hate speech, Amharic	The new dataset created by the authors uses a total of 6,497 files	DL (LSTM, BiLSTM, GRU, and BiGRU)	Improved accuracy 0.84 and 0.8815, respectively		
[86]	Fake news, Amharic	(ETH_FAKE)	BERT	0.9883,	-	0.98
[69]	Hateful and abusive	L-SHAB and *T-HSAB*	CNN, Grey Wolf Optimizer (GWO)	0.9783		0.978

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 9. Comparative Analysis of Deep Learning Models Across Datasets

Ref.	Year	Dataset used	Data Scale	Classification type (Binary vs Multiclass)	Data types	Techniques used	Evaluation metrics		
							ACC	R	F1
[73]	2025	ISOT Fake vs. Real News	45,000	Binary	Textual and visuals	CNN and RNN	0.9816	0.98	0.98
		WELFake	72,134	Binary					
		FA-KES	804	Binary					
[86]	2022	ETH_FAKE	6,834	Binary	Textual	BERT	0.9883,	-	0.9866
[35]	2020	Custom dataset	30,000	Binary	Textual	Deep learning (LSTM and GRU)	0.979	-	-
[69]	2024	L-SHAB	5,846	Multiclass	Textual	CNN and Grey Wolf Optimizer (GWO)	0.9783		0.9783
		T-HSAB	6,039	Multiclass	Textual				
[74]	2025	ISOT	45,000	Binary	Textual and visuals	NLP and Bi-LSTM	0.971	0.967	0.979
		LIAR	12,800	Multiclass					
		COVID-19 Fake News Dataset	10,700	Binary					
[75]	2024	WELFake	72,134	Binary	Textual	CNN and LSTM	0.92	-	0.88
		FakeNewsPrediction	6,400	Binary					
[89]	2025	Twitter & Stormfront	10,944	Binary	Textual	BERT & BiLSTM	0.84		0.80 0.75
[77]	2022	Twitter	13924	Binary	Textual and visual	BERT and VGG-19 model	0.81	0.78	0.80
		Weibo	9528	Binary					
[34]	2018	Custom dataset	6, 120	Binary	Textual	Machine learning (NB, and RF)	0.7983		0.6534
[82]	2025	Custom dataset	6,626	Multiclass	Textual and image	LMM (GPT-4V, LLaVA, and Qwen-VL)	0.7217	-	0.7029
[84] [85]	2025	Custom dataset	30,000	Binary (Bilingual)	Textual	CNN, BiLSTM, CNN-BiLSTM, and BiGRU	0.7805	0.78	0.78
[33]	2021	Afaan Oromo Hate Speech Detection	13600	Binary	Textual	ML (SVM, NB, LR, DT, and RF)	0.67		0.64
[88]	2025	LT-EDI-EACL	3114,2978	Multiclass (Bilingual)	Textual	BERT and RoBERT	-	-	0.95

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Studies combining textual and visual data often demonstrated improved robustness and performance, emphasizing the benefit of multimodal approaches. Recent studies (2024–2025) highlight the growing use of hybrid and transformer-based architectures, with reported F1-scores reaching up to 0.98. Overall, the reviewed studies indicate that substantial research efforts have explored information disorder detection using machine learning and deep learning approaches across public and custom datasets. The analysis suggests that deep learning models provide promising approaches for both unimodal and multimodal information disorder detection tasks.

Table 10. Comparison of Deep Learning and Machine Learning Model Performance

Ref	Year	Techniques / Models Used	Dataset Type	Accuracy	Recall	F1-Score
[73]	2025	CNN + RNN	Textual & Visual	**0.9816**	0.980	0.980
[86]	2022	BERT	Textual	**0.9883**	–	0.9866
[35]	2020	LSTM + GRU	Textual	**0.979**	–	–
[69]	2024	CNN + Grey Wolf Optimizer (GWO)	Textual	**0.9783**	–	0.9783
[74]	2025	Bi-LSTM	Textual & Visual	**0.971**	0.967	0.979
[75]	2024	CNN + LSTM	Textual	**0.920**	–	0.880
[76]	2024	Linear SVM, RF, DT	Textual	**0.920**	–	0.990
[89]	2025	BERT + BiLSTM	Textual	**0.840**	–	0.750
[77]	2022	BERT + VGG-19	Textual & Visual	**0.810**	0.780	0.800
[34]	2018	NB + RF	Textual	**0.7983**	–	0.6534
[84][85]	2025	CNN, BiLSTM, CNN-BiLSTM, BiGRU	Textual	**0.7805**	0.780	0.780
[33]	2021	SVM, NB, LR, DT, RF	Textual	**0.670**	–	0.640
[95]	2024	Graph Convolutional Network (GCN)	Textual	–	–	–
[88]	2025	BERT + RoBERTa	Textual	–	–	0.950

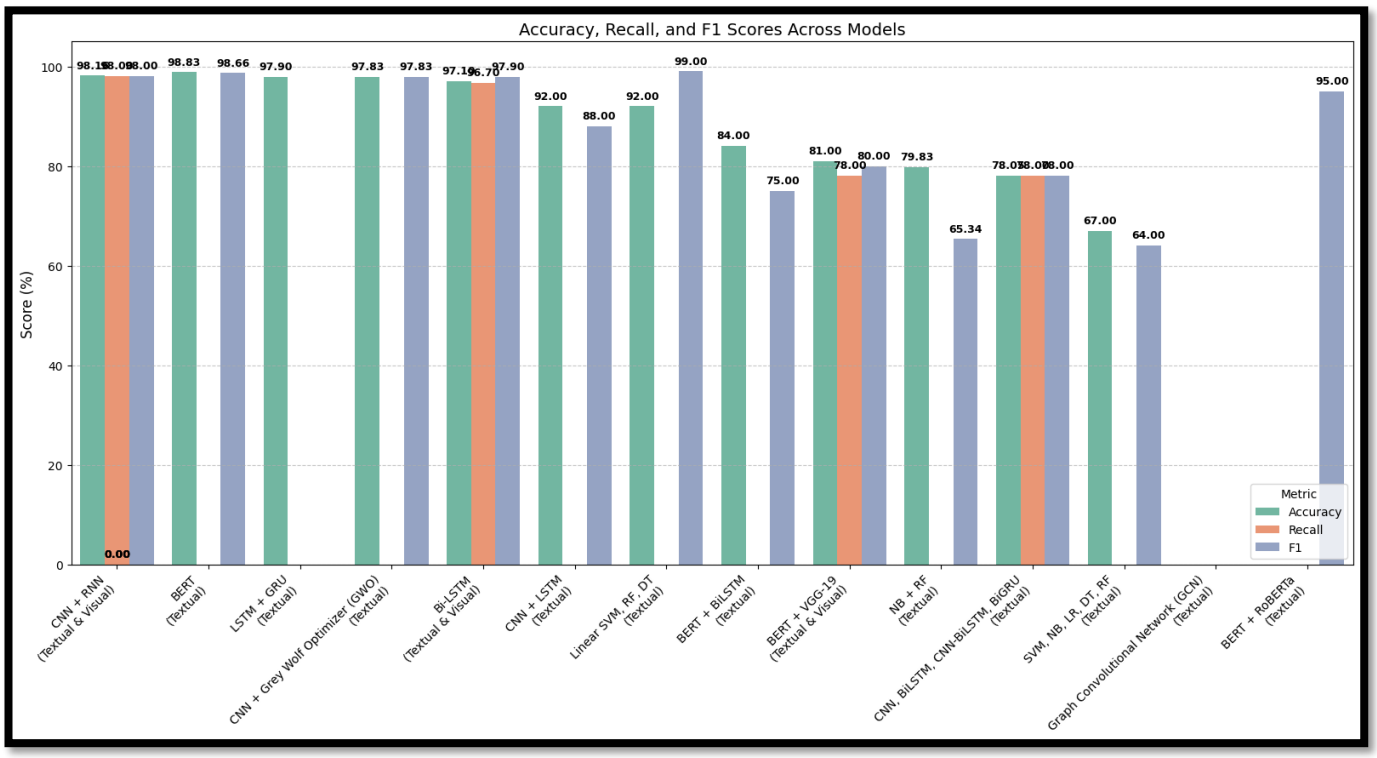

Fig. 9. Evaluation of deep learning and machine learning models based on Accuracy, Recall, and F1

As presented in Fig. 9, the evaluation results show that deep learning models, such as CNN-RNN and Bi-LSTM, perform strongly on combined text and visual data, while BERT and BERT–RoBERTa perform strongly on textual data by capturing contextual semantics.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 11. Summary of Multimodal Studies with Evaluation Metrics, Key Contributions, and Research Gaps

Ref.	Year	Language Used	Dataset Used	Data Types	Techniques Used	Evaluation Metrics	Key Contribution	Research Gaps
[97]	2024	Multilingual	MDAM3-DB	Text, Image, Audio, Video	Multimodal annotations; misinformation fabrication strategies	–	Created >7M multimodal dataset enabling multiple tasks (deep fake, out-of-context detection)	The dataset is still new; it lacks fine-grained multimodal alignment baselines
[73]	2023	English	ISOT, Fake vs Real, WELFake, FA-KES, Twitter	Text + Image	Multi-level fusion CNN + dual conv + RNN	Acc: 0.9725, 0.9108, 0.9816, 0.5403, 0.9163	Developed early and late fusion to improve semantic representation	The study lacks the integration and fusion of video/audio and memes
[74]	2023	English	ISOT, LIAR, COVID-19	Text + Image	MM-FND (Word2Vec, TF-IDF, Bi-LSTM, NER GloVe)	Acc: 0.963, 0.956, 0.971	Integrates temporal, spatial, and global features	Lacks in addressing cross-modal attention; no large-scale dataset
[98]	2023	English	Fake news dataset	Text + Image	O-BiLSTM + AWS Algorithm; VGG16; ResNet50	Acc: 0.9651	Developed an Optimized multimodal fusion with swarm intelligence	High computational cost; limited generalizability
[99]	2024	Multilingual	Systematic Review (2020–2024)	Text + Image + Hybrid	ML, DL, Hybrid, Transformers	–	Presents the need for XAI and multilingual datasets	Missing multimodal benchmarks for low-resource languages
[100]	2023	Multilingual	Survey (Multimodal)	Text, Image, Video	Deep learning, Graph NN, Transformers	–	Tracks the evolution of multimodal fusion techniques	Lacks a unified evaluation framework
[101]	2023	Chinese/ English	Weibo, Twitter, Pheme	Text + Image	MIGCL + hierarchical graph contrastive learning	Weibo (Acc:0.90 P:0.895 R:0.911 F1:0.903)	Develops cross-modal alignment and GCN for better fusion	The study lacks testing on multilingual datasets
[1]	2023	English	PolitiFact, MMHS150K, MultiOFF	Text + Image	Inter-modal attention, LAVIS captions, EasyOCR, GloVe	Acc: 0.940 F1: 0.939	Converts multimodal content into a unified text representation	The study lacks a detailed analysis of fusion with the respective modalities.
[102]	2024	Amharic	2,000 Amharic Memes	Text + Image	VGG16 + Word2Vec; LSTM/ Bi-LSTM/CNN	Acc: Text 0.63, Multimodal 0.75	First multimodal Amharic meme hate speech model developed	The study lacks detailed fusion analysis, a tiny dataset, and OCR accuracy issues

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

Table 11 presents related multimodal studies along with evaluation metrics, key contributions, and research gaps. Classical machine learning models such as SVM, Random Forest, and Naive Bayes show moderate to low performance, particularly on complex datasets. The reviewed studies report that several high-performing models achieved accuracy values in the range of 98–99%, with CNN+RNN, BERT, LSTM+GRU, and CNN+GWO reporting values above 98%. BiLSTM and CNN+LSTM remain competitive with reported performance in the range of 92–97%, while BERT+BiLSTM and BERT+VGG19 report lower values in the range of 75–84%. Machine learning models such as NB+RF and SVM/LR/DT/RF perform lower, ranging from 64% to 80%. The Graph Convolutional Network reported comparatively lower performance, while BERT+RoBERTa achieved the highest reported F1-score at 95%. Overall, the reviewed studies indicate strong performance of textual deep-learning models; however, comparisons across modalities and datasets should be interpreted cautiously.

As presented in Table 11, the comparative analysis indicates extensive use of deep learning models for multimodal data representation and detection. Furthermore, limited research effort was observed on multimodal content and low-resource languages for both textual and multimodal analysis, and no standard dataset was identified for text–image fusion tasks. In addition, lower performance was observed for low-resource languages compared with studies evaluated using benchmark datasets such as Weibo, ISOT, and Twitter. To develop multimodal models, a wide range of data modalities, including text, images, memes, audio, and video, are shared across social media platforms. Effectively integrating these multimodalities is still a major problem, even though the accessibility of such diverse information enhances content representation. Several fusion strategies, including early fusion, intermediate fusion, late fusion, and cross-attention-based multimodal methods, have been proposed in the literature to solve this problem [103]-[105].

In this study, early fusion (feature-level fusion) is considered because of its reported effectiveness in multimodal tasks such as sentiment and text–image analysis [106]. Compared with separate unimodal approaches, it has been reported to improve meme sentiment classification performance [107]. This technique integrates inputs from several modalities into a single feature that is subsequently fed into a learning model. Techniques like concatenation, pooling, and gated units can be used for this. Type I early fusion combines original characteristics, whereas Type II early fusion combines features extracted by a different neural network. This approach captures correlations between modalities at an early stage, enabling richer representation learning [108],[109]. A unified feature vector is first created and then used as input to a deep learning or machine learning classifier [110]. It preserves original unified information of each modality, simplifies model architecture by merging data before learning, improves cross-modal correlation learning, and reduces computational complexity [104].

Joint Fusion (Intermediate Fusion) often referred to as intermediate fusion, uses inputs from multiple modalities and learned feature representations from neural network intermediate layers as input to a final model. Unlike early fusion, joint fusion allows the neural networks to gradually improve the feature representations with each iteration by backpropagating the training loss. In particular, by extracting and combining feature representations from all modalities, Type I joint fusion supports multimodal interaction and may improve model performance [111]. Late or Decision-level fusion uses aggregation techniques like averaging, weighted voting, majority voting, or stacking to generate a final prediction by combining predictions from several models. Late fusion involves training a separate unimodal model using each modality, then integrating the predictions. This method works especially well when one modality dominates or when each unimodal model performs well on its own.

Because late fusion optimizes each model for its particular modality, it allows for the learning of strong marginal representations. By reducing the influence of individual model errors, it may improve overall performance; nevertheless, significant improvements are only made when the unimodal models offer complementary data [112]. As shown in Fig. 10, a multimodal architecture is presented that combines text and visual components to classify Amharic social media content. Text features are proposed to be extracted using transformer models (BERT, RoBERTa, ALBERT, DistilBERT), whilst picture and meme features are extracted using deep learning models (ResNet-18/50, EfficientNet-B0, ViT-B/16). Embedded text is captured by OCR. The architecture incorporates early (feature-level), joint (intermediate), and late (decision-level) fusion strategies.

RQ3: What are the contextual research gaps for linguistic localization of scalable social media datasets, and deep learning model accuracy for handling futuristic multimodal, multilingual, and multi-class classification tasks?

As illustrated in Table 12, and Fig. 11, existing linguistic localized studies on Amharic and Afaan Oromo information disorder and disruption detection predominantly rely on textual and small-scale datasets, on specific types of abusive content, i.e., offense, hate, fake, with limited multimodal exploration.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

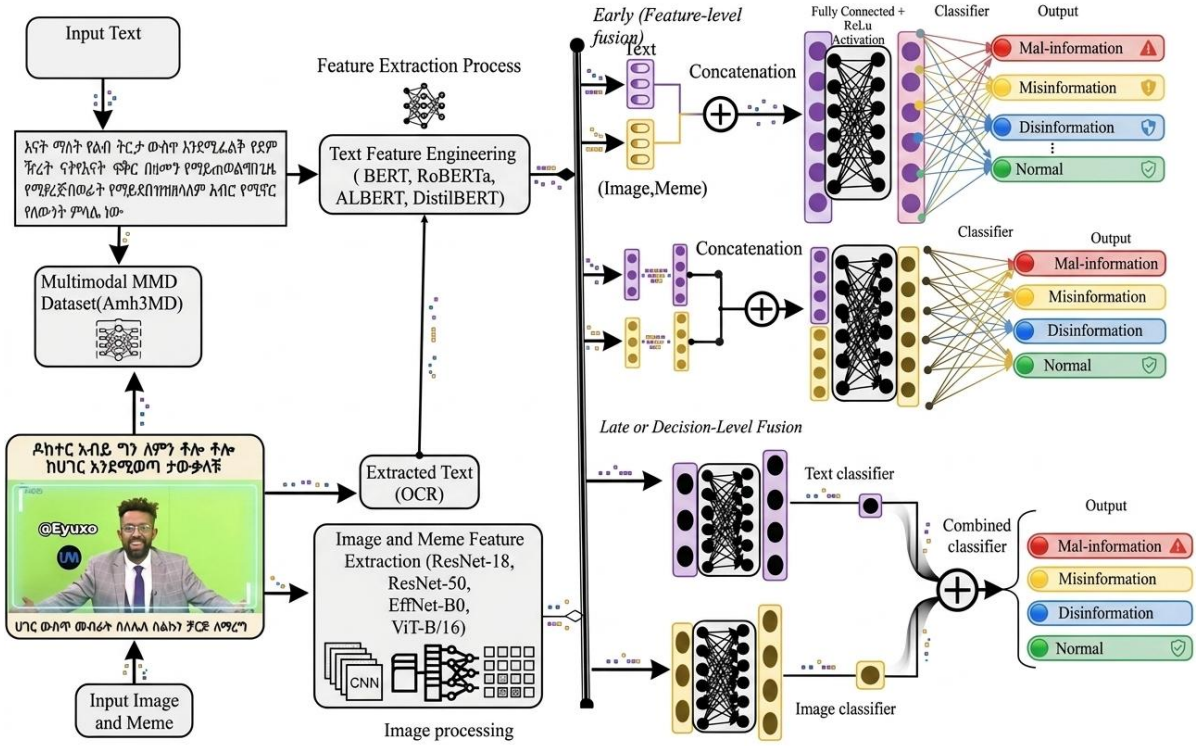

Fig. 10. Multimodal Fusion Architecture for Amharic Information Disorder Classification

Table 12. Summarizing the Literature Survey for Linguistic Localization and Deep Learning Models toward information disruption

Ref.	Problem Addressed	Dataset Used	Data Types	Techniques Used	ACC	R	F1	Remarks
[86]	Fake news, Amharic	ETH_FAKE	Textual	DL (BERT)	0.98	–	0.98	No fact-check validation; lacks multimodal and bi-classification; limited inclusion of other abusive information.
[36]	Hate speech Amharic)	6,497 files	Audio & Text	DL (LSTM, BiLSTM, GRU, BiGRU)	0.84–0.881	–	–	Small dataset; lacks generalization and multimodal expansion beyond audio-text; no coverage of other abusive speech types.
[35]	Hate speech, Amharic	Custom dataset (30,000)	Textual	DL (LSTM, GRU)	0.979	–	–	Lacks comparison with BiLSTM; absence of public datasets; lacks model compression and multimodal integration.
[34]	Hate speech, Amharic	Custom dataset (6,120)	Textual	ML (NB, RF)	0.7983	–	0.653	Small dataset size; lacks comparison with other ML models and multimodal integration.
[84] [85]	Hate speech, Afaan Oromo	Custom dataset	Textual	CNN, BiLSTM, CNN-BiLSTM, BiGRU	0.7805	0.78	0.78	Lacks bilingual hate speech detection (Amharic–Afaan Oromo); no handling of code-mixed language.
[33]	Hate speech Afaan Oromo	Afaan Oromo Hate Speech Detection	Textual	SVM, NB, LR, DT, RF	0.67	–	0.64	Lower accuracy and F1 score; lacks analysis and comparison with state-of-the-art; lacks multimodal integration and bi-classification.
[37]	Hate speech, Amharic	2,000 annotated social media comments	Textual	DL (CNN)	–	–	0.813	Small dataset; lacks full metric evaluation (only F1); lacks multimodal and bi-classification integration.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

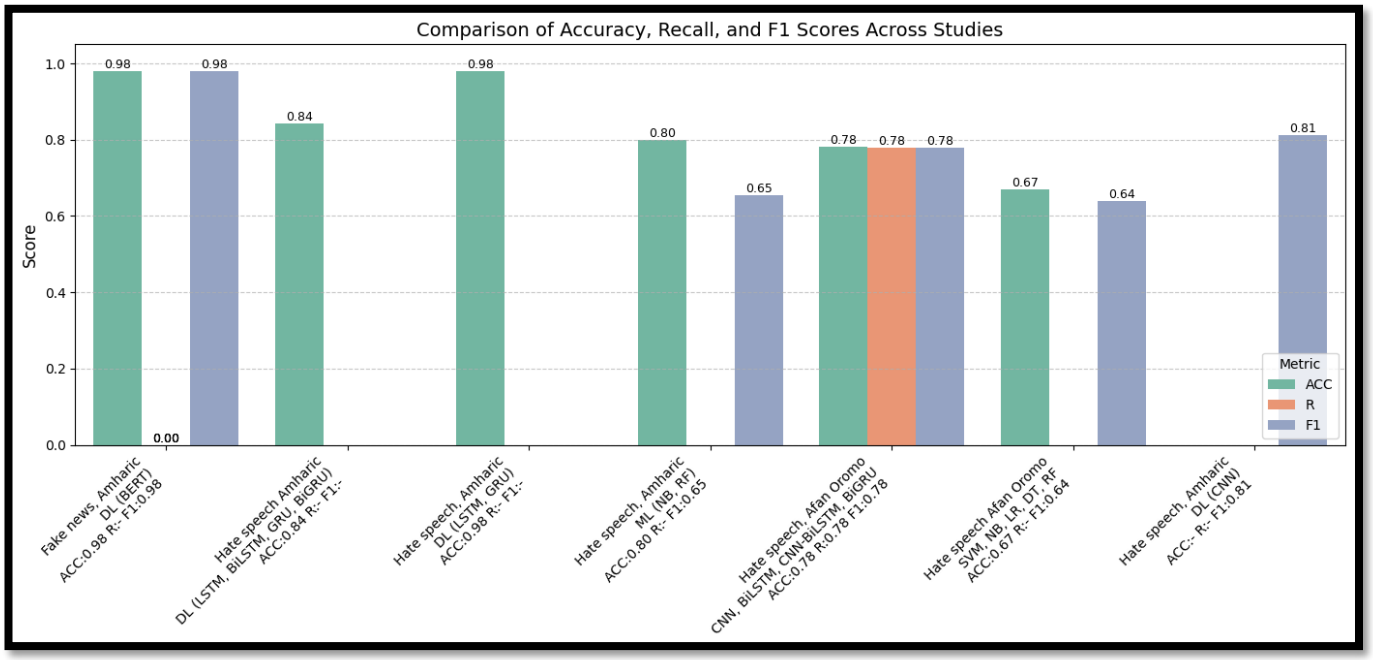

Fig. 11. Summarizing literature survey for Linguistic Localization and Deep Learning Model

As presented in Fig. 11, the summary results of Linguistic Localization and Deep Learning Models toward information disruption show that deep learning models, such as CNN, LSTM, GRU, and BERT, have achieved a comparatively high accuracy. The results show that Amharic models perform strongest, with BERT achieving 98% accuracy, recall, and F1. Other Amharic models also perform well, scoring accuracies ranging from 84% to 90% accuracy. In contrast, Afaan Oromo models perform lower, with the BiLSTM/CNN model at 78% across all metrics and the CNN+BiRF model between 64-67%. Overall, Amharic models consistently outperform Afaan Oromo models on publicly available datasets.

However, the limited research effort has done on consolidation of information disorder and disruption on social media content and lack of scalable social media datasets in the Ethiopian context, localized context-aware deep learning models' capabilities towards handling multimodal, i.e., visual and textual, multilingual, multi-class classification with an optical optimization technique and its appropriate annotations tasks persist as serious research gaps and can be explored as future work. In addition, the emergence and widespread adoption of large language models and multimodal vision-language models represent a pressing concern in the context of information disorder, necessitating focused investigation in future work.

6 Discussion

The systematic review reveals that a methodological shift from machine learning models toward deep learning and transformer-based learning models, particularly CNNs, LSTMs, and BiLSTMs, and transformer-based architectures like BERT, consistently achieve high accuracy in information disorder detection, often exceeding 97% on large benchmark datasets like ISOT, WELFake, and L-SHAB. Hybrid models, including CNN-BiLSTM and CNN with Grey Wolf Optimizer, also show competitive performance. For low-resource languages such as Amharic and Afaan Oromo, deep learning approaches outperform classical ML, offering higher robustness and generalizability even on smaller datasets, suggesting that transformer-based and recurrent models are preferable for developing reliable detection. Due to the ability to learn contextual and hierarchical representations, which is crucial for morphologically rich and syntactically flexible handling of noisy, informal, and context-dependent social media content.

However, transformer-based models' substantial computational inefficiencies and storage requirements present resource deployment challenges. To overcome these challenges, transformers need to be further explored for a multiclass classification task for multimodal data. Fig. 11, shows a comparative analysis of state-of-the-art approaches, emphasizing the performance of advanced learning models. The evaluation of global versus localized models across diverse data modalities demonstrates the effectiveness of deep learning, particularly in multimodal data. The review and quantitative comparative analysis result shows that there are significant research gaps in existing studies, particularly in the integration and fusion of multimodal components such as text, image, memes, video, and audio, and a lack of a standard publicly available dataset with appropriate annotation. In addition, the existing research studies focused on binary classification for low-resource languages. Therefore, this study presents an urgent need for innovative approaches that address these challenges, ensuring comprehensive solutions for combating misinformation, disinformation, and mal-information with multiclass classification in today's digital landscape.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

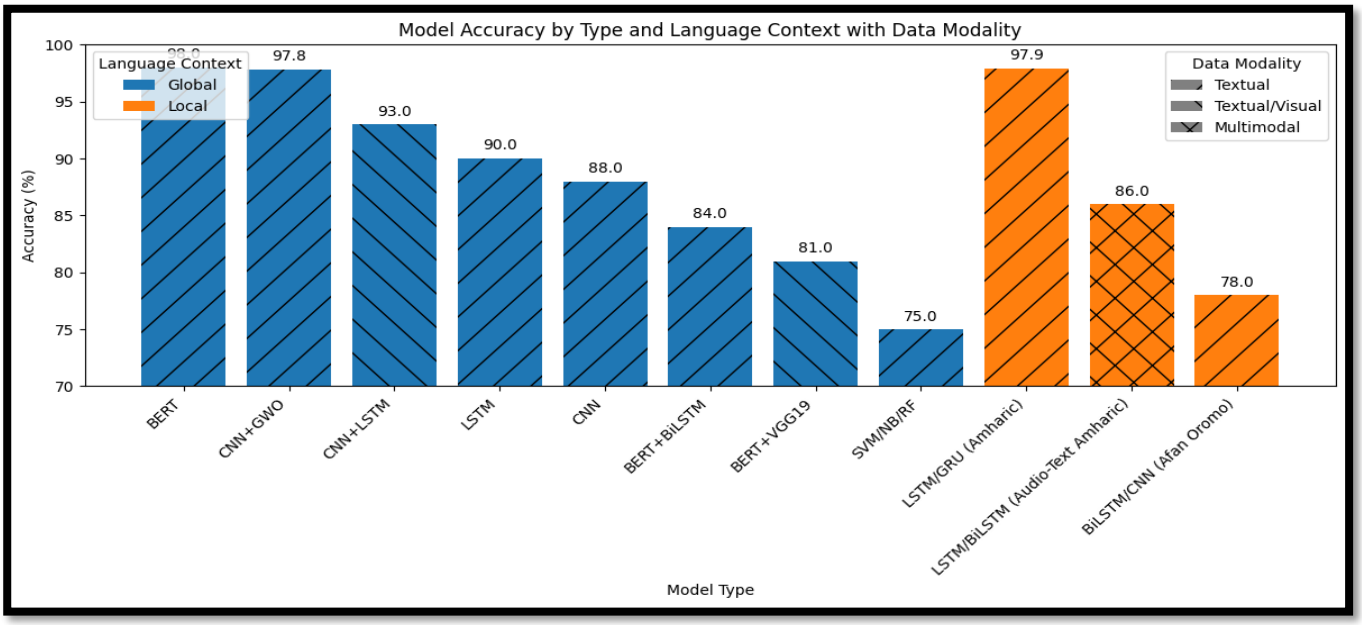

Fig. 12. Summary of Linguistic content, data modalities, and Deep learning models Towards information disruption

As illustrated in Fig. 12, the BERT model achieves the highest accuracy at 98.2%, followed closely by the CNN+GWO model at 97.8%. The local linguistic language local Amharic language by the LSTM/BiLSTM model performs strongly, achieving 97.9% and nearly matching the top global models. In contrast, multimodal and vision-based approaches show reduced performance, with SWIN+BiRF recording the lowest accuracy at 75% and the Afaan Oromo BiLSTM/CNN model reaching 78%.

7 Conclusion

This systematic review provides a comprehensive assessment of state-of-the-art deep learning, machine learning, and ensemble models used for detecting misinformation, disinformation, and mal-information on social media platforms. The systematic literature review with quantitative comparative analysis indicates that deep learning models, particularly Transformer, CNN, RNN, LSTM, BERT, and GCN architectures, demonstrated strong performance across various datasets using standard evaluation metrics. Existing research studies primarily focused on binary classification, particularly using unimodal textual data. However, significant challenges remain, including limited resources, the lack of standardized public datasets, and limited integration of multimodal data types such as text, images, memes, video, and audio, particularly for low-resource languages such as Amharic and Afaan Oromo. This study provides insights relevant to practical deployment and policy development that may support Ethiopian language-inclusive digital governance initiatives, fact-checking efforts, and social media monitoring for information disorder detection. Future research may focus on developing scalable, context-aware deep learning models for handling multimodal data. Additionally, emphasis may be placed on creating large, well-annotated datasets for low-resource languages, such as Amharic and Afaan Oromo, and exploring feature fusion-based models to improve detection. Furthermore, integrating LLMs with explainable AI (XAI) methods, such as feature importance analysis and attention visualization, may support accountability, improve model reliability, and enable better validation of classification outcomes. In addition, investigating multilingual models and refining existing frameworks to better handle the dynamic nature of social media language remain important future research directions.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare no conflicts of interest related to this study.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

REFERENCES

[1] E. F. Ayetiran and Ö. Özgöbek, "An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection," *Information Systems*, vol. 123, p. 102378, Mar. 2024, doi: 10.1016/j.is.2024.102378.

[2] "Press conference by Secretary-General António Guterres at United Nations Headquarters | UN meetings coverage and press releases," Jun. 24, 2024. https://press.un.org/en/2024/sgsm22284.doc.htm

[3] P. Muñoz, F. Díez, and A. Bellogín, "Modeling disinformation networks on Twitter: structure, behavior, and impact," *Applied Network Science*, vol. 9, no. 1, Jan. 2024, doi: 10.1007/s41109-024-00610-w.

[4] H. R. Saeidnia, E. Hosseini, B. Lund, M. A. Tehrani, S. Zaker, and S. Molaei, "Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches," *Knowledge and Information Systems*, vol. 67, no. 4, pp. 3139–3158, Jan. 2025, doi: 10.1007/s10115-024-02337-7.

[5] P. Akhtar *et al.*, "Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions," *Annals of Operations Research*, vol. 327, no. 2, pp. 633–657, Nov. 2022, doi: 10.1007/s10479-022-05015-5.

[6] H. Wang, R. Czerminski, and A. C. Jamieson, "Neural Networks and Deep Learning," in *The Machine Age of Customer Insight*, 2021, pp. 91–101. doi: 10.1108/978-1-83909-694-520211010.

[7] I. H. Sarker, "Deep Learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[8] A. Tursunbayeva, M. Franco, and C. Pagliari, "Use of social media for e-Government in the public health sector: A systematic review of published studies," *Government Information Quarterly*, vol. 34, no. 2, pp. 270–282, Apr. 2017, doi: 10.1016/j.giq.2017.04.001.

[9] S. Kaur, S. Singh, and S. Kaushal, "Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 104–122, Jan. 2024, doi: 10.1016/j.ijcce.2024.02.002.

[10] Hemant Kumar Soni, Sanjiv Sharma, G. R. Sinha, *Text and Social Media Analytics for Fake News and Hate Speech Detection*. Chapman and Hall/CRC, 2025, doi: 10.1201/9781003409519.

[11] S. Harris, H. J. Hadi, N. Ahmad, and M. A. Alshara, "Fake News Detection Revisited: An Extensive Review of Theoretical Frameworks, Dataset Assessments, Model Constraints, and Forward-Looking Research Agendas," *Technologies*, vol. 12, no. 11, 2024, doi: 10.3390/technologies12110222.

[12] S. Yadav, A. Kesharwani, and D. Sharma, "Blurred Boundaries of Truth: A review of deepfakes and fake news," *Journal of Internet Commerce*, vol. 25, no. 2, pp. 240–262, Dec. 2025, doi: 10.1080/15332861.2025.2598809.

[13] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, Feb. 2023, doi: 10.1007/s13278-023-01028-5.

[14] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, no. 1, p. 82, Sep. 2020, doi: 10.1007/s13278-020-00696-x.

[15] C. L. Bockting, E. A. M. Van Dis, R. Van Rooij, W. Zuidema, J. Bollen, "Living guidelines for generative AI-why scientists must oversee its use," *Nature*, vol. 622, no. 7984, pp. 693–696, 2023,doi: 10.1038/d41586-023-03266-1.

[16] M. Hutson, "Rules to keep AI in check: nations carve different paths for tech regulation.," *Nature*, vol. 620, no. 7973, pp. 260–263, 2023, doi:10.1038/d41586-023-02491-y.

[17] T. Nagasako, "Global disinformation campaigns and legal challenges," *International Cybersecurity Law Review*, vol. 1, no. 1–2, pp. 125–136, Oct. 2020, doi: 10.1365/s43439-020-00010-7.

[18] M. Chakraborty *et al.*, "FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W Question-Answering," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 15282–15322, 2023, doi:10.18653/v1/2023.emnlp-main.945.

[19] I. H. Sarker, "Deep Learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[20] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, Feb. 2023, doi: 10.1007/s13278-023-01028-5.

[21] J. Baptista and A. Gradim, "A working definition of fake news," *Encyclopedia*, vol. 2, no. 1, pp. 632–645, Mar. 2022, doi: 10.3390/encyclopedia2010043.

[22] Z. Bastick, "Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation," *Computers in Human Behavior*, vol. 116, p. 106633, Nov. 2020, doi: 10.1016/j.chb.2020.106633.

[23] R. Cohen-Almagor, "Freedom of expression v. social responsibility: Holocaust denial in Canada," *Repository@Hull (Worktribe) (University of Hull)*, Jan. 2013, doi: 10.1080/08900523.2012.746119.

[24] F. Mehmood, H. Ghafoor, M. N. Asim, M. U. Ghani, W. Mahmood, and A. Dengel, "Passion-Net: a robust precise and explainable predictor for hate speech detection in Roman Urdu text," *Neural Computing and Applications*, vol. 36, no. 6, pp. 3077–3100, Nov. 2023, doi: 10.1007/s00521-023-09169-6.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

[25] N. Alkiviadou, "Platform liability, hate speech and the fundamental right to free speech," *Information & Communications Technology Law*, vol. 34, no. 2, pp. 207–217, Oct. 2024, doi: 10.1080/13600834.2024.2411799.

[26] L. Anderson and M. Barnes, "Hate Speech," *The Stanford Encyclopedia of Philosophy* (Summer 2025 Edition), Metaphysics Research Lab, Stanford University, 2023. https://plato.stanford.edu/archives/sum2025/entries/hate-speech/

[27] J. L. Imbwaga, N. B. Chittaragi, and S. G. Koolagudi, "Automatic hate speech detection in audio using machine learning algorithms," *International Journal of Speech Technology*, vol. 27, no. 2, pp. 447–469, Jun. 2024, doi: 10.1007/s10772-024-10116-6.

[28] E. Hashmi and S. Y. Yayilgan, "Multi-class hate speech detection in the Norwegian language using FAST-RNN and multilingual fine-tuned transformers," *Complex & Intelligent Systems*, vol. 10, no. 3, pp. 4535–4556, Mar. 2024, doi: 10.1007/s40747-024-01392-5.

[29] A. Mousa, I. Shahin, A. B. Nassif, and A. Elnagar, "Detection of Arabic offensive language in social media using machine learning models," *Intelligent Systems With Applications*, vol. 22, p. 200376, Apr. 2024, doi: 10.1016/j.iswa.2024.200376.

[30] United Nations, "What is hate speech? | United Nations," *United Nations*. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[31] B. Belay, T. Habtegebrial, M. Meshesha, M. Liwicki, G. Belay, and D. Stricker, "Amharic OCR: an End-to-End Learning," *Applied Sciences*, vol. 10, no. 3, p. 1117, Feb. 2020, doi: 10.3390/app10031117.

[32] S. T. Abate *et al.*, "Large vocabulary read speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta," *ACL Anthology*, May 01, 2020. https://aclanthology.org/2020.lrec-1.513/

[33] Naol Bakala Defersha, Kula Kekeba Tune, "Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Approach," *Indian Journal of Science and Technology*, vol. 14, no. 31, pp. 2567–2578, Aug. 2021, doi: 10.17485/ijst/v14i31.1019.

[34] Z. Mossie, J.-H. Wang, and others, "Social network hate speech detection for Amharic language," *Computer Science & Information Technology*, Academy & Industry Research Collaboration Center, pp. 41–55, 2018, doi:10.5121/csit.2018.80604.

[35] S. G. Tesfaye and K. Kakeba, "Automated Amharic hate speech posts and comments detection model using recurrent neural network," *Research square*, Dec. 2020, doi: 10.21203/rs.3.rs-114533/v1.

[36] A. G. Debele and M. M. Woldeyohannis, "Multimodal Amharic Hate Speech Detection Using Deep Learning," *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, Bahir Dar, Ethiopia, 2022, pp. 102-107, doi: 10.1109/ICT4DA56482.2022.9971436.

[37] Z. Abebaw, A. Rauber, and S. Atnafu, "Multi-channel convolutional neural network for hate speech detection in social media," in *Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2022, pp. 603–618. doi: 10.1007/978-3-030-93709-6_41.

[38] A. Minaye and T. Megersa, "Ethnic-based online hate speech in Ethiopia: its typology and context," *National Academic Digital Repository of Ethiopia*, Jun. 2023, doi: 10.20372/ejss.v9i1.1643.

[39] W. B. Demilie and A. O. Salau, "Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches," *Journal of Big Data*, vol. 9, no. 1, p. 66, May 2022, doi: 10.1186/s40537-022-00619-x.

[40] F. Gereme, W. Zhu, T. Ayall, and D. Alemu, "Combating fake news in 'Low-Resource' languages: amharic fake news detection accompanied by resource crafting," *Information*, vol. 12, no. 1, p. 20, Jan. 2021, doi: 10.3390/info12010020.

[41] "Global Risks Report 2024 | World Economic Forum," *World Economic Forum*, Aug. 25, 2025. https://www.weforum.org/publications/global-risks-report-2024/

[42] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere," *Applied Sciences*, vol. 10, no. 23, p. 8614, Dec. 2020, doi: 10.3390/app10238614.

[43] T. T. Aurpa, R. Sadik, and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," *Social Network Analysis and Mining*, vol. 12, no. 1, Dec. 2021, doi: 10.1007/s13278-021-00852-x.

[44] S. R. Dube *et al.*, "Childhood verbal abuse as a child maltreatment subtype: A systematic review of the current evidence," *Child Abuse & Neglect*, vol. 144, p. 106394, Aug. 2023, doi: 10.1016/j.chiabu.2023.106394.

[45] T. T. Aurpa, R. Sadik, and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," *Social Network Analysis and Mining*, vol. 12, no. 1, Dec. 2021, doi: 10.1007/s13278-021-00852-x.

[46] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A comprehensive review of DeepFake detection using advanced machine learning and fusion methods," *Electronics*, vol. 13, no. 1, p. 95, Dec. 2023, doi: 10.3390/electronics13010095.

[47] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *International Journal of Surgery*, vol. 88, p. 105906, Mar. 2021, doi: 10.1016/j.ijsu.2021.105906.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

[48] C. Sohrabi *et al.*, "PRISMA 2020 statement: What's new and the importance of reporting guidelines," *International Journal of Surgery*, vol. 88, p. 105918, Mar. 2021, doi: 10.1016/j.ijsu.2021.105918.

[49] J. R. Polanin, T. D. Pigott, D. L. Espelage, and J. K. Grotpeter, "Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses," *Research Synthesis Methods*, vol. 10, no. 3, pp. 330–342, May 2019, doi: 10.1002/jrsm.1354.

[50] C. Hamel *et al.*, "Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses," *BMC Medical Research Methodology*, vol. 21, no. 1, p. 285, Dec. 2021, doi: 10.1186/s12874-021-01451-2.

[51] G. K. Frampton, B. Livoreil, and G. Petrokofsky, "Eligibility screening in evidence synthesis of environmental management topics," *Environmental Evidence*, vol. 6, no. 1, Sep. 2017, doi: 10.1186/s13750-017-0102-2.

[52] S. Jena *et al.*, "Developing a negative speech emotion recognition model for safety systems using deep learning," *Journal of Big Data*, vol. 12, no. 1, Mar. 2025, doi: 10.1186/s40537-025-01090-0.

[53] M. Chabbouh, S. Bechikh, E. Mezura-Montes, and L. B. Said, "Evolutionary optimization of the area under precision-recall curve for classifying imbalanced multi-class data," *Journal of Heuristics*, vol. 31, no. 1, Jan. 2025, doi: 10.1007/s10732-024-09544-z.

[54] V. J. G. Genovés and M. J. B. Arrojo, "El control de la agresión sexual: manual para el terapeuta," *Dialnet (Universidad De La Rioja)*, vol. 6, no. 14, p. eaay3539, Jan. 1996, doi: 10.1126/sciadv.aay3539.

[55] V. K. Singh, I. Ghosh, and D. Sonagara, "Detecting fake news stories via multimodal analysis," *Journal of the Association for Information Science and Technology*, vol. 72, no. 1, pp. 3–17, May 2020, doi: 10.1002/asi.24359.

[56] S. Chen, L. Xiao, and A. Kumar, "Spread of misinformation on social media: What contributes to it and how to combat it," *Computers in Human Behavior*, vol. 141, p. 107643, Dec. 2022, doi: 10.1016/j.chb.2022.107643.

[57] G. Di Domenico, J. Sit, A. Ishizaka, and D. Nunan, "Fake news, social media and marketing: A systematic review," *Journal of Business Research*, vol. 124, pp. 329–341, Dec. 2020, doi: 10.1016/j.jbusres.2020.11.037.

[58] R. P. Bringula, A. E. Catacutan-Bangit, M. B. Garcia, J. P. S. Gonzales, and A. M. C. Valderama, "Who is gullible to political disinformation?' : predicting susceptibility of university students to fake news," *Journal of Information Technology & Politics*, vol. 19, no. 2, pp. 165–179, Jul. 2021, doi: 10.1080/19331681.2021.1945988.

[59] Z. Bastick, "Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation," *Computers in Human Behavior*, vol. 116, p. 106633, Nov. 2020, doi: 10.1016/j.chb.2020.106633.

[60] Y. Wu, E. W. T. Ngai, P. Wu, and C. Wu, "Fake news on the internet: a literature review, synthesis and directions for future research," *Internet Research*, vol. 32, no. 5, pp. 1662–1699, Mar. 2022, doi: 10.1108/intr-05-2021-0294.

[61] T. D. Adjin-Tettey, "Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education," *Cogent Arts and Humanities*, vol. 9, no. 1, Feb. 2022, doi: 10.1080/23311983.2022.2037229.

[62] M. Hameleers, A. Brosius, and C. H. De Vreese, "Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media," *European Journal of Communication*, vol. 37, no. 3, pp. 237–268, Feb. 2022, doi: 10.1177/02673231211072667.

[63] A. P. Weiss, A. Alwan, E. P. Garcia, and J. Garcia, "Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking," *International Journal for Educational Integrity*, vol. 16, no. 1, Feb. 2020, doi: 10.1007/s40979-019-0049-x.

[64] X. Zhou and R. Zafarani, "A survey of fake news," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, Jul. 2020, doi: 10.1145/3395046.

[65] F. Mehmood, H. Ghafoor, M. N. Asim, M. U. Ghani, W. Mahmood, and A. Dengel, "Passion-Net: a robust precise and explainable predictor for hate speech detection in Roman Urdu text," *Neural Computing and Applications*, vol. 36, no. 6, pp. 3077–3100, Nov. 2023, doi: 10.1007/s00521-023-09169-6.

[66] M. K. Singh, J. Ahmed, M. A. Alam, K. K. Raghuvanshi, and S. Kumar, "A comprehensive review on automatic detection of fake news on social media," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 47319–47352, Oct. 2023, doi: 10.1007/s11042-023-17377-4.

[67] Kai Shu, Suhang Wang, Dongwon Lee, Huan Liu, *Disinformation, Misinformation, and Fake News in Social Media*. Lecture Notes in Social Networks (LNSN), 2020, doi: 10.1007/978-3-030-42699-6.

[68] V. K. Sharma, R. Garg, Q. Caudron, "A systematic literature review on deepfake detection techniques," *Multimedia Tools and Applications*, vol. 84, no. 20, pp. 22187–22229, Aug. 2024, doi: 10.1007/s11042-024-19906-1.

[69] A. Aljohani, N. Alharbe, R. E. A. Mamlook, and M. M. Khayyat, "A hybrid combination of CNN Attention with optimized random forest with grey wolf optimizer to discriminate between Arabic hateful, abusive tweets," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 2, p. 101961, Feb. 2024, doi: 10.1016/j.jksuci.2024.101961.

[70] S. Eelmaa, "Sexualization of children in Deepfakes and hentai," *Trames*, vol. 26, no. 2, pp. 229–248, 2022, doi:10.3176/tr.2022.2.07.

[71] G. Xu, M. Qian, and L. Meng, "Misinformation dissemination on social media: key research themes and evolutionary paths between 2013 and 2023," *Humanities and Social Sciences Communications*, vol. 12, no. 1, Nov. 2025, doi: 10.1057/s41599-025-06067-1.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 36-66 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 07 Feb 2026 | Revised: 05 May 2026 | Accepted: 17 May 2026 | Published: 21 May 2026

[72] H. R. Saeidnia, E. Hosseini, B. Lund, M. A. Tehrani, S. Zaker, and S. Molaei, "Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches," *Knowledge and Information Systems*, vol. 67, no. 4, pp. 3139–3158, Jan. 2025, doi: 10.1007/s10115-024-02337-7.

[73] F. Abbas and A. Taeihagh, "A multi-level fusion-based framework for multimodal fake news classification using semantic feature extraction," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 9, pp. 6531–6560, May 2025, doi: 10.1007/s13042-025-02633-w.

[74] E. Alsuwat and H. Alsuwat, "An improved multi-modal framework for fake news detection using NLP and Bi-LSTM," *The Journal of Supercomputing*, vol. 81, no. 1, Nov. 2024, doi: 10.1007/s11227-024-06671-z.

[75] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," in *IEEE Access*, vol. 12, pp. 44462-44480, 2024, doi: 10.1109/ACCESS.2024.3381038.

[76] A. Aslam *et al.*, "Advancements in Fake News Detection: A comprehensive machine learning approach across varied datasets," *SN Computer Science*, vol. 5, no. 5, May 2024, doi: 10.1007/s42979-024-02943-w.

[77] N. Xiang, "Deep Learning-Based Fake Information Detection and Influence Evaluation," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–8, Feb. 2022, doi: 10.1155/2022/8514430.

[78] X. Lei, "Network Rumor Detection Method using deep learning in big data environment," *Mobile Information Systems*, vol. 2022, pp. 1–8, May 2022, doi: 10.1155/2022/6725840.

[79] H. Liu, W. Wang, H. Sun, A. Rocha and H. Li, "Robust Domain Misinformation Detection via Multi-Modal Feature Alignment," in *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 793-806, 2024, doi: 10.1109/TIFS.2023.3326368.

[80] H. Lin, J. Ma, M. Cheng, Z. Yang, L. Chen, and G. Chen, "Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10035–10047, Jan. 2021, doi: 10.18653/v1/2021.emnlp-main.786.

[81] Y. Liang, T. Tohti, and A. Hamdulla, "False information detection via multimodal feature fusion and Multi-Classifier Hybrid Prediction," *Algorithms*, vol. 15, no. 4, p. 119, Mar. 2022, doi: 10.3390/a15040119.

[82] H. Lin, Z. Luo, B. Wang, R. Yang, and J. Ma, "GOAT-Bench : Safety Insights to Large Multimodal Models through Meme-Based Social Abuse," *ACM Transactions on Intelligent Systems and Technology*, vol. 17, no. 4, pp. 1–25, Apr. 2025, doi: 10.1145/3729239.

[83] H. Lin, Z. Luo, J. Ma, and L. Chen, "Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9114–9128, 2023, doi: 10.18653/v1/2023.findings-emnlp.611.

[84] T. M. Ababu, M. M. Woldeyohannis, and E. B. Getaneh, "Bilingual hate speech detection on social media: Amharic and Afaan Oromo," *Journal of Big Data*, vol. 12, no. 1, Feb. 2025, doi: 10.1186/s40537-024-01044-y.

[85] G. O. Ganfure, "Comparative analysis of deep learning based Afaan Oromo hate speech detection," *Journal of Big Data*, vol. 9, no. 1, Jun. 2022, doi: 10.1186/s40537-022-00628-w.

[86] F. Gereme, W. Zhu, T. Ayall, and D. Alemu, "Combating fake news in 'Low-Resource' languages: amharic fake news detection accompanied by resource crafting," *Information*, vol. 12, no. 1, p. 20, Jan. 2021, doi: 10.3390/info12010020.

[87] H. Lin, J. Ma, L. Chen, Z. Yang, M. Cheng, and C. Guang, "Detect rumors in microblog posts for Low-Resource domains via adversarial contrastive learning," *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2543–2556, Jan. 2022, doi: 10.18653/v1/2022.findings-naacl.194.

[88] N. K, H. Sabaha, S. Rajiakodi, and B. Sivagnanam, "Detecting Homophobic and Transphobic Comments on Social media in Malayalam and English Languages," *Procedia Computer Science*, vol. 258, pp. 2479–2489, Jan. 2025, doi: 10.1016/j.procs.2025.04.510.

[89] H. S. Alatawi, A. M. Alhothali and K. M. Moria, "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT," in *IEEE Access*, vol. 9, pp. 106363-106374, 2021, doi: 10.1109/ACCESS.2021.3100435.

[90] N. Dufour *et al.*, "AMMEBA: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild," *arXiv.org*, May 19, 2024. https://arxiv.org/abs/2405.11697

[91] E. Broda and J. Strömbäck, "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review," *Annals of the International Communication Association*, vol. 48, no. 2, pp. 139–166, Mar. 2024, doi: 10.1080/23808985.2024.2323736.

[92] A. Sandu, I. Ioanăș, C. Delcea, L.-M. Geantă, and L.-A. Cotfas, "Mapping the Landscape of Misinformation Detection: A Bibliometric approach," *Information*, vol. 15, no. 1, p. 60, Jan. 2024, doi: 10.3390/info15010060.

[93] N. Navarro-Sierra, S. Magro-Vela, and R. Vinader-Segura, "Research on Disinformation in Academic Studies: Perspectives through a Bibliometric Analysis," *Publications*, vol. 12, no. 2, p. 14, May 2024, doi: 10.3390/publications12020014.

[94] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," *Multimedia Tools & Applications*, vol. 83, no. 17, pp. 51009–51067, Nov. 2023, doi: 10.1007/s11042-023-17470-8.

[95] M. Asif, M. Al-Razgan, Y. A. Ali, and L. Yunrong, "Graph convolution networks for social media trolls detection use deep feature extraction," *Journal of Cloud Computing Advances Systems and Applications*, vol. 13, no. 1, Feb. 2024, doi: 10.1186/s13677-024-00600-4.

[96] R. M. K. Saeed, S. Rady, and T. F. Gharib, "An ensemble approach for spam detection in Arabic opinion texts," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 1, pp. 1407–1416, Oct. 2019, doi: 10.1016/j.jksuci.2019.10.002.

[97] Q. Xu *et al.*, "M3A: A multimodal misinformation dataset for media authenticity analysis," *Computer Vision and Image Understanding*, vol. 249, p. 104205, Oct. 2024, doi: 10.1016/j.cviu.2024.104205.

[98] V. Kishore and M. Kumar, "Enhanced Multimodal Fake News Detection with Optimal Feature Fusion and Modified Bi-LSTM Architecture," *Cybernetics & Systems*, vol. 56, no. 6, pp. 684–714, Feb. 2023, doi: 10.1080/01969722.2023.2175155.

[99] A. Saeed and E. A. Solami, "Fake news detection using machine learning and deep learning methods," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 77, no. 2, pp. 2079–2096, Jan. 2023, doi: 10.32604/cmc.2023.030551.

[100] J. Lv, Y. Gao, L. Li, L. Shi, and S. Li, "Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 37, no. 9, Nov. 2025, doi: 10.1007/s44443-025-00317-7.

[101] W. Cui and M. Shang, "MIGCL: Fake news detection with multimodal interaction and graph contrastive learning networks," *Applied Intelligence*, vol. 55, no. 1, Dec. 2024, doi: 10.1007/s10489-024-05883-3.

[102] M. A. Jigar, A. A. Ayele, S. M. Yimam, and C. Biemann, "Detecting hate speech in Amharic using multimodal analysis of social media memes," *ACL Anthology*, May 01, 2024. https://aclanthology.org/2024.trac-1.10/

[103] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 6, Sep. 2021, doi: 10.1007/s00138-021-01249-8.

[104] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A comprehensive survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and applications," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 80, no. 1, pp. 1–35, Jan. 2024, doi: 10.32604/cmc.2024.053204.

[105] S. Hangloo and B. Arora, "Multimodal fusion techniques: Review, data representation, information fusion, and application areas," *Neurocomputing*, vol. 649, p. 130827, Jun. 2025, doi: 10.1016/j.neucom.2025.130827.

[106] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Multi-Model Fusion Framework using Deep Learning for Visual-Textual Sentiment Classification," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 76, no. 2, pp. 2145–2177, Jan. 2023, doi: 10.32604/cmc.2023.040997.

[107] M. Velmala, S. Rajiakodi, K. Pannerselvam, and B. Sivagnanam, "Multimodal Sentiment Analysis of Online Memes: Integrating text and image features for enhanced classification," *Procedia Computer Science*, vol. 258, pp. 355–364, Jan. 2025, doi: 10.1016/j.procs.2025.04.272.

[108] S. K. Hamed, M. Juzaiddin Ab Aziz and M. Ridzwan Yaakub, "Improving Data Fusion for Fake News Detection: A Hybrid Fusion Approach for Unimodal and Multimodal Data," in *IEEE Access*, vol. 12, pp. 112412-112425, 2024, doi: 10.1109/ACCESS.2024.3443092.

[109] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, Feb. 2024, doi: 10.1145/3649447.

[110] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Multi-Model Fusion Framework using Deep Learning for Visual-Textual Sentiment Classification," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 76, no. 2, pp. 2145–2177, Jan. 2023, doi: 10.32604/cmc.2023.040997.

[111] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *Npj Digital Medicine*, vol. 3, no. 1, p. 136, Oct. 2020, doi: 10.1038/s41746-020-00341-z.

[112] V. A. T. Caceres, K. Duffaut, A. Yazidi, F. Westad, and Y. B. Johansen, "Automated well log depth matching: Late fusion multimodal deep learning," *Geophysical Prospecting*, vol. 72, no. 1, pp. 155–182, Apr. 2022, doi: 10.1111/1365-2478.13200.