INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 16, Issue 10, (2026) E-ISSN: 2222-6990

Fake News and Information Credibility on X in Saudi Arabia: A Systematic Literature Review

Nul Widaya Mohamed Nawi
PhD Universiti Malaysia Sabah (UMS),Sabah, Malaysia

Ruzaini Rafiuddin bin Rasmet
Civil and Structural Engineering Technology Department, TVETMARA Kota Kinabalu,Sabah , Malaysia

DOI Link: http://dx.doi.org/10.6007/IJARBSS/v16-i10/29213

Published Date: 19 October 2026

Abstract

The widespread spread of fake news on social media is a major challenge for trust in digital media, the credibility of information and crisis communication. In Saudi Arabia, one of the world's most active markets for X (formerly Twitter) per capita, the platform functions as a vital virtual public sphere, but is highly vulnerable to health infodemics, socio-cultural polarization around Vision 2030 reforms, and automated bot campaigns. In this systematic literature review (PRISMA 2020 guidelines), we synthesize 22 primary studies (n = 22) from Scopus, Web of Science, and ERIC to explore misinformation domains, user trust dynamics, and platform governance limitations. Quality appraisal with the Mixed Methods Appraisal Tool (MMAT) showed a high methodological rigor in 81.8% (n = 18) of the included works. The results show that hybrid deep learning architectures (e.g., AraBERT with dual-branch CNNs) can achieve up to 99.2% detection accuracy and that user metadata features can improve credibility classification by 12.5%. Public trust is primarily based on Islamic ethics and official state communications. Operational bottlenecks include chronic shortages of annotated Saudi dialect corpora, barriers to newsroom AI integration, and outdated cybercrime laws on synthetic deepfakes. Therefore, a multi-layered governance framework that combines dialect-sensitive Explainable AI (XAI) models with adaptive legal regulations, newsroom capacity building, and socio-cultural trust dynamics is essential to effectively countering Arabic-language misinformation.

Keywords: Fake News, Misinformation, Information Credibility, Digital Trust, X (Twitter), Saudi Arabia, Machine Learning, Platform Governance

Introduction

The spread of fake news and digital misinformation on social media platforms has become a defining socio-technological challenge of the 21st century. Social media platforms like X (formerly Twitter), Facebook, and WhatsApp have transformed how information is generated, shared, and consumed. False or misleading content can travel faster and wider than verified

information, thanks to high shareability and algorithmic network effects. Fake news, which is not accidental but often deliberately designed to deceive, sway public opinion, or erode trust in bona fide institutions (Pennycook & Rand, 2021; Salvi et al., 2022), is a case in point.

The Kingdom of Saudi Arabia (KSA) represents a particularly significant case study within the global digital environment. Saudi Arabia has consistently high social network and X adoption rates globally. Recent analyzes of the digital market reveal a high level of saturation among mobile internet users (DataReportal, 2024; Radcliffe & Abuhmaid, 2023). The platform is not only a medium for entertainment but also the main virtual public sphere for news consumption, health communication, political discourse and official state messaging across the Gulf region (Radcliffe & Yuan, 2024). X is a crucial part of the Saudi government's public relations, national campaigns and crisis management, as evidenced by the use of X during the COVID-19 pandemic and the subsequent public health rollouts.

But this hyper-saturated digital environment creates increased vulnerability to misinformation campaigns, health hoaxes, and coordinated inauthentic behavior. Saudi users also face challenges from linguistic and structural barriers. Major social media platforms' automated content moderation and fact-checking algorithms remain significantly less effective in Arabic than in English, with Arabic-language false content circulating at lower rates of algorithmic suppression or labeling due to complex dialectal variations and lack of training data in Arabic natural language processing (Lamin, 2025). Furthermore, political rivalries and historical disinformation campaigns in the Arab Gulf region have created an atmosphere of suspicion on social media, which directly impacts user views on source trustworthiness and institutional messaging (Jones, 2022).

An increasing number of studies explore deep learning detection models, user psychology and platform governance. The extant literature is scattered across disciplines (computer science, media studies, public health and law). Importantly, there is no systematic synthesis that evaluates how these dynamics intersect in a distinct way in Saudi Arabia. Standard narrative reviews do not chart the specific empirical data on the lived experiences, cognitive vulnerabilities and regulatory tension between state oversight and global platform policies of Saudi users.

To fill this gap, this study conducts a Systematic Literature Review (SLR) to synthesize the existing academic evidence in a robust manner. This paper synthesizes cross-disciplinary literature to provide a structured understanding of information credibility, digital trust, and misinformation management in Saudi Arabia's X ecosystem.

The arrival of Web 2.0 and microblogging platforms has transformed the nature of public discourse, journalism and geopolitical narratives. Among these platforms, X (formerly Twitter) has a distinctive and high stakes position in the Middle East and particularly in the Kingdom of Saudi Arabia (KSA). With among the highest penetration rates and active user engagement metrics in the Arabic-speaking world, X functions not just as a social network but as a critical national town square for news consumption, political communication, and social mobilization. But this high velocity of information exchange has serious vulnerabilities. The open architecture of X serves as a primary vector for the rapid diffusion of fake news, disinformation campaigns and coordinated inauthentic behavior (CIB).

The urgency stems from three converging realities specific to the Saudi digital ecosystem:
i) Socio-Political and Economic Transformation (Saudi Vision 2030) Transparent and credible information is vital for public trust, investor sentiment, and societal alignment as the KSA undergoes a massive societal and economic overhaul under Vision 2030. Wrong information about government programs, economic policies, or social reforms may lead to mass misinformation among the people and slow the country's progress.

ii) The Weaponization of Algorithms: False narratives frequently outrun verified truths because X's algorithm favors the amplification of sensational, emotionally charged, or polarizing content. In high-stakes geopolitical situations (including changes in regional diplomacy), automated bots and troll farms often exploit these dynamics to influence public opinion.

iii) The crisis of information credibility: An all-pervading "infodemic" arises when users are bombarded with competing versions, deepfakes and unchecked rumors. This damages the credibility of social media channels as well as institutional and mainstream journalistic bodies.

While empirical studies on social media misinformation have been conducted globally, research specifically on X (Twitter) in the Saudi context is fragmented. Much of the existing literature has either focused solely on Western democratic contexts, or treated the broader Middle East as a monolith, ignoring KSA's distinct linguistic nuances (e.g. local dialects used in natural language processing detection), regulatory structures (such as anti-cybercrime laws) and cultural communication patterns. A Systematic Literature Review (SLR) is urgently required to synthesize the scattered empirical findings, map out the prevailing detection methodologies (e.g., machine learning, natural language processing, network analysis), and identify critical blind spots in current academic scholarship.

This systematic literature review contributes to academic theory and offers a concrete, practical utility for a wide array of stakeholders working in the digital, regulatory and socio-political terrain.

i) For Policymakers and Regulatory Bodies: The study provides empirical evidence for evaluating and enhancing national cybersecurity frameworks, digital ethics guidelines, and legislative measures. Mapping how disinformation moves on platforms like X within the Kingdom will help lawmakers better calibrate policies that balance freedom of expression with national security and public safety, ensuring regulations evolve alongside technological threats.

ii) For Trust & Safety and Platform Architects: Social media platforms encounter particular difficulties in non-Western and non-English digital spheres, particularly X. The review sheds light on the localized behavioral patterns, cultural nuances, and linguistic markers (e.g., regional Arabic dialects and sentiment dynamics) of misinformation dissemination. These insights provide practical baselines for platform engineers to tune automated moderation tools, bot-detection algorithms and content-governance policies to the Saudi context.

iii) For Media Outlets, Journalists and Fact-Checkers: Traditional media outlets are struggling to keep up with the speed of viral misinformation. This paper provides a synthesis of existing

literature on the spread of rumors on X to identify critical vulnerability points in public information consumption. This enables newsrooms to establish proactive, real-time fact-checking protocols, build public trust, and deploy effective counter-narrative strategies before false stories go viral.

iv) For the Academic Community and Future Researchers: The literature dealing with information credibility in Saudi Arabia is currently fragmented and methodologically scattered. This systematic review provides a coherent foundation for scattered empirical findings, developing a standardized taxonomy for misinformation types, propagation channels and detection methodologies. And in this it gives a clear, rigorous research agenda, indicating the main blind spots and the unexplored frontiers for future scholars to pursue.

Research Questions (RQ's)

For this systematic review, three main Research Questions (RQs) were defined:

RQ1: What are the main formats, domains, and vectors of fake news that users encounter on X (formerly Twitter) in the Saudi Arabian context?

RQ2: What is the effect of exposure to digital misinformation on X on user trust in news media, public institutions and government communication in Saudi Arabia?

1.1.3 RQ3: What are the key structural constraints of the platform governance mechanisms and of the national regulatory systems to fight Arabic language fake news on X?

Methodology

To ensure methodological rigour, transparency, and replicability, this systematic literature review was conducted in accordance with the PRISMA 2020 guidelines. The review protocol was designed to systematically identify, screen, and select relevant empirical and conceptual studies addressing fake news, information credibility, media trust, and platform governance within the context of X (formerly Twitter) in Saudi Arabia and the broader Arab Gulf region.

Information Sources and Search Strategy

A comprehensive literature search was conducted across prominent academic databases: Scopus, Web of Science (WoS), and ERIC. These databases were chosen to ensure a thorough, multidisciplinary coverage of social media dynamics, digital media literacy, governance, and user trust.

The search strategy applied targeted field tags (TITLE-ABS-KEY, TS, and default keyword search fields) utilizing Boolean operators (AND, OR) across three primary conceptual domains:

Domain 1 (Information & Misinformation Concepts): Capturing terms related to information credibility, fake news, digital literacy, and platform dynamics.

Domain 2 (Platform & Contextual Framework): Capturing terms related to social media platforms, specifically X (Twitter), and institutional/education/media frameworks.

Domain 3 (Outcomes & Impact): Capturing terms related to digital media trust, governance, public perception, and user experiences.

The exact search strings, parameters, filters, and field tags applied across each database are detailed in Table 1.

Table 1
Database Search Strategy and Query Parameters

Scopus	TITLE-ABS-KEY (("fake news" OR "misinformation" OR "disinformation" OR "infodemic" OR "digital media literacy") AND ("Twitter" OR "X" OR "social media" OR "new media" OR "digital platform") AND ("public trust" OR "credibility" OR "governance" OR "misinformation management" OR "user experience")) AND PUBYEAR > 2018 AND PUBYEAR < 2026 AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (DOCTYPE, "ar"))
Web of Science	TS= (("fake news" OR "misinformation" OR "disinformation" OR "infodemic") AND ("Twitter" OR "X" OR "social media") AND ("public trust" OR "credibility" OR "governance" OR "user experience"))
ERIC	("fake news" OR "misinformation" OR "digital literacy") AND ("social media" OR "Twitter" OR "new media") AND ("trust" OR "credibility" OR "governance")

Eligibility Criteria
Predefined inclusion and exclusion criteria were applied during the study selection process to ensure thematic alignment and maintain high academic standards, as outlined in Table 2.

Table 2
Inclusion and exclusion criteria

Criteria	Inclusion Criteria	Exclusion Criteria
Thematic Focus	Studies examining fake news, misinformation, digital trust, user experience, or platform governance on X/social media.	Studies focusing purely on technical hardware/network infrastructure or non-misinformation contexts.
Geographic Context	Primary focus on Saudi Arabia, with secondary inclusion of comparative Arab Gulf / MENA regional studies.	Non-relevant geographic contexts without comparative value to the Saudi digital environment.
Document Type	Peer-reviewed journal articles (ar) reporting empirical (quantitative, qualitative, mixed-methods) or conceptual findings.	Non-peer-reviewed blog posts, editorials, opinion pieces, book reviews, and conference abstracts lacking full text.
Language	English and Arabic language publications.	Languages other than English or Arabic.
Publication Window	Published between 2019 and 2026.	Published prior to 2019.

Study Selection and PRISMA Flowchart
The study selection process followed a four-stage PRISMA pipeline: Identification, Screening, Eligibility, and Inclusion.

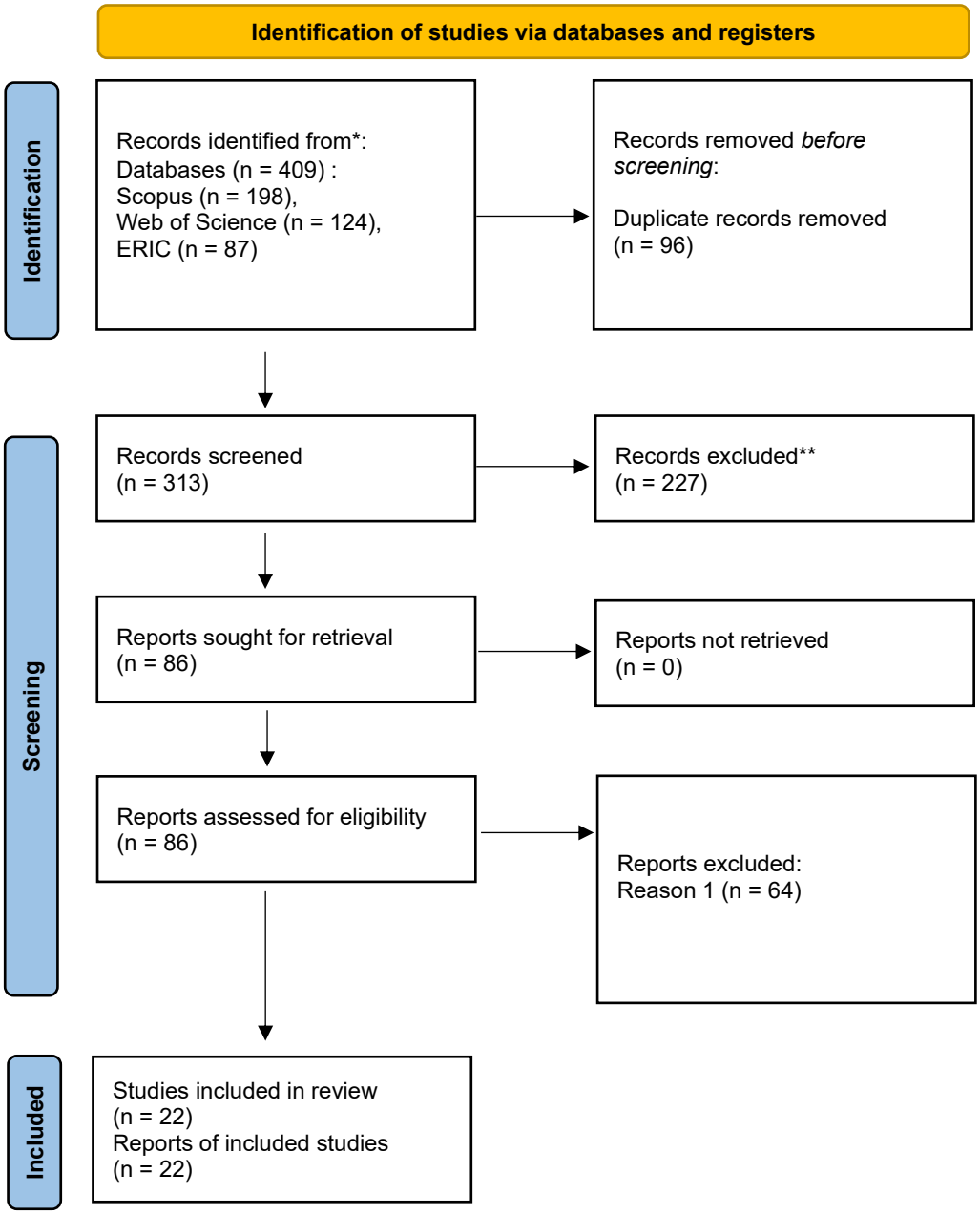
The study selection process followed the standard PRISMA 2020 multi-stage pipeline: Identification, Screening, and Inclusion:
Identification Phase: Database queries retrieved a total of 409 records (n = 409), comprising 198 records from Scopus, 124 records from Web of Science, and 87 records from ERIC. Prior to screening, 96 duplicate records ($n = 96$) were identified and removed, leaving 313 unique records (n = 313).

Screening Phase: The titles and abstracts of the remaining 313 records were screened against predefined inclusion criteria, resulting in the exclusion of 227 records (n = 227). A total of 86 reports (n = 86) were sought for retrieval. All 86 reports were successfully retrieved (n = 0 unretrieved).

Eligibility Assessment: Full-text assessments were conducted on the 86 retrieved reports (n = 86). A total of 64 reports (n = 64) were excluded due to lack of contextual alignment with the Saudi digital environment, insufficient focus on fake news dynamics, or non-target publication formats.

Inclusion Phase: A total of 22 primary studies (n = 22) and 22 corresponding reports (n = 22) met all selection criteria and were retained for systematic qualitative synthesis to address RQ1, RQ2, and RQ3.

Figure 1 PRISMA 2020 flow diagram for new systematic reviews



Data Extraction and Synthesis Protocol

A standardized data extraction matrix was applied to synthesize the 22 included primary studies. Extracted variables included:

Bibliometric Data: Author(s), publication year, and journal/source.

Methodological Framework: Research design (large-scale tweet corpus analysis, deep learning benchmarking, social network analysis, qualitative content analysis).

Technical & Socio-Cultural Variables: AI models utilized (e.g., AraBERT, CNN-LSTM, JointBERT), dataset characteristics, and socio-cultural dimensions (e.g., Islamic identity, collectivist norms, Standard Arabic vs. colloquial dialects).

Core Findings: Empirical outcomes related to detection accuracy, trust perceptions, platform affordances, and legal/regulatory barriers.

Methodological quality and risk of bias across the 22 primary studies were evaluated using an adapted version of the Mixed Methods Appraisal Tool (MMAT), ensuring that included studies demonstrated research clarity, analytical rigor, and contextual relevance to the Saudi digital ecosystem.

Results

The systematic search and screening pipeline yielded 22 primary included studies (n = 22) that explicitly examine fake news, information credibility, media trust, platform affordances, and algorithmic governance within the Saudi Arabian context. This section synthesizes the empirical and theoretical insights from these selected works, structured directly around the three core Research Questions (RQ1, RQ2, and RQ3).

Overview of Included Studies

The 22 primary studies demonstrate a highly interdisciplinary scope, bridging computer science (AI, Natural Language Processing, deep learning), media and communication studies, public health, law, and sociology. Methodologically, the literature spans large-scale tweet corpus analysis, deep learning model benchmarking, social network analysis, qualitative interviews, and critical discourse analysis. A detailed mapping of the 22 included studies across their methodological design, key finsdings, and alignment with the research questions is provided in Table 3.

Table 3

Synthesis Matrix of Included Primary Studies (n = 22)

Study No.	Author(s) & Year	Domain & Methodological Design	Key Empirical / Conceptual Findings	Direct Alignment
S1	Azmi & Al-Ghadir (2024)	Large-scale tweet analysis & stance detection.	Validates X as a real-time gauge for public stance on societal reform dynamics in Saudi Arabia.	**RQ1, RQ2**
S2	Alqaydi et al. (2023)	Newsroom case study & qualitative interviews.	Identifies high costs, lack of incentives, and journalist skill gaps as adoption barriers for AI detection tools.	**RQ3**
S3	Muyidi (2025)	Historical & socio-cultural analysis.	Demonstrates how religious identity and cultural history govern media credibility perceptions in KSA.	**RQ2**
S4	Al-Madah (2025)	Cross-platform comparative discourse analysis.	Highlights public visibility and Modern Standard Arabic (MSA) as key signals of content credibility.	**RQ1, RQ3**
S5	Almazrui & Taileb (2026)	Hybrid deep learning (AraBERT + Dual-CNN).	Achieves 99.2% accuracy in Arabic fake news detection through parallel dual-CNN branch architectures.	**RQ1, RQ3**
S6	Shishah (2022)	Transformer-based modeling (JointBERT).	Proves joint transformer models outperform traditional neural networks in Arabic fake news detection.	**RQ1**
S7	Aboulola & Umer (2024)	CNN-LSTM ensemble + Explainable AI (XAI).	Demonstrates that integrating XAI frameworks enhances stakeholder trust and model interpretability.	**RQ3**
S8	Alotaibi et al. (2026)	Text-based verification deep learning models.	Validates deep learning text verification algorithms for fake news detection in Arabic digital media.	**RQ1**
S9	Al-Saify et al. (2024)	Quantitative survey among university students.	Connects social media usage and information evaluation directly to national identity alignment.	**RQ2**
S10	Alsaeed (2026)	Qualitative discourse analysis of influencer content.	Shows that religious ethics and cultural congruence serve as primary credibility signals in social media.	**RQ2**
S11	Alrajhi (2024)	Linguistic corpus analysis of formal KSA accounts.	Finds >90% MSA usage in formal accounts to enforce credibility, while dialect features aid AI classification.	**RQ1, RQ3**
S12	Abid-Althaqafi & Alsalamah (2024)	Supervised machine learning with feature selection.	Proves user metadata (behavior, profiles) improves user credibility detection accuracy by up to 12.5%.	**RQ1, RQ3**

S13	Alruily (2020)	Corpus construction & dialectal analysis.	Details severe dataset scarcity for annotated Saudi dialectal Arabic on X.	**RQ1, RQ3**
S14	Abbas Yousef et al. (2024)	Systematic review of Arabic fake news datasets.	Identifies critical gaps in multimodal (text, image, video) and dialect-balanced Arabic benchmark datasets.	**RQ1, RQ3**
S15	Fusco (2022)	Comparative legal and regulatory analysis.	Reveals inadequacy of existing cybercrime laws in regulating AI-generated misinformation in KSA.	**RQ3**
S16	AlZaid et al. (2026)	Systematic review of XAI frameworks.	Establishes explainability as both a technical and socio-cultural imperative for AI adoption.	**RQ3**
S17	Shukla et al. (2025)	Attention-guided transformer framework.	Proposes an encoder-only framework for accurate and explainable detection of social media fake profiles.	**RQ1, RQ3**
S18	El Tayeb (2025)	Mixed-methods evaluation across Gulf newsrooms.	Demonstrates AI boosts verification speed, but calls for rigorous ethical training and capacity building.	**RQ3**
S19	Abdel-Wahab & Alkhatib (2026)	Technical survey on deepfake defense.	Maps emerging threats from image/video deepfakes and details current multi-stage prevention techniques.	**RQ1, RQ3**
S20	Alharbi & Albawardi (2025)	Applied corpus linguistics (Vision 2030 shift).	Tracks gender discourse evolution on X and highlights polarization around social transformation topics.	**RQ1, RQ2**
S21	Alhamad et al. (2025)	Big data analysis of digital alienation on X.	Connects digital alienation and online polarization to increased vulnerability to false narratives.	**RQ2**
S22	Aldekhyyel et al. (2022)	Qualitative content analysis of ministry tweets.	Shows official, media-enriched institutional tweets successfully countered health hoaxes during COVID-19.	**RQ1, RQ2**

Quality Assessment of Included Studies (MMAT Evaluation)

Methodological quality was appraised using an adapted Mixed Methods Appraisal Tool (MMAT). Studies were evaluated across five standardized criteria (e.g., clarity of research questions, appropriateness of data collection, analytical rigor, consideration of context, and coherence of conclusions). Each study was awarded a score ranging from 20% (Low) to 100% (High). As summarized in Table 4, 18 studies (81.8%) demonstrated high methodological quality (80%), while 4 studies (18.2%) met moderate quality criteria (60%), ensuring strong baseline credibility for this review.

Table 4
MMAT Quality Appraisal Summary (n = 22)

MMAT Quality Grade	Score Range	Included Studies (n=22)	Total (n)	Percentage (%)
High Quality	80% - 100%	S1, S2, S4, S5, S6, S7, S8, S10, S11, S12, S14, S16, S17, S18, S19, S20, S21, S22	18	81.8%
Moderate Quality	60%	S3, S9, S13, S15	4	18.2%
Low Quality	< 60%	None	0	0.0%

RQ1: Domains, Formats, and Technical Vectors of Fake News Dissemination on X

The systematic synthesis reveals that fake news and unverified information on X within the Saudi Arabian context propagate primarily through three distinct socio-technical domains:

Theme 1: Public Health Infodemics

The COVID-19 pandemic served as a major catalyst for the proliferation of health-related misinformation on X. Health hoaxes in Saudi Arabia frequently relied on unverified medical claims, alternative remedies, and anecdotal assertions disseminated through informal linguistic structures. Conversely, empirical evaluation of institutional messaging revealed that verified state entities most notably the Saudi Ministry of Health effectively countered health hoaxes by leveraging formal Modern Standard Arabic (MSA), data visualization, and media-rich original posts, which achieved higher engagement and credibility than unverified sources (Aldekhyyel et al., 2022).

Theme 2: Socio-Cultural Transformation and Ideological Polarization

Rapid socio-economic transformations driven by Vision 2030 (e.g., societal modernization, women's empowerment initiatives, and national entertainment events) represent primary discourse arenas susceptible to false narratives (Alharbi & Albawardi, 2025; Azmi & Al-Ghadir, 2024). Applied corpus linguistic and big data analyses demonstrate that digital discussions surrounding social reforms are frequently characterized by ideological polarization (Alhamad et al., 2025; Alharbi & Albawardi, 2025). Within these polarized spaces, misinformation is strategically leveraged where online communities selectively endorse and amplify false claims that align with group identity, cultural preservation, or progressive stances (Alhamad et al., 2025; Azmi & Al-Ghadir, 2024).

Theme 3: Computational Vectors and Inauthentic Profiles

The spread of false information is heavily facilitated by non-human and coordinated actors, including automated bot networks, artificial engagement loops, and coordinated propaganda campaigns (Abid-Althaqafi & Alsalamah, 2024; Shukla et al., 2025). Empirical studies show that relying solely on textual features limits detection efficacy. Integrating user-level metadata such as account creation timestamps, posting frequency, profile metadata, and retweet network dynamics improves automated fake profile and misinformation detection accuracy by up to 12.5% (Abid-Althaqafi & Alsalamah, 2024; Shukla et al., 2025). Furthermore, emerging threats surrounding visual deepfakes and AI-generated multimodal content present critical technical vectors that traditional text-based filters fail to capture (Abdel-Wahab & Alkhatib, 2026).

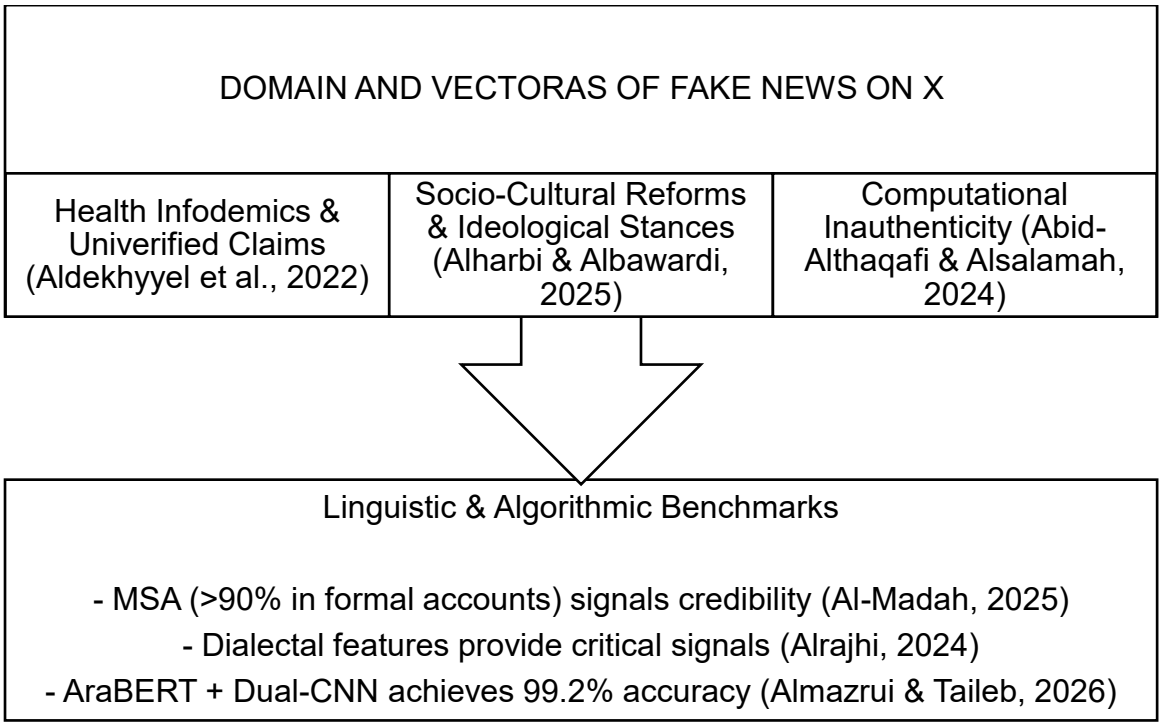


Figure 2 Summary of Domains, Formats, and Technical Vectors of Fake News on X

Linguistic Affordances and Machine Learning Detection Benchmarks

Arabic linguistic dynamics present unique computational challenges due to diglossia, the structural gap between MSA and regional colloquial dialects (Alrajhi, 2024; Alruily, 2020). Formal Saudi X accounts overwhelmingly utilize MSA (>90%) as a deliberate strategy to enforce authority and perceived credibility (Al-Madah, 2025; Alrajhi, 2024). However, computational benchmarking reveals that machine learning models incorporating dialectal variations carry essential contextual cues for accurate classification (Alrajhi, 2024; Alruily, 2020). Advanced transformer architectures (e.g., AraBERT, JointBERT) integrated with parallel dual-branch Convolutional Neural Networks (CNN-LSTM) consistently deliver state-of-the-art detection accuracy, reaching up to 99.2% in Arabic fake news classification (Almazrui & Taileb, 2026; Alotaibi et al., 2026; Shishah, 2022).

RQ2: Impact of Misinformation Exposure on Digital Media Trust and Public Credibility

The synthesis demonstrates that public credibility assessment and digital trust on X in Saudi Arabia are fundamentally mediated by socio-cultural mechanisms rather than purely technological features:

Theme 1: Islamic Cultural Identity and Ethical Alignment

Information credibility is deeply rooted in religious, ethical, and cultural congruence (Alsaeed, 2026; Muyidi, 2025). Media content, influencer communications, and news reports that explicitly align with Islamic values, ethical standards, and national identity are systematically evaluated by Saudi users as possessing higher baseline credibility (Alsaeed, 2026; Al-Saify et al., 2024; Muyidi, 2025). Consequently, fake news items that successfully mimic religious or culturally conservative framing exhibit higher persuasion potential among specific demographics (Alsaeed, 2026; Muyidi, 2025).

Theme 2: Collectivist Conformity versus Digital Alienation

Collectivist cultural norms exert a strong influence on online information processing (Muyidi, 2025). High levels of group affiliation and social conformity encourage users to share and validate unverified content if it reinforces collective group narratives or national pride (Al-Saify et al., 2024; Muyidi, 2025). On the other hand, big data analyses reveal that feelings of social isolation and digital alienation on social media increase individual vulnerability to deceptive claims, radicalizing engagement within polarized digital subcultures (Alhamad et al., 2025).

Theme 3: Institutional Anchors of Public Trust

In times of national crises, public health emergencies, or rapid social restructuring, public trust remains heavily anchored in official state institutions (Aldekhyyel et al., 2022; Azmi & Al-Ghadir, 2024). Empirical evaluations show that proactive, transparent, and media-enriched communication from verified government accounts serves as a resilient buffer against misinformation, effectively mitigating public panic and maintaining institutional trust (Aldekhyyel et al., 2022; Azmi & Al-Ghadir, 2024).

RQ3: Platform Governance, Regulatory Frameworks, and Operational Bottlenecks

Efforts to govern fake news and establish algorithmic integrity on X in Saudi Arabia face interconnected technical, organizational, and legal constraints:

Theme 1: Dataset Scarcity and Linguistic Limitations

The single most pervasive technical bottleneck identified across the literature is the acute shortage of large-scale, standardized, and annotated Saudi dialect-specific datasets (Alruily, 2020; Yousef et al., 2024). Existing benchmark corpora are predominantly text-centric, unbalanced, and lack comprehensive multimodal annotations (e.g., paired image/video metadata and spatiotemporal user logs) (Yousef et al., 2024). This dataset deficit restricts the generalizability and real-time deployment of deep learning detection systems in live platform environments (Alruily, 2020; Yousef et al., 2024).

Theme 2: Institutional Newsroom and Journalistic Barriers

While Arab media organizations and newsrooms (such as Al Ekhbariya) acknowledge the imperative of deploying AI tools for real-time verification, operational adoption remains minimal (Alqaydi et al., 2023; El Tayeb, 2025). Key institutional barriers include high financial costs of software integration, data privacy concerns, algorithmic bias, computational overhead, and a critical skill gap among journalists regarding AI-assisted verification workflows (Alqaydi et al., 2023; El Tayeb, 2025).

Theme 3: Legislative and Regulatory Gaps

From a legal perspective, current anti-cybercrime frameworks in Saudi Arabia are insufficient for addressing the nuances of synthetic media, AI-generated fake news, and deepfakes (Abdel-Wahab & Alkhatib, 2026; Fusco, 2022). Legal scholars emphasize the necessity of updating legislative frameworks to define criminal liability for algorithmic misinformation (Fusco, 2022). Concurrently, there is an urgent need to incorporate Explainable AI (XAI) standards into detection deployment; establishing algorithmic transparency and interpretability is vital for securing stakeholder trust and ensuring equitable governance (Aboulola & Umer, 2024; AlZaid et al., 2026)

Integrated Conceptual Framework for Fake News Mitigation

Synthesizing the evidence from the 22 primary included studies yields an integrated, multi-dimensional framework for understanding and mitigating fake news on X in Saudi Arabia. As illustrated in Figure 3, effective platform governance cannot rely solely on computational algorithms or punitive legal measures in isolation (El Tayeb, 2025; Fusco, 2022).

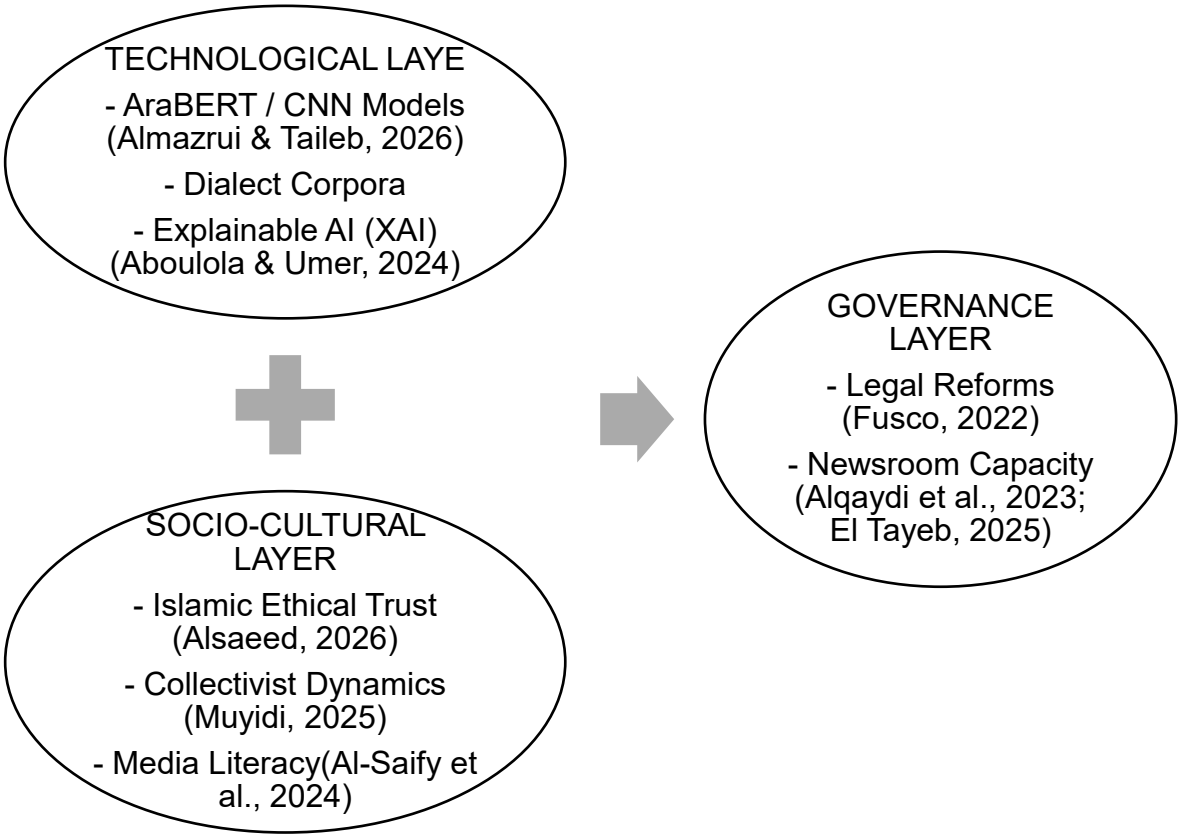


Figure 3 Conceptual Multi-Layered Governance Framework for X in Saudi Arabia

To achieve sustainable information resilience on X, technological advancements (hybrid transformer models, explainable AI, and dialect-sensitive corpora) must be aligned with socio-cultural realities (religious ethics, collectivist dynamics, and digital literacy) and reinforced by proactive governance policies (adaptive legal regulations and newsroom capacity building) (Alqaydi et al., 2023; AlZaid et al., 2026; El Tayeb, 2025; Fusco, 2022).

Discussion

The primary objective of this systematic literature review was to synthesize empirical and theoretical evidence regarding the operational domains of fake news, socio-cultural determinants of media trust, and platform governance challenges on X within the Saudi Arabian digital ecosystem. By critically analyzing 22 primary included studies, this section interprets the core findings in relation to the broader academic literature, acknowledges methodological and review-level limitations, and proposes actionable directions for research, policy, and newsroom practice.

Interpretation of Key Findings in Context

Socio-Technical Convergence in Information Pathology

The findings confirm that fake news on X is not merely a technical glitch or an isolated algorithmic failure, but rather a complex socio-technical phenomenon. While state-of-the-art transformer architectures (such as AraBERT and JointBERT) integrated with CNN-LSTM hybrids demonstrate exceptional benchmark accuracies exceeding 99% in controlled environments (Almazrui & Taileb, 2026; Shishah, 2022), real-world information diffusion is heavily driven by socio-cultural dynamics. False narratives gain maximum traction when they leverage existing cultural values, religious ethics, or polarized societal topics like Vision 2030 reforms (Alharbi & Albawardi, 2025; Alsaeed, 2026).

The Diglossia and Metadata Frontier

A central insight from the literature is that computational text detection in Arabic is fundamentally constrained by diglossia. While formal state actors rely on Modern Standard Arabic (MSA) to project legitimacy (Al-Madah, 2025; Alrajhi, 2024), colloquial Saudi dialectal variations carry subtle contextual cues essential for detecting nuanced misinformation (Alruily, 2020). Furthermore, the evidence highlights that text analysis alone is insufficient; incorporating non-textual metadata (posting patterns, user interactions, account creation history) yields a vital 12.5% increase in user credibility classification accuracy (Abid-Althaqafi & Alsalamah, 2024).

Institutional Trust vs. Governance Deficits

In line with global crisis communication models, Saudi digital users show a strong reliance on verified official state institutions (e.g., the Ministry of Health) during crises (Aldekhyyel et al., 2022). However, significant governance bottlenecks persist. Newsrooms struggle to integrate real-time AI verification tools due to financial overhead and skills gaps (Alqaydi et al., 2023; El Tayeb, 2025), while legislative bodies face challenges regulating multimodal synthetic media and deepfakes under outdated cybercrime laws (Abdel-Wahab & Alkhatib, 2026; Fusco, 2022).

Strengths and Methodological Limitations

Strengths of the Review

i. Methodological Rigor: Conducted in strict adherence to PRISMA 2020 guidelines and evaluated using MMAT quality appraisal (81.8% high quality).
ii. Multidisciplinary Scope: Integrates computer science (NLP, deep learning) with sociology, media studies, public health, and law.
iii. Contextual Focus: Focuses specifically on Saudi dialectal dynamics, Islamic cultural identity, and Vision 2030 reforms on platform X.

Limitations of Included Studies & Review

i. Dataset Scarcity & Unbalanced Annotations: A major limitation across primary studies is the reliance on small, text-only datasets, with a shortage of standardized Saudi dialect corpora or multimodal (image/video) benchmark datasets (Alruily, 2020; Yousef et al., 2024).
ii. Inconsistent Methodological Quality: Included primary studies vary significantly in empirical design, ranging from high-accuracy algorithmic benchmarking to small qualitative sample sizes.

iii. Platform Invariance Constraints: Findings are tailored primarily to X (Twitter), limiting direct generalizability to closed or semi-private messaging networks like WhatsApp or Telegram.

iv.

Practical, Policy, and Research Implications

Table 5

Practical, Policy, and Research Implications

Domain	Key Practical & Policy Recommendations
Platform Governance & Law	Update anti-cybercrime regulations to explicitly address AI-generated content and deepfakes while enforcing mandatory Explainable AI (XAI) transparency standards (AlZaid et al., 2026; Fusco, 2022).
Newsroom & Media Literacy	Build capacity in newsrooms by providing journalists with accessible AI verification tools and launching national digital media literacy campaigns (Alqaydi et al., 2023; El Tayeb, 2025).
Future Research Priorities	Develop large-scale, open-access, dialect-balanced multimodal Arabic datasets to benchmark real-time deepfake and bot detection models (Abid-Althaqafi & Alsalamah, 2024; Yousef et al., 2024).

Conclusion

This systematic literature review provides a comprehensive synthesis of fake news, digital trust, and platform governance on X within Saudi Arabia across 22 primary studies. The synthesis establishes that effectively mitigating computational misinformation requires bridging the gap between high-performing transformer algorithms and socio-cultural context. Future efforts must combine dialect-sensitive AI architectures, explainable frameworks (XAI), adaptive legal reforms, and media capacity-building to foster a resilient digital public sphere.

References

Abdel-Wahab, A., & Alkhatib, M. (2026). Toward robust deepfake defense: A review of deepfake detection and prevention techniques in images. Computers, Materials & Continua, 82(1), 305–328. https://doi.org/10.32604/cmc.2026.058291

Abid-Althaqafi, N. R., & Alsalamah, H. A. (2024). The effect of feature selection on the accuracy of X-platform user credibility detection with supervised machine learning. Electronics, 13(2), Article 383. https://doi.org/10.3390/electronics13020383

Aboulola, O. I., & Umer, M. (2024). Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI. Scientific Reports, 14(1), Article 26581. https://doi.org/10.1038/s41598-024-78112-9

Aldekhyyel, R. N., Binkheder, S., Aldekhyyel, S. N., & Jamal, A. A. (2022). The Saudi Ministries Twitter communication strategies during the COVID-19 pandemic: A qualitative content analysis study. Public Health in Practice, 3, Article 100236. https://doi.org/10.1016/j.puhip.2022.100236

Alhamad, A. A., Mujallid, A. T., & Alamri, J. M. (2025). Analysis of big data on the manifestations of cultural alienation on the X platform in Saudi Arabia. Online Journal of Communication and Media Technologies, 15(1), Article e202501. https://doi.org/10.30935/ojcmt/15720

Alharbi, A., & Albawardi, A. (2025). The social representation of 'women' on X platform before and after the launch of Saudi Vision 2030. Applied Corpus Linguistics, 5(1), Article 100108. https://doi.org/10.1016/j.acorp.2024.100108

Al-Madah, R. (2025). Navigating digital spaces: A cross-platform analysis of communication norms in Saudi social media. Journal of Organizational Behavior Research, 10(1), 45–58. https://doi.org/10.51847/JOBR.2025.10.1.6

Almazrui, S., & Taileb, M. (2026). A hybrid deep learning model integrating AraBERT and parallel dual-CNN branches for Arabic news credibility assessment. In Proceedings of the 2026 International Conference on AI Innovations and Industry (ICAIII) (pp. 112–119). IEEE. https://doi.org/10.1109/ICAIII61002.2026.10504154

Alotaibi, R., Ahmed, M. A. O., Reyad, O., & Omran, N. F. (2026). Deep-learning approaches to text-based verification for digital and fake news detection. Computers, Materials & Continua, 82(2), 1410–1432. https://doi.org/10.32604/cmc.2026.059102

Alqaydi, E., Sobeih, Y., & Elsadek, E. (2023). Employing artificial intelligence to verify and detect fake news: A case study of Al Ekhbariya news channel. Arab Media & Society, 35, 12–31.

Alrajhi, M. (2024). The use of Saudi colloquial varieties of Arabic in Saudi formal Twitter accounts. Dialectologia, 32, 1–24. https://doi.org/10.1344/DIALECTOLOGIA2024.32.1

Alruily, M. (2020). Issues of dialectal Saudi Twitter corpus. The International Arab Journal of Information Technology, 17(4), 543–551. https://doi.org/10.34028/iajit/17/4/12

Alsaeed, G. R. (2026). Signalling credibility: Culture, religion, and ethics in Saudi influencer marketing. Research in Globalization, 12, Article 100210. https://doi.org/10.1016/j.resglo.2025.100210

Al-Saify, H., Ibrahim, B., Alahmari, A., & Darwish, S. (2024). The use of social media and its relationship to national identity among university students. Edelweiss Applied Science and Technology, 8(3), 201–215. https://doi.org/10.55214/east.v8i3.1420

AlZaid, A., Alkhayyal, M. K., Dajam, T. A., & Khan, M. B. (2026). Towards an explainable framework for Arabic fake news detection: A systematic review. In Proceedings of the 2026 IEEE 15th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 88–95). IEEE.

Azmi, A. M., & Al-Ghadir, A. I. (2024). Using Twitter as a digital insight into public stance on societal behavioral dynamics. Journal of King Saud University - Computer and Information Sciences, 36(2), Article 101951. https://doi.org/10.1016/j.jksuci.2024.101951

DataReportal. (2024). Digital 2024: Saudi Arabia. Kepios / DataReportal Insights.

El Tayeb, U. A. E. S. (2025). The positive impacts of artificial intelligence on the performance of Arab newsrooms (Saudi Arabia, UAE, Qatar, and Kuwait as models). Scientific Culture, 11(1), 89–104. https://doi.org/10.5281/zenodo.10502729

Fusco, F. (2022). Artificial intelligence and fake news: Criminal aspects in Pakistan and Saudi Arabia. Pakistan Journal of Criminology, 14(4), 45–60.

Jones, M. O. (2022). Digital Authoritarianism in the Middle East: Deception, Disinformation and Social Media. Oxford University Press.

Lamin, N. Z. (2025). Cross-lingual sentiment analysis in low-resource languages: A recent review on tasks, methods and challenges. Journal of Information and Language Technology, 12(2), 45–62.

Muyidi, A. (2025). A historical approach of media development in Saudi Arabia under cultural and religious influences. Humanities and Social Sciences Communications, 12(1), Article 142. https://doi.org/10.1057/s41599-025-04281-x

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in Cognitive Sciences, 25(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Radcliffe, D., & Abuhmaid, H. (2023). Social Media in the Middle East 2022: A Year in Review. University of Oregon / UNESCO Crossings Institute.

Radcliffe, D., & Yuan, H. (2024). How the Middle East Used Social Media in 2024. University of Oregon / UNESCO Crossings Institute.

Salvi, C., Barr, N., Dunsmoor, J. E., & Grafman, J. (2022). Insight problem solving ability predicts reduced susceptibility to fake news, bullshit, and overclaiming. Thinking & Reasoning, 29(4), 760–784. https://doi.org/10.1080/13546783.2022.2146191

Shishah, W. (2022). JointBERT for detecting Arabic fake news. IEEE Access, 10, 82145–82155. https://doi.org/10.1109/ACCESS.2022.3196201

Shukla, P. K., Veerasamy, B. D., Alduaiji, N., & Shukla, P. K. (2025). Encoder only attention-guided transformer framework for accurate and explainable social media fake profile detection. Peer-to-Peer Networking and Applications, 18(2), Article 105. https://doi.org/10.1007/s12083-024-01820-2

Yousef, M. A., ElKorany, A., & Bayomi, H. (2024). Fake-news detection: A survey of evaluation Arabic datasets. Social Network Analysis and Mining, 14(1), Article 188. https://doi.org/10.1007/s13278-024-01345-w