A Bibliometric Mapping of Transformer-Based Misinformation Detection: Trends and Gaps

Borhan Ab Rahman, Mohd Zakree Ahmad Nazri, Mohd Ridzwan Yaakub
Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia

***Abstract*—The rapid spread of misinformation on social platforms has intensified research on automated detection, with Transformer-based architectures becoming a primary technical foundation. However, the volume and diversity of publications make it difficult to track the field's evolution. It is also challenging to identify the contributors, influential studies, and remaining gaps that shape its knowledge base. This study presents a bibliometric mapping of Transformer-based research on misinformation detection, based on Scopus records retrieved on 3rd March 2026 (n = 3,637). Performance indicators and science mapping analysis are used to profile publication growth, subject-area distribution, leading countries, institutions, and authors, citation impact, and thematic structure. Results show that conference proceedings (47.18%) and journals (41.93%) dominate dissemination, reflecting both rapid-cycle publishing and archival consolidation. Computer Science accounts for most publications (84.99%), while Social Sciences and Medicine contribute non-trivial shares, indicating applied and societal engagement. China, the United States, and India lead in national output, and institutional productivity is concentrated in a small set of research hubs. Citation indicators (h-index = 82; g-index = 154) suggest a broad influence with a distinct, highly cited core. Keyword mapping confirms a stable emphasis on NLP-driven detection in social media and shows recent growth in multimodal approaches and generative-AI-related topics. The findings provide an evidence-based overview of current trends and emerging directions, helping researchers position future work and prioritize underexplored problems.**

***Keywords*—Bibliometric analysis; misinformation detection; Transformer architectures; keyword co-occurrence; citation analysis; Scopus**

I. Introduction

Over the past decade, decentralized social media platforms have increased the reach and velocity of online communication. The same properties also increase exposure to misinformation by lowering moderation barriers and enabling rapid cross-community spread, including across languages. The consequences extend beyond online discourse and affect democratic decision-making, public health behavior, and social cohesion. These risks have intensified demand for automated detection systems that can support early identification and timely mitigation [1], [2]. In this study, misinformation is used as an umbrella term, while fake news and disinformation are treated as closely related labels reported in the computational literature.

Research in Natural Language Processing (NLP) has, therefore, moved from feature engineering and classical classifiers toward Transformer-based architectures for contextual representation learning. Early misinformation detection pipelines were typically built around handcrafted lexical or syntactic features with classical classifiers [3]. These approaches remain computationally efficient, but they often degrade under implicit claims, context-dependent semantics, and distribution shifts across topics, platforms or time [4]. Transformer architectures have changed the technical baseline by learning contextual representations through attention mechanisms [5]. Models such as BERT, RoBERTa and more recently, large language models (LLMs) have become standard choices for misinformation detection because they achieve strong performance on benchmark datasets and transfer effectively across varied text genres [6], [7]. The research frontier has also moved beyond text-only classification toward multimodal detection, explainability and robustness under real-world conditions [8], [9].

Despite the volume of technical work, the field lacks a precise macro-level account of Transformer-based misinformation detection as a distinct research stream. Existing reviews often treat misinformation or disinformation as a broad phenomenon and prioritize diffusion patterns, societal implications or general computational approaches [10], [11]. This leaves several important questions unresolved. It remains difficult to identify which venues and publications define the core knowledge base, how collaboration and influence are structured and which themes are consolidating versus emerging, especially given the recent surge of LLM-related research and the growing reliance on multimodal content. Without this structured view, researchers risk duplicating effort, selecting outdated baselines and evaluating systems in settings that do not reflect current threat models [12].

This study addresses that gap through a focused bibliometric mapping of Transformer-based architectures for misinformation detection. Using Scopus records retrieved on 3rd March 2026 and science mapping techniques, the analysis integrates performance indicators with network-based mapping of collaboration patterns, citation impact and thematic evolution [13], [14]. The study also synthesizes research gaps that follow directly from observed publication and topic trends, supporting clearer prioritization of future work.

A. Related Work

Bibliometric and survey studies have examined misinformation and disinformation research from broad perspectives, including publication growth, influential sources and topic structures. For example, Wang et al. [11] provide a bibliometric review of disinformation research, while Xu et al. [10] map evolutionary themes in misinformation dissemination on social media. Barve and Mulay [13] survey incremental learning for false-information text classification, reflecting continued interest in efficiency and continual adaptation.

These works provide useful coverage of the wider misinformation landscape, but they do not isolate Transformer-based architectures for misinformation detection as a distinct technical stream. Recent reviews often aggregate heterogeneous model families and settings, which can blur how Transformer-centric methods, multimodal pipelines and LLM-related topics are reshaping evaluation practices and research priorities [14]. A focused mapping that integrates performance indicators with science mapping is therefore needed to clarify the intellectual base, contributor structure and thematic evolution specific to Transformer-based misinformation detection.

This study addresses the following research questions. (RQ1): How has the publication output and citation impact of Transformer-based research on misinformation detection evolved? RQ2: Which articles and sources have the greatest impact in this domain? RQ3: Which countries and institutions contribute the most to the literature? RQ4: Who are the most productive authors and what collaboration patterns dominate the field? RQ5: What thematic clusters dominate the domain and how have they evolved based on keyword co-occurrence and overlay analysis? RQ6: How is the literature distributed across subject areas and what does this distribution suggest about interdisciplinarity? RQ7: What research gaps and future directions emerge from the observed trends in methods, applications and evaluation practices?

Section II details the Scopus data source and search strategy, the inclusion workflow and the procedures used for performance profiling and science mapping. Section III presents the results aligned with the research questions and Section IV summarizes contributions and future directions.

II. Materials and Methods

The data source, search strategy, inclusion criteria and analysis workflow used for this bibliometric study are described in this section. The differences between prior bibliometric reviews and this study are summarized in Table I and the retrieval and screening workflow is shown in Fig. 1.

TABLE I. Comparison of Previous Studies of Misinformation Detection and Our Study

Elements of comparison	[13]	[11]	[10]	**Our study**
Coverage year	2000–2019	2002–2021	2013–2023	2017 – 2026 (as on 03 March 2026)
Database	Scopus and Web of Science	Web of Science (SCI-E and SSCI)	Web of Science Core Database	Scopus
Keywords	Master keyword “incremental learning” with primary keywords including “false information” OR “misinformation” OR “text classification” etc. and secondary exclusion terms	disinformation OR misinformation OR “fake news” OR “infodemic” OR “information pollut*” OR deepfake* OR “rumor propagation”	TS query combining misinformation terms with social media/social network and diffusion verbs	Misinformation terms combined with Transformer terms
Focus of study	Maps incremental learning for text classification in false information detection and reports publication statistics, collaboration, network and citation analysis plus techniques used	Bibliometric analysis of disinformation research, identifying influential works and central topics using co-citation and keyword co-occurrence	Topic evolution of misinformation dissemination on social media; identifies hotspots and trends using a staged framework (community detection, TOPSIS, visualization)	Bibliometric mapping of Transformer-based architectures for misinformation detection: intellectual structure, dominant clusters, trend trajectory and research gaps
Methodology	Bibliometric survey with publication statistics, collaboration degree, collaboration visualization and citation analysis	Bibliometric analysis: performance analysis and science mapping	Scoping review framework combined with bibliometric analysis	Bibliometric analysis: performance analysis + science mapping
Protocol/Tools	Not explicitly tool-focused in the study, but the objectives include collaboration visualization and citation analysis.	Derwent Data Analyzer (DDA 10) for cleaning, mining and visualization	ITGinsight for preprocessing; integrates community detection, AHP, TOPSIS and visualization	VOSviewer v1.6.20 (mapping and visualization) and Harzing’s Publish or Perish (performance indicators)

Table I positions the present work relative to recent bibliometric studies by comparing their coverage periods, database selections, keyword strategies, analytical focus and toolchains. Prior studies [10], [11] and [13] vary in scope and objectives, including broad misinformation-dissemination mapping and incremental-learning-oriented surveys and they often rely on Web of Science or a combination of databases, with different preprocessing pipelines. In contrast, the current study targets a narrower technical stream by focusing specifically on Transformer-based architectures for misinformation detection. It adopts a consistent Scopus-only retrieval strategy covering 2017–2026 (as of 3 March 2026). Scopus was selected as the sole data source because it offers broad coverage of computer science and engineering venues and provides structured bibliographic metadata required for performance analysis and science mapping (e.g., author affiliations, source titles, references and citation counts) under a consistent indexing policy. Using a single database also reduces duplication and record-matching errors that can arise when merging records from multiple sources with different metadata schemas.

Nevertheless, Scopus-only retrieval may omit publications indexed exclusively elsewhere and may underrepresent some regional outlets. Accordingly, the results should be interpreted as representing the Scopus-indexed landscape during the retrieval period. Future work may triangulate Scopus with Web of Science or other databases to test the stability of the identified trends and networks.

The keyword formulation also differs by explicitly combining misinformation terms with Transformer-related terms, enabling the dataset to capture Transformer-centric model families and related evaluation settings rather than general diffusion or platform-level discourse. Methodologically, the study integrates performance analysis with science mapping, supported by the tool configuration reported in Table I and the retrieval workflow shown in Fig. 1.

A. Data Source and Search Scope

Records were retrieved from Scopus on 3rd March 2026 using the TITLE-ABS-KEY fields. The search covered all years, was restricted to English, and applied no restrictions on source type or document type. The workflow is summarized in Fig. 1. The query is intentionally inclusive, so the retrieved set may contain adjacent high-impact LLM-related work that shapes misinformation threat models even when it is not a detection-model paper.

B. Search Strategy and Query Design

The query combined (1) misinformation-related terms (e.g., misinformation, disinformation, fake news, rumor) with (2) Transformer-related terms (e.g., Transformer, attention mechanism, BERT, RoBERTa, LLM). The complete Boolean query is provided in Fig. 1.

C. Study Selection and Dataset Construction

The search returned 3,637 records. Screening was limited to database-level filters (TITLE-ABS-KEY and English). No post-retrieval exclusions were applied; therefore, 0 records were removed, and 3,637 documents were included (Fig. 1). To clarify the dataset scope, the 3,637 records are treated as a query-defined Scopus corpus rather than a manually screened systematic review corpus. The query was designed to capture both detection-oriented studies and related Transformer/LLM work that informs misinformation threat models and research trends. Therefore, no post-retrieval topical exclusion was applied, and the dataset should be interpreted as the Scopus-indexed landscape captured by the stated search strategy, not as a set limited only to detection-architecture papers.

D. Data Export and Preprocessing

All retrieved records were exported from Scopus, including available bibliographic metadata such as publication year, document title, authors, affiliations, source title, abstracts, author keywords, indexed keywords, cited references, and citation counts. Before analysis, the dataset was checked for basic consistency to reduce noise in network construction. This included harmonizing common variants of author names and keywords, where applicable, and consolidating obvious duplicates (for example, singular and plural forms or abbreviations representing the same concept). The cleaned files were then prepared in formats compatible with the selected tools.

E. Bibliometric Tools and Analysis Procedures

VOSviewer (v1.6.20) was used to construct and visualize co-authorship networks involving authors, institutions, and countries, as well as keyword co-occurrence and overlay maps. Citation indicators were examined using Harzing's Publish or Perish to support publication-level and source-level impact analysis.

F. Outcome Measures and Reporting

Results are reported using (1) performance indicators describing publication growth, major contributors, leading sources, and highly cited documents, and (2) science mapping outputs that visualize collaboration structure and thematic organization. The overall workflow is summarized in Fig. 1.

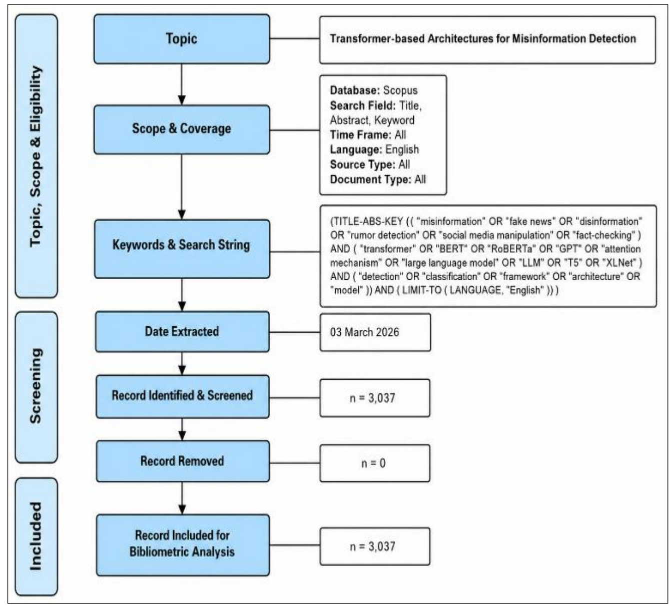


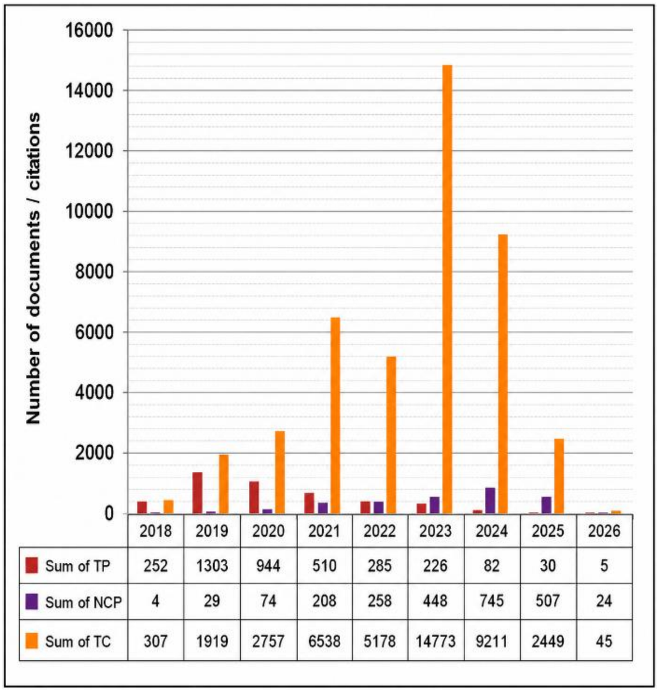
Fig. 1. Flowchart illustrating the literature search and data analysis process in the study.

III. Results and Discussion

A. Publication Output and Citation Trend

The temporal pattern of research activity and impact is described using three indicators: total publications (TP), number of cited publications (NCP), and total citations (TC). As shown in Fig. 2, publication output increases from 252 papers in 2018 to a peak of 1,303 papers in 2019, followed by 944 papers in 2020. Output then declines in subsequent years (510 in 2021, 285 in 2022, 226 in 2023, 82 in 2024, 30 in 2025 and 5 in 2026). The reduction in the most recent years is consistent with a shorter publication and indexing window, as the dataset was extracted in early 2026.

Indexing delay can explain part of the reduction in 2024–2026, but it does not fully account for the sustained decline between 2019 and 2023. A plausible interpretation is that the 2019 peak reflects an early surge of BERT-family adoption in misinformation detection, while subsequent work increasingly shifted toward adjacent framings (e.g., fact-checking, claim verification, stance and content authenticity) and reporting conventions that do not always foreground the same Transformer keywords in TITLE-ABS-KEY fields. Accordingly, the post-2019 decline should be interpreted as a query- and indexing-conditioned pattern within the retrieved Scopus dataset rather than a definitive indicator of reduced Transformer-based research activity.



Note: TP = Total Publications; NCP = Cited Publications; TC = Total Citations.

Fig. 2. Annual publication output and citation impact.

Citation indicators exhibit a lag relative to publication volume. TC rises from 307 (2018) and 1,919 (2019) to 2,757 (2020) and 6,538 (2021), remains high in 2022 (5,178) and peaks in 2023 (14,773). TC then decreases in 2024 (9,211), drops further in 2025 (2,449) and declines in 2026 (45), reflecting limited time for recent outputs to accumulate citations. A similar pattern is observed for NCP, which increases from 4 (2018) and 29 (2019) to 74 (2020) and 208 (2021), reaches 258 (2022) and 448 (2023), peaks in 2024 (745) and declines in 2025 (507) and 2026 (24).

Overall, the year-wise patterns indicate that publication volume is concentrated in earlier years within this dataset, while citation impact becomes more visible after a maturation period. This lag implies that citation-based influence should be interpreted using multi-year windows rather than recent-year counts alone, since newer publications have not had sufficient time to accumulate citations. In practical terms, the most recent declines are better viewed as recency and indexing effects than as evidence of reduced scholarly attention. This answers RQ1 by describing how publication output and citation impact evolve. Having established this temporal profile, the next subsection evaluates the impact more directly by identifying influential publications and sources.

B. Citation Impact of Publications and Sources

Citation impact is assessed at two levels: publication-level influence, using citation indicators and highly cited papers and source-level influence, indicating where influential work is disseminated and where recent attention is concentrated. Publication-level impact is summarized in Table II.

As shown in Table II, the dataset contains 3,637 papers and 43,207 citations, with an average of 11.88 citations per paper, an h-index of 82, and a g-index of 154. These indicators suggest broad influence with a highly cited core, which motivates a closer look at the publications driving this impact. The most cited publications are listed in Table III. Because the dataset follows a query-defined bibliometric scope, Table III is interpreted as showing the most-cited publications within the retrieved corpus rather than as a list restricted to detection-architecture papers. Some records are not detection papers in a narrow methodological sense, but they are retained because they represent influential Transformer- and LLM-related work that shapes current misinformation research, especially around generative AI, information integrity and emerging threat models.

TABLE II. CITATION METRICS

Metrics	Data
Papers	3637
Number of Citations	43207
Years	8
Citations per Year	5400.88
Citations per Paper	11.88
Cites_Author	11525.38
Papers_Author	1170.09
Authors_Paper	4.1
h_index	82
g_index	154

Table III should be read as reflecting the most influential publications within the retrieved scope, rather than as a list restricted to detection-architecture papers. It indicates two major citation drivers. Several highly cited items focus on LLMs and ChatGPT, reflecting strong cross-domain attention to generative AI and its implications for information integrity and high-stakes settings. A second stream comprises methodological contributions, including detection baselines and multimodal frameworks that are repeatedly cited as reference implementations or comparative anchors in subsequent studies. This split suggests that citation impact is shaped by both the rapid uptake of LLM-related discourse and the persistent reliance on foundational detection frameworks. Importantly, this pattern implies that influence in the domain is determined not only by advances in modelling but also by how research engages with emerging threat models and deployment-relevant concerns. As a result, studies that connect detection performance with robustness, transparency, or real-world risk contexts are more likely to achieve sustained citation visibility.

Source-level dissemination is examined in Fig. 3, which contrasts Total Publications (TP) with citations in 2025 across leading venues grouped by source type. While Table II and Table III provide a publication-level view of impact, it is also important to examine the sources/venues through which this research is disseminated.

TABLE III. TOP 10 CITED PUBLICATIONS

No.	Authors	Title	Cites	Cites per Year
1	Y.K. Dwivedi et al. (2023) [15]	So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy	3217	15.67
2	J. Clusmann et al. (2023) [16]	The future landscape of large language models in medicine	695	19
3	R.K. Kaliyar et al. (2021) [17]	FakeBERT: Fake news detection in social media with a BERT-based deep learning approach	673	11
4	S. Singhal et al. (2019) [18]	SpotFake: A multi-modal framework for fake news detection	553	10
5	L. De Angelis et al. (2023) [19]	ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health	501	12.17
6	A. Abd-alrazaq et al. (2023) [20]	Large Language Models in Medical Education: Opportunities, Challenges and Future Directions	438	7.44
7	J. Dempere et al. (2023) [21]	The impact of ChatGPT on higher education	343	9.83
8	S. Harrer (2023) [22]	Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine	334	13.5
9	A. Choudhury, H. Shamszare (2023) [23]	Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis	326	5.88
10	P. Qi, J. Cao, T. Yang, J. Guo, J. Li (2019) [24]	Exploiting multi-domain visual information for fake news detection	324	4.6

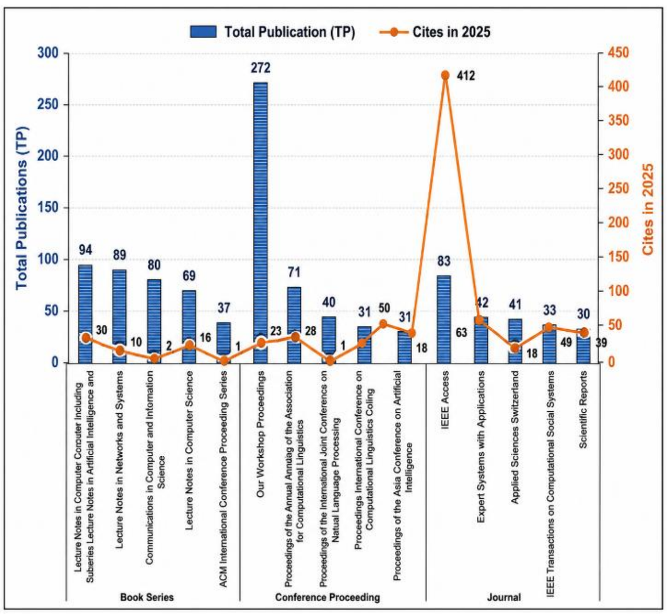


Fig. 3. Publication output (TP) and citations in 2025 across leading sources.

In Fig. 3, the highest-output venues differ by source type. Within Book Series, Lecture Notes in Computer Science (including LNAI/LNBI) leads output (TP = 94), followed by Lecture Notes in Networks and Systems (TP = 89), and Communications in Computer and Information Science (TP = 80). Within Conference Proceedings, CEUR Workshop Proceedings dominates (TP = 272), followed by the ACL Annual Meeting Proceedings (TP = 71). Within Journals, IEEE Access records the highest output (TP = 83), followed by Expert Systems with Applications (TP = 42), and Applied Sciences (Switzerland) (TP = 41). Citations in 2025 are used as a fixed one-year window to reduce the advantage of older publications and to highlight venues receiving current attention within the same citation period.

Taken together, the source-level profile separates dissemination capacity (TP) from recent visibility (2025 citations). This distinction is useful because high output does not necessarily imply high short-window citation uptake and it provides a venue-focused complement to publication-level impact indicators. Table II, Table III and Fig. 3, therefore, provide publication and venue-level evidence of impact. This answers RQ2 by identifying the most influential publications and the main sources through which impactful work is disseminated. A limitation is that citations in a single-year window and source-level counts are sensitive to indexing coverage and retrieval scope. Accordingly, these indicators should be interpreted as reflecting venue visibility within the Scopus-indexed records captured by the query and retrieval date.

C. Subject Area Distribution and Interdisciplinarity

The distribution of publications across Scopus subject areas is used to characterize the disciplinary base of Transformer-based misinformation detection research and to assess the degree of interdisciplinarity. Because Scopus can assign a single document to multiple subject areas, percentages may sum to more than 100 per cent; the overlap reflects multi-disciplinary indexing rather than duplication in the dataset.

As shown in Table IV, Computer Science accounts for 3,091 publications (84.99%), indicating that the field is primarily anchored in computing and NLP research. Engineering accounts for 1,018 publications (27.99%), followed by Mathematics with 796 (21.89%), aligning with the emphasis on model development, optimization and evaluation methodology. Beyond the technical core, Social Sciences records 560 publications (15.40%), Decision Sciences contributes 404 (11.11%) and Medicine accounts for 324 (8.91%), reflecting sustained interest in misinformation as a socio-technical problem with decision and health implications.

The overall pattern is summarized visually in Fig. 4, which focuses on the top 10 subject areas and clearly shows the dominance of Computer Science while also highlighting meaningful participation from adjacent disciplines. Taken together, these results suggest that Transformer-based misinformation detection is rooted in computer science research but increasingly shaped by interdisciplinary applications, particularly in social and health domains.

TABLE IV. SUBJECT AREA

Subject Area	TP	%
Computer Science	3,091	84.99%
Engineering	1,018	27.99%
Mathematics	796	21.89%
Social Sciences	560	15.40%
Decision Sciences	404	11.11%
Medicine	324	8.91%
Arts and Humanities	206	5.66%
Materials Science	185	5.09%
Physics and Astronomy	140	3.85%
Business, Management and Accounting	106	2.91%
Multidisciplinary	77	2.12%
Energy	72	1.98%
Chemical Engineering	47	1.29%
Neuroscience	44	1.21%
Psychology	39	1.07%
Health Professions	38	1.04%
Biochemistry, Genetics and Molecular Biology	36	0.99%
Environmental Science	34	0.93%
Chemistry	18	0.49%
Agricultural and Biological Sciences	16	0.44%
Economics, Econometrics and Finance	15	0.41%
Dentistry	14	0.38%
Nursing	13	0.36%
Earth and Planetary Sciences	10	0.27%
Pharmacology, Toxicology and Pharmaceutics	7	0.19%
Immunology and Microbiology	3	0.08%
Veterinary	1	0.03%

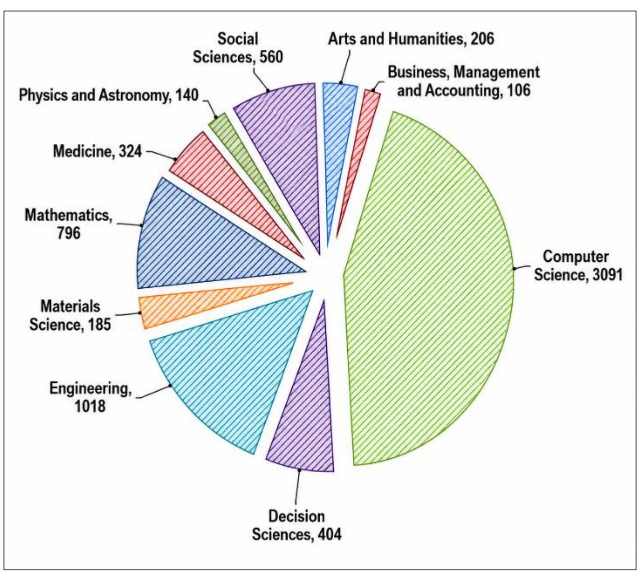


Fig. 4. Top ten subject areas.

Overall, the subject-area profile suggests that Transformer-based misinformation detection is anchored in Computer Science and draws substantial contributions from Engineering and Mathematics, reflecting a technical emphasis on model design, optimization and learning foundations. At the same time, the visible presence of Social Sciences, Decision Sciences and Medicine indicates that the field is increasingly shaped by applied and societal contexts, where misinformation is studied not only as a classification problem but also as a phenomenon involving human behavior, risk and domain-specific consequences. This interdisciplinary pattern provides useful context for interpreting the thematic clusters and collaboration networks in later analyses by explaining why research topics often connect algorithmic advances with application-driven concerns. These findings directly address RQ6 by showing that the field remains anchored in technical disciplines while drawing increasing attention from applied and societal domains.

Having established the disciplinary footprint of the literature, the next subsection examines country-level publication output to characterize the geographic distribution of research activity.

D. Country Contribution and Geographic Distribution

Country-level publication output is used to identify the geographic concentration of research activity in Transformer-based misinformation detection. The top contributing countries are summarized in Table V and their spatial distribution is illustrated in Fig. 5.

TABLE V. TOP 10 COUNTRIES WITH THE MOST TOTAL PUBLICATIONS

Country	TP	%
China	839	23.07%
United States	708	19.47%
India	645	17.73%
United Kingdom	199	5.47%
Germany	169	4.65%
Italy	142	3.90%
Australia	126	3.46%
Spain	116	3.19%
Saudi Arabia	115	3.16%
Canada	113	3.11%

As shown in Table V, publication output is concentrated among a small set of countries. China ranks first with 839 publications (23.07%), followed by the United States with 708 (19.47%) and India with 645 (17.73%). The next contributors are the United Kingdom (199, 5.47%), and Germany (169, 4.65%), followed by Italy (142, 3.90%), and Australia (126, 3.46%). Spain (116, 3.19%), Saudi Arabia (115, 3.16%), and Canada (113, 3.11%) complete the top ten.

Fig. 5 shows the same ranking, where shading intensity reflects publication volume. The concentration in a small number of countries suggests that a limited set of national research ecosystems drives output. At the same time, the presence of contributors across Asia, North America, Europe and the Middle East indicates broad international participation.

This distribution provides context for later analysis of collaboration, since geographic concentration can influence co-authorship density and cross-border research ties. This answers RQ3 by identifying the countries that contribute most to the literature.

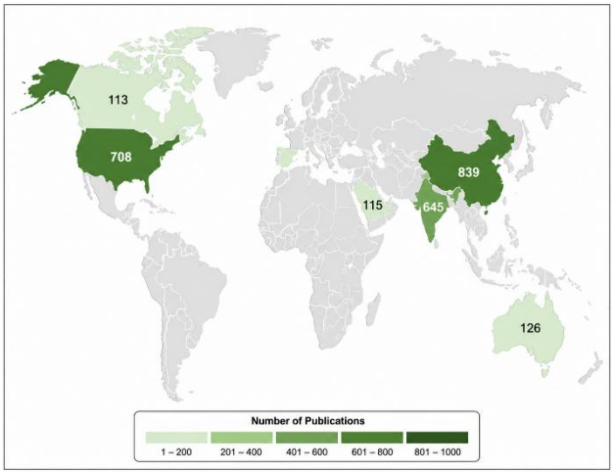


Fig. 5. Geographic distribution of the top 10 countries by total publications.

E. Institutional Contribution

Institutional-level publication output is summarized to identify the primary organizational hubs contributing to Transformer-based misinformation detection research. The distribution of the most productive institutions is shown in Fig. 6.

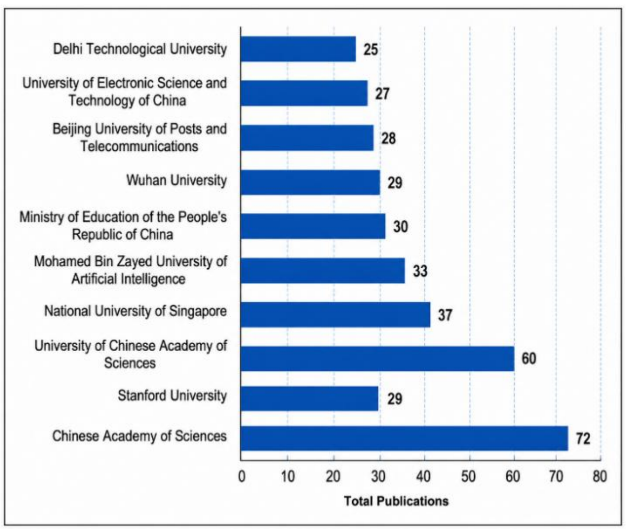


Fig. 6. Top 10 institutions by total publications.

As shown in Fig. 6, output is concentrated in a small set of institutions. The Chinese Academy of Sciences (CAS) ranks first with 72 publications, followed by the University of Chinese Academy of Sciences (UCAS) with 60 publications. These two entities are closely linked within the same research ecosystem. Still, they are reported separately because they are indexed as distinct affiliations and authors frequently list research institute and university affiliations independently. In the next tier, the National University of Singapore records 37 publications and Mohamed Bin Zayed University of Artificial Intelligence records 33 publications. Additional contributors include the Ministry of Education of the People's Republic of China (30), Wuhan University (29), Stanford University (29), Beijing University of Posts and Telecommunications (28), University of Electronic Science and Technology of China (27), and Delhi Technological University (25).

Overall, the institutional profile indicates that publication activity is strongly represented by China-based organizations, with major contributors from the United States, Singapore, UAE, and India. This answers RQ3 by identifying the institutions with the highest publication output in the dataset. Having established institutional concentration, the next subsection examines author productivity and collaboration patterns to characterize how research communities are organized at the researcher level.

F. Author Productivity and Collaboration Patterns

The collaboration structure is illustrated by the co-authorship network in Fig. 7, complemented by the productivity ranking of leading authors in Table VI. The most productive authors are summarized in Table VI and the co-authorship network is illustrated in Fig. 7.

TABLE VI. LEADING AUTHORS BY TOTAL PUBLICATIONS

Author full names	Affiliations	Author	TP
Nakov, Preslav Ivanov	Mohamed Bin Zayed University of Artificial Intelligence	United Arab Emirates	33
Alam, Firoj	Hamad Bin Khalifa University	Qatar	17
Zhang, Min	Peking University	China	12
Martín, Antonio	Universidad Politécnica de Madrid	Spain	11
Camacho, David	Universidad Autónoma de Madrid	Spain	11
Hoque, Md. Moshiur	University of Dhaka	Bangladesh	11
Ma, Jing	The Hong Kong Polytechnic University	Hong Kong	10
Madasamy, Anand	Indian Institute of Technology	India	10
Wasim, Muhammad	Lahore University of Management Sciences	Pakistan	10
Alghamdi, Jaber	King Abdulaziz University	Saudi Arabia	10

Note: TP=total number of publications

As shown in Table VI, publication output is concentrated among a small group of authors. Nakov, Preslav Ivanov, ranks first with 33 publications, followed by Alam, Firoj, with 17 and Zhang, Min, with 12. The next contributors include Martín, Antonio (11), Camacho, David (11) and Hoque, Md. Moshiur (11). Several authors contribute at a comparable level with 10 publications each, including Ma, Jing, Madasamy, Anand, Wasim, Muhammad and Alghamdi, Jaber. This distribution indicates that author productivity is top-heavy, with a small set of frequent contributors and a broader group of steady contributors.

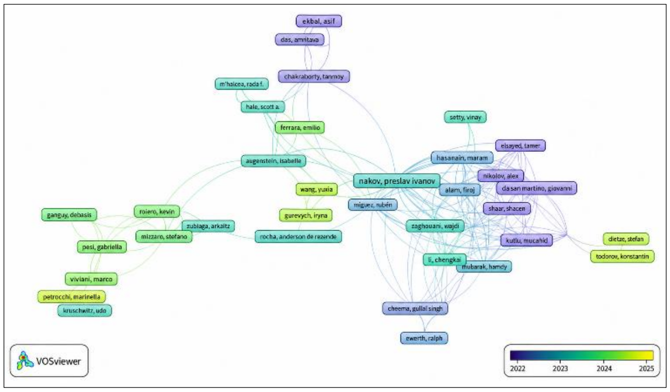


Fig. 7. Co-authorship network visualization.

The co-authorship network in Fig. 7 provides a structural view of collaboration. Nodes represent authors and links indicate co-authored publications. Larger nodes indicate greater publication presence in the dataset, while denser link patterns indicate repeated collaboration. The network forms distinct clusters, indicating collaboration communities in which authors frequently publish together. Some authors are embedded within tightly connected clusters, consistent with stable research groups. In contrast, others occupy bridging positions between clusters, suggesting cross-group collaboration pathways that can facilitate the spread of methods and datasets across the field. The color gradient adds a temporal signal by distinguishing earlier versus more recent activity within the displayed period.

Taken together, the author productivity ranking and co-authorship structure provide evidence of concentrated productivity and cluster-based collaboration patterns. This answers RQ4 by identifying the most productive authors and the dominant collaboration structure in the field. Having established the leading contributors and collaboration patterns, the next subsection maps the literature's thematic structure using keyword co-occurrence and overlay analysis.

G. Keyword Co-occurrence and Thematic Evolution

The thematic structure of Transformer-based misinformation detection research is characterized through keyword co-occurrence patterns and their temporal evolution. The co-occurrence structure is shown in Fig. 8, while the overlay visualization in Fig. 9 highlights topic recency based on average publication year. Keyword labels such as *fake news* and *disinformation* are reported as indexed terms in the dataset and are interpreted within the broader scope of misinformation detection defined in the Introduction.

As shown in Fig. 8, the literature is organized around a dense core of high-frequency keywords, with strong connectivity among Natural Language Processing, social media and misinformation-related terms. This configuration indicates that a substantial portion of the field is framed as NLP-based detection in social media environments. The map also shows connected sub-themes that reflect methodological diversification, including multimodal and cross-modal directions (e.g., terms associated with multimodal fusion and vision-based processing) and application-driven contexts such as COVID-19, which continues to appear as an anchoring topic. The presence of both baseline classifier terms and Transformer-family model terms indicates that comparative evaluation against conventional baselines remains common alongside modern representation learning approaches.

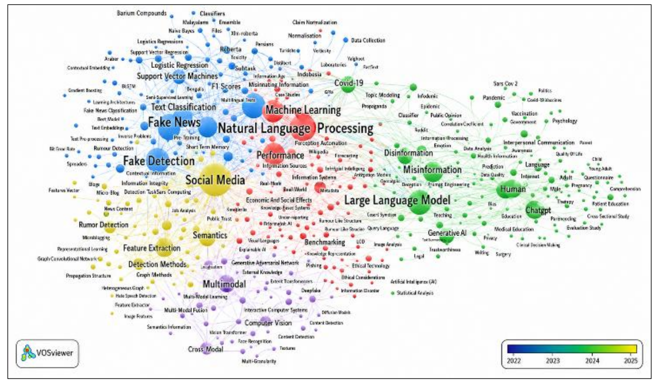


Fig. 8. Keyword co-occurrence map.

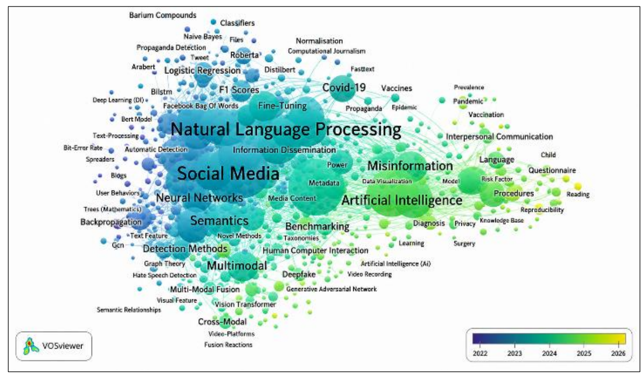


Fig. 9. Keyword overlay visualization map.

The overlay view in Fig. 9 provides a time-sensitive interpretation of keyword activity. Earlier topics are more strongly associated with general detection and evaluation vocabulary, while more recent activity shifts toward keywords linked to contemporary AI developments and deployment-oriented concerns. The appearance of terms related to generative systems and research-practice issues (e.g., reproducibility and procedure-oriented keywords) suggests that recent work is expanding beyond benchmark performance toward questions of reliability, reporting rigor and applicability in real-world conditions. At the same time, core detection topics remain active, indicating continuity in the main problem setting while also highlighting emerging directions. This temporal shift implies that evaluation designs may need to move beyond static benchmark settings toward protocols that reflect emerging threat models, including synthetic content and multimodal manipulation. It also indicates that research visibility is increasingly linked to deployment-oriented themes, which can guide future work toward robustness, transparency and real-world validation.

Specifically, the earlier activity in Fig. 9 is more closely associated with classical detection and evaluation vocabulary, such as fake news, text classification, support vector machines, logistic regression, rumor detection and F1 score. In contrast, more recent keywords include ChatGPT, generative AI, prompt engineering, reproducibility, Vision Transformer (ViT) and multimodal fusion, indicating a shift toward LLM-driven authenticity concerns, multimodal settings and deployment-oriented research practices.

Overall, the combined evidence from Fig. 8 to Fig. 9 identifies dominant thematic clusters and clarifies how attention has shifted over time. This answers RQ5 by characterizing the domain's thematic structure and its evolution through keyword co-occurrence and overlay analysis. Having established the thematic structure and its temporal shifts, the next subsection synthesizes research gaps and future directions derived from the observed trends in methods, applications and evaluation practices.

A limitation of keyword co-occurrence mapping is its dependence on author keywording practices and indexing variation, which can affect term visibility and apparent connectivity. Therefore, Fig. 8 should be read as indicating dominant thematic signals in the retrieved dataset rather than an exhaustive representation of all topics.

H. Research Gaps and Future Directions

The thematic structure and temporal overlay indicate that the field is expanding rapidly, yet several gaps remain that limit the reliability and deployability of Transformer-based misinformation detection. The key gaps and recommended future directions derived from the bibliometric findings are summarized in Table VII.

TABLE VII. RESEARCH GAPS AND FUTURE DIRECTIONS IN TRANSFORMER-BASED MISINFORMATION DETECTION

Gap area	Evidence from this study	Why it matters	Future directions (recommended)
Evaluation of realism and robustness	Keyword maps emphasize benchmark and evaluation terms (Fig. 8) and recent attention to reproducibility (Fig. 9).	Benchmark gains may not translate to deployment when data shifts across time, topics and platforms.	Use cross-domain and cross-platform evaluation, temporal holdout testing, robustness checks under distribution shift and transparent reporting of splits and baselines.
LLM-generated misinformation	Highly cited works are dominated by publications on LLMs/ChatGPT (Table III).	Detection pipelines often assume human-authored content, while LLM text can be prompt-sensitive and rapidly evolving.	Develop detection strategies for synthetic text (prompt variation, paraphrasing, mixed human–AI authorship), integrate verification signals and evaluate against evolving generator models.
Multimodal misinformation detection	Multimodal themes appear as a connected topic area (Fig. 8) and several influential publications are multimodal (Table III).	Real-world misinformation often combines text and images, so text-only models may miss important signals or be susceptible to context manipulation.	Expand multimodal datasets, evaluate cross-modal consistency and robustness and design fusion architectures that handle missing or manipulated modalities.
Explainable and trustworthy detection	Explainability-related directions appear as emerging concerns alongside deployment-oriented themes (Fig. 9).	High-stakes settings require interpretability and trust, not only accuracy, to support verification and decision-making.	Incorporate explanation methods (faithful rationales, uncertainty estimates), audit bias and failure modes and evaluate explanation quality with human-centered protocols.
Cross-lingual and low-resource settings	The thematic core is strongly anchored in mainstream NLP and social media detection terminology (Fig. 8), suggesting a dominance of high-resource settings.	Models trained on English-heavy benchmarks often generalize poorly to other languages, dialects and code-switching contexts.	Build multilingual and low-resource benchmarks, develop adaptation methods (transfer, multilingual pretraining, PEFT) and evaluate fairness across languages and regions.
Reproducibility and transparency	"Reproducibility" emerges as a recent keyword signal in the overlay map (Fig. 9).	Without consistent reporting and reusable pipelines, performance improvements may reflect experimental variance rather than real advances.	Release code and data splits, standardize preprocessing and baselines, report hyperparameters and adopt reproducible evaluation toolchains.

Taken together, Table VII indicates that recent progress is not limited by model capacity alone, but by evaluation realism, modality coverage and reporting discipline. This implies that future contributions will be more defensible when they adopt shift-aware evaluation protocols and provide reproducible pipelines alongside improvements in accuracy.

First, the literature remains strongly benchmark-driven, with frequent emphasis on standard classification performance and metric reporting, as reflected by the prominence of evaluation terms in the keyword map and its overlay evolution (Fig. 8 and Fig. 9). While benchmark progress supports comparability, the field would benefit from evaluation protocols that better reflect real-world distribution shifts over time, across platforms and across topics. Future work should prioritize cross-domain and cross-platform testing, temporal generalization and robustness assessment in the face of changing narratives and emerging events. Prior evidence also shows that dataset limitations, feature representation choices and simplistic fusion strategies remain recurring failure points in misinformation detection systems, reinforcing the need for evaluation designs that reflect these constraints [25], [26].

Second, the rise of LLMs and ChatGPT is reshaping the misinformation landscape and requires explicit methodological responses. Highly cited publications emphasize the influence and societal implications of generative conversational AI [11], including risks in sensitive domains such as public health and education [15]-[18]. However, detection pipelines often assume human-authored content and may not be calibrated to synthetic text produced by variations in prompts and model updates. Future studies should develop detection and verification strategies specifically designed for LLM-generated misinformation, including robustness under prompt manipulation, synthetic paraphrasing and mixed human–AI authorship settings, while also considering human trust and adoption behaviors that shape how such content is consumed [19].

Third, multimodal misinformation remains under-addressed relative to its prevalence in online ecosystems. Although multimodal cues have emerged as a theme and influential works have demonstrated the value of integrating visual information [14], [20], the literature still concentrates heavily on text-only inputs. Future research should move toward multimodal and cross-modal architectures evaluated on realistic mixed-media datasets, including scenarios with incomplete, noisy, or manipulated visual context. In parallel, explainable and trustworthy AI should be treated as a first-class requirement, particularly for high-stakes environments and public-facing deployments. Recent work on explainable multilingual and multimodal detection supports this direction [6], [7], but broader adoption remains limited. Beyond modality fusion, incorporating affective signals from news content and user responses has also been shown to improve detection performance on social platforms, suggesting an additional direction for feature enrichment in real-world settings [27].

Fourth, low-resource and cross-lingual settings remain comparatively underrepresented. The keyword landscape suggests dominance of mainstream NLP terms and widely used benchmarks, which often privilege English and high-resource contexts. To improve real-world applicability, future work should invest in multilingual datasets, domain adaptation techniques, and evaluation settings that emphasize fairness and generalization across languages, including code-switching and regional dialects [28]. Noisy, non-standard social media language and spelling variation can further degrade cross-lingual transfer, motivating normalization-aware pipelines and curated resources for low-resource settings [29], [30].

Finally, reproducibility and methodological transparency require stronger emphasis. The overlay map highlights "reproducibility" as a more recent concern (Fig. 9), indicating growing awareness of the need for reporting rigor. Future studies should adopt clearer documentation of preprocessing, data splits, hyperparameter configurations, and baseline selection. They should publish reusable evaluation pipelines to ensure that progress reflects genuine methodological improvement rather than experimental variance [31].

In summary, these gaps collectively answer RQ7 by identifying priorities for advancing reliable and deployable misinformation detection: 1) realistic evaluation and robustness, 2) explicit handling of LLM-generated misinformation, 3) stronger multimodal and explainable frameworks, 4) broader cross-lingual coverage, and 5) improved reproducibility and transparency.

IV. Conclusion

This study presents a Scopus-based bibliometric analysis of Transformer-based misinformation detection, drawing on 3,637 records, and provides an evidence-driven account of how the field is currently organized and where it is moving. The results show that publication output is concentrated in conference proceedings and journals and that the research base is firmly anchored in Computer Science, with sustained contributions from applied disciplines that reflect deployment-facing concerns. At the contributor level, productivity is concentrated among a small set of leading countries, institutions and co-authorship communities, indicating that research capacity is not evenly distributed. Citation metrics further confirm broad influence with a distinct, highly cited core. At the same time, keyword mapping identifies a stable focus on NLP-driven misinformation detection in social media and a clear recent shift toward multimodal pipelines and LLM-related themes.

The primary contribution of this study is a consolidated, reproducible map of productivity, influence and thematic evolution that reduces fragmentation in the literature and supports a more defensible positioning for future work. Future research may further deepen intellectual-structure analysis through citation, co-citation, and bibliographic coupling networks, refine time-sliced thematic tracking to capture topic turnover, and strengthen external validity through reproducible evaluation protocols, cross-domain and cross-platform robustness testing, and expanded coverage of low-resource and cross-lingual settings.

Acknowledgment

This study is a research product supported and funded by the Fundamental Research Grant Scheme (FRGS) of the Ministry of Higher Education Malaysia (Code: FRGS/1/2022/ICT02/ UKM/02/3).

Declaration of Generative AI in Scientific Writing

Generative AI tools were used only for language editing to improve clarity and readability. All AI-assisted revisions were reviewed and applied by the authors. The authors take full responsibility for the accuracy and integrity of the work. No AI tools were used for conceptualization, study design, data analysis, result interpretation or the development of the scientific content. In accordance with journal and ethical guidelines, no AI tool is listed as an author or co-author.

References

[1] G. Nogara et al., "A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch," Online Social Networks and Media, vol. 51, Art. no. 100342, 2026, doi: 10.1016/j.osnem.2025.100342.

[2] C. Bono, N. Liu, G. Russo and F. Pierri, "Self-moderation in the decentralized era: decoding blocking behavior on Bluesky," arXiv:2505.01174, 2025. [Online]. Available: https://arxiv.org/abs/2505.01174

[3] D. Mouratidis, A. Kanavos and K. Kermanidis, "From misinformation to insight: Machine learning strategies for fake news detection," Information, vol. 16, no. 3, Art. no. 189, 2025, doi: 10.3390/info16030189.

[4] Q. Su, M. Wan, X. Liu and C.-R. Huang, "Motivations, methods and metrics of misinformation detection: An NLP perspective," Natural Language Processing Research, vol. 1, nos. 1-2, Art. no. 1, 2020, doi: 10.2991/nlpr.d.200522.001.

[5] A. O. Mohamed, I. Asim, I. Eltayeb, A. Ruslan and N. E. Jeffry, "Fake news detection: A review of conventional and state-of-the-art approaches," Journal of Applied Science, Technology and Computing, vol. 2, no. 1, pp. 43-54, 2025. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jastec/article/view/23956

[6] C. Wang, M. Gao, Z. Wang, J. Yin, K. Shu and C. Lin, "Prompt-induced linguistic fingerprints for LLM-generated fake news detection," arXiv:2508.12632, 2025. [Online]. Available: https://arxiv.org/abs/2508.12632

[7] L. Pangtey, A. Bhatnagar, S. Bansal, S. S. Dar and N. Kumar, "Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions," arXiv:2505.08464, 2025. [Online]. Available: https://arxiv.org/abs/2505.08464

[8] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash and S. Jadhav, "Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation," Frontiers in

Artificial Intelligence, vol. 8, Art. no. 1690616, 2025, doi: 10.3389/frai.2025.1690616.

[9] C. A. Anser Pasha and S. Sampath, "X-FakeNet: Explainable multimodal fusion with meta-reasoning for fake news detection," International Journal of Intelligent Engineering and Systems, vol. 18, no. 11, pp. 645-659, 2025, doi: 10.22266/ijies2025.1231.40.

[10] 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, Art. no. 1775, 2025, doi: 10.1057/s41599-025-06067-1.

[11] S. Wang, F. Su, L. Ye and Y. Jing, "Disinformation: A bibliometric review," International Journal of Environmental Research and Public Health, vol. 19, no. 24, Art. no. 16849, 2022, doi: 10.3390/ijerph192416849.

[12] A. Alarabid, M. Ali, E. B. Abbad, A. Al Bloushi and H. Heba, "Leveraging large language models for misinformation detection: A focus on public health misinformation on social media," in Lecture Notes in Networks and Systems, vol. 1539, pp. 561-573, 2026, doi: 10.1007/978-981-96-9248-4_42.

[13] Y. N. Barve and P. Mulay, "Bibliometric survey on incremental learning in text classification algorithms for false information detection," Library Philosophy and Practice (e-journal), Art. no. 4560, 2020. [Online]. Available: https://digitalcommons.unl.edu/libphilprac/4560/

[14] S. Rastogi and D. Bansal, "A review on fake news detection 3T's: Typology, time of detection, taxonomies," International Journal of Information Security, vol. 22, no. 1, pp. 177-212, 2023, doi: 10.1007/s10207-022-00625-3.

[15] Y. K. Dwivedi et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," International Journal of Information Management, vol. 71, Art. no. 102642, 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[16] J. Clusmann et al., "The future landscape of large language models in medicine," Communications Medicine, vol. 3, no. 1, 2023, doi: 10.1038/s43856-023-00370-1.

[17] R. K. Kaliyar, A. Goswami and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools and Applications, vol. 80, no. 8, pp. 11765-11788, 2021, doi: 10.1007/s11042-020-10183-2.

[18] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in Proc. IEEE Int. Conf. Multimedia Big Data (BigMM), 2019, pp. 39-47, doi: 10.1109/BigMM.2019.00-44.

[19] L. De Angelis et al., "ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health," Frontiers in Public Health, vol. 11, Art. no. 1166120, 2023, doi: 10.3389/fpubh.2023.1166120.

[20] A. Abd-alrazaq et al., "Large language models in medical education: Opportunities, challenges and future directions," JMIR Medical Education, vol. 9, Art. no. e48291, 2023, doi: 10.2196/48291.

[21] J. Dempere, K. Modugu, A. Allam and L. K. Ramasamy, "The impact of ChatGPT on higher education," Frontiers in Education, vol. 8, Art. no. 1206936, 2023, doi: 10.3389/feduc.2023.1206936.

[22] S. Harrer, "Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine," eBioMedicine, vol. 90, Art. no. 104512, 2023, doi: 10.1016/j.ebiom.2023.104512.

[23] A. Choudhury and H. Shamszare, "Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis," Journal of Medical Internet Research, vol. 25, Art. no. e47184, 2023, doi: 10.2196/47184.

[24] P. Qi, J. Cao, T. Yang, J. Guo and J. Li, "Exploiting multi-domain visual information for fake news detection," in Proc. IEEE Int. Conf. Data Mining (ICDM), 2019, pp. 518-527, doi: 10.1109/ICDM.2019.00062.

[25] S. Kh. Hamed, M. J. Ab Aziz and M. R. Yaakub, "A review of fake news detection models: Highlighting the factors affecting model performance and the prominent techniques used," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 7, pp. 379-390, 2023.

[26] S. Kh. Hamed, M. J. Ab Aziz and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation and data fusion," Heliyon, vol. 9, Art. no. e20382, 2023, doi: 10.1016/j.heliyon.2023.e20382.

[27] S. Kh. Hamed, M. J. Ab Aziz and M. R. Yaakub, "Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments," Sensors, vol. 23, Art. no. 1748, 2023, doi: 10.3390/s23041748.

[28] X. Wang, W. Zhang and S. Rajtmajer, "Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey," arXiv:2410.18390, 2024. [Online]. Available: https://arxiv.org/abs/2410.18390

[29] S. N. A. N. Ariffin and S. Tiun, "Rule-based text normalization for Malay social media texts," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 10, pp. 156-162, 2020.

[30] Z. Mansur, N. Omar, S. Tiun and E. M. Alshari, "A hybrid BERT-based normalization framework for hate speech detection," Asia-Pacific Journal of Information Technology and Multimedia, vol. 14, no. 2, pp. 472-495, 2025, doi: 10.17576/apjitm-2025-1402-09.

[31] I. Magnusson, N. A. Smith and J. Dodge, "Reproducibility in NLP: What have we learned from the checklist?," in Findings of the Association for Computational Linguistics: ACL 2023, 2023. [Online]. Available: https://aclanthology.org/2023.findings-acl.809/