Reproducibility and Validity at Risk: Reporting and Evaluation Inconsistencies in Fake News Detection Studies

Anonymous authors
Paper under double-blind review

Abstract

The growing body of fake news detection (FND) research has led to the development of numerous machine learning and deep learning (ML/DL) approaches. However, the increasing volume of publications has not always been accompanied by corresponding methodological rigor. As a result, concerns regarding reproducibility, reliability, and validity have attracted increasing attention. In particular, inconsistencies in dataset construction and reporting, inadequate evaluation practices, methodological flaws, and insufficient experimental transparency may compromise the credibility of reported results and impede meaningful comparison across studies. This study was motivated by observations made during a systematic mapping study of the FND literature. The review followed a comprehensive search process across major digital libraries, initially identifying 1,030 publications. After applying predefined inclusion and exclusion criteria, 138 studies remained for full-text assessment. During the second screening stage, each excluded study was manually examined and its exclusion reason was documented individually. This process resulted in the exclusion of 77 studies, whose reasons were subsequently categorized into high-level groups. Although the analysis was conducted within the context of FND, the identified methodological, reporting, and evaluation deficiencies extend far beyond this particular research area. Since most empirical ML/DL studies follow a common experimental pipeline involving dataset acquisition or selection, model development, and performance evaluation, these shortcomings may affect research conducted across diverse application domains. To address these recurring issues, we propose a practical checklist derived from the systematic analysis of the excluded studies to support the transparent, reproducible, and methodologically sound design, implementation, and evaluation of empirical ML/DL research. The checklist is intended to assist authors, reviewers, and editors in assessing the methodological quality, transparency, and reproducibility of ML/DL studies.

1 Introduction

Disinformation has emerged as one of the most significant challenges of the modern information ecosystem. According to the definition proposed by the High-Level Expert Group (HLEG) (European Commission. Directorate General for Communications Networks, Content and Technology., 2018), disinformation refers to false, inaccurate, or misleading information that is intentionally created and disseminated to cause public harm or generate economic gain. Unlike accidental misinformation, disinformation involves a deliberate intent to deceive. In both academic research and public discourse, the term fake news is frequently used to describe deceptive or fabricated information, often interchangeably with disinformation. While fake news is generally more familiar to the broader public, disinformation is considered a more precise term because it explicitly emphasizes the intentional nature of deception. Throughout this study both terms are used interchangeably.

The increasing prevalence of disinformation is closely associated with the rise of digital communication technologies, which have fundamentally transformed how people consume and share information. Rather

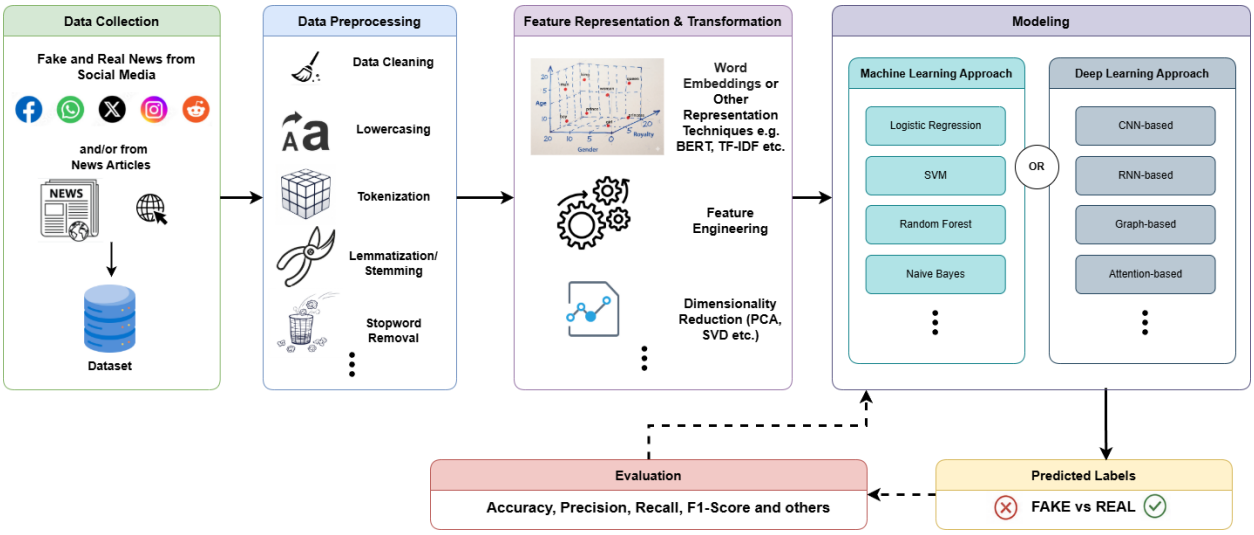

Figure 1: Representative experimental pipeline commonly adopted in machine learning and deep learning-based FND studies.

than accessing news through the official websites of individual news organizations, many users increasingly rely on social media platforms, where content from news agencies, public institutions, independent creators, and ordinary users is aggregated into a single information stream. While this shift has improved accessibility and increased the speed of information dissemination, it has also created an environment in which information of varying quality, credibility, and intent can spread rapidly to large audiences.

The challenge has become even more pronounced as social media platforms have evolved into multimodal information ecosystems where text, images, videos, audio recordings, and other forms of content coexist and are consumed simultaneously. Consequently, users of all ages can be exposed to vast amounts of unverified information across multiple languages, topics, and modalities through a single platform. The scale and speed of this information flow make manual verification increasingly difficult and often impractical.

As the volume of online content continues to grow, relying solely on human fact-checkers has become insufficient for addressing the scale of the problem. This limitation has motivated researchers to explore automated approaches for fake news and disinformation detection using ML/DL techniques. Since the surge of interest in the term "fake news" during the COVID-19 pandemic in 2019, numerous studies have proposed models designed to distinguish truthful information from deceptive content by leveraging textual, visual, social, and contextual signals. Despite differences in architectures and learning paradigms, most of these approaches follow a broadly similar experimental pipeline involving dataset acquisition or selection, data preprocessing, feature representation, model development and training, and performance evaluation. Figure 1 illustrates a representative ML/DL workflow commonly adopted in FND research. Although individual studies may incorporate additional components, the overall structure remains largely consistent across the literature.

However, the widespread adoption of ML/DL techniques has not always been accompanied by corresponding methodological rigor. While the number of published studies continues to grow, concerns regarding reproducibility, transparency, and validity have become increasingly evident. In particular, inconsistencies in dataset reporting, deficiencies in experimental design, errors in evaluation procedures, and insufficient methodological transparency may compromise the reliability of reported findings and impede meaningful comparison across studies.

To better understand these issues, this paper examines the studies excluded during the systematic mapping study (SMS) we conducted on FND research. In this paper, rather than focusing on the studies that satisfied the review criteria, we investigate the methodological, reporting, and conceptual problems identified in the excluded literature. By systematically analyzing these shortcomings, we aim to highlight recurring pitfalls

that threaten research validity and reproducibility and to provide practical insights to improve the quality of future ML/DL studies.

2 Related work

Concerns regarding the validity, transparency, and reproducibility of ML/DL research are not unique to FND. As the adoption of data-driven methods has accelerated across disciplines, several studies have examined methodological shortcomings, reporting deficiencies, and threats to reproducibility in different application domains.

One notable example is a critique of the study titled *Lightweight Hybrid Deep Learning Models for Accurate Classification of Respiratory Conditions from Raw Lung Sounds* by Lei (2026). The authors questioned the credibility of the reported results, particularly the near-perfect classification performance achieved on an eleven-class respiratory disease classification task. Their analysis highlighted several concerns, including insufficient dataset description, contradictory claims regarding model complexity, and inadequate reporting of experimental procedures and validation protocols. The study concluded that the combination of exceptionally high performance and limited methodological transparency raised substantial doubts regarding the robustness and reliability of the reported findings.

Similar concerns have been reported in the medical domain through broader evidence synthesis studies. For example, the systematic review by Huang et al. (2022) evaluated deep learning studies developed for neurosurgical outcome prediction. The authors identified dataset-related limitations, reporting deficiencies, and reproducibility challenges as major threats to validity. Their findings emphasized that, despite promising predictive performance, insufficient transparency in model development and reporting practices continues to pose a substantial barier to independent validation and reproducibility.

Methodological concerns have also been investigated in educational research (Anglin, 2024). The authors examine supervised machine learning applications such as automated essay scoring, graduation prediction, and automatic grading systems. Recognizing the rapid adoption of machine learning methods in education research, they argued that domain-specific methodological guidance is needed to ensure valid interpretations of model outputs. The study adapts classical validity concepts from the social sciences to supervised learning and proposes a practical checklist intended to help both researchers and reviewers identify common threats to validity.

Within the domain of FND, the study most closely related to our work is *Assessing Reproducibility and Accessibility in Hate Speech and Fake News Detection Datasets: A Literature Review (2023–2024)* (Đurić et al., 2025). That review focuses on six primary studies whose datasets satisfy predefined criteria related to reproducibility, accessibility, and documentation quality. The authors discussed challenges related to annotation bias, class imbalance, limited data availability, multilinguality, dataset maintenance, and documentation practices. Their analysis highlighted the importance of well-documented datasets and transparent annotation procedures for reproducible research.

Despite these valuable contributions, existing studies exhibit several limitations. First, prior work in other domains primarily focuses on specific application areas such as healthcare or education, rather than fake news detection. Second, the only closely related study within the FND literature concentrates primarily on dataset accessibility, annotation quality, and reproducibility characteristics of a small set of selected datasets. In contrast, our work adopts a broader perspective by examining methodological shortcomings across the entire machine learning pipeline, including dataset reporting, labeling practices, experimental design, evaluation procedures, metric computation, potential data leakage, and conceptual task formulation.

Furthermore, while previous reviews typically analyze studies or datasets that satisfy predefined quality criteria, our analysis focuses on studies excluded during an SMS. By investigating the reasons why peer-reviewed studies did not meet the methodological, reporting or conceptual requirements, we aim to identify common methodological problems that may otherwise remain hidden in reviews of the literature. To the best of our knowledge, this is the first study to systematically analyze exclusion decisions in FND research as a means of uncovering broader methodological challenges affecting ML/DL studies.

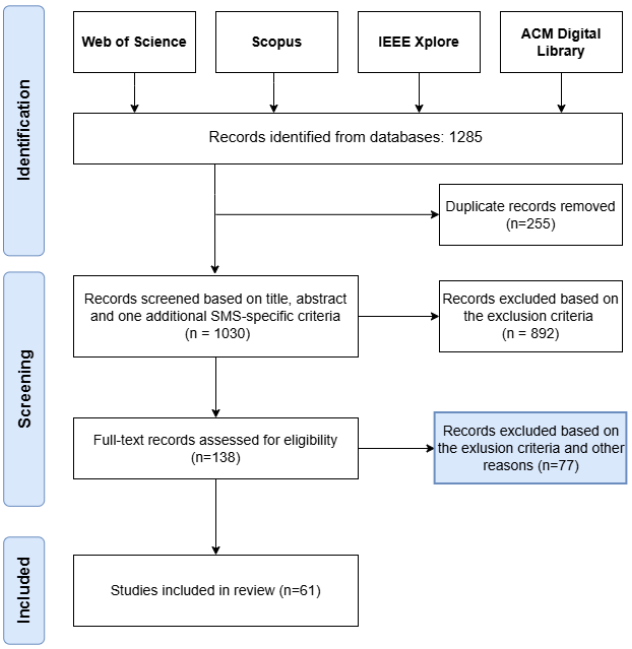

Figure 2: Study selection process and screening outcomes from the systematic mapping study that served as the foundation of this research.

3 Methodology

This study builds upon a previously conducted SMS on FND research covering the period from 2022 to 2025. The original review followed a structured and reproducible search and screening process across ACM Digital Library, IEEE Xplore, Scopus, and Web of Science (WoS), resulting in an initial pool of 1,030 records. Through successive screening stages and predefined eligibility criteria, the candidate set was progressively refined, ultimately yielding 60 primary studies for detailed analysis. The overall study selection workflow is summarized in Figure 2.

The focus of the present study is the set of papers excluded during the full-text assessment stage. More specifically, rather than examining studies excluded solely due to scope-related inclusion and exclusion criteria defined for the original SMS, we concentrate on articles that exhibited methodological and reporting shortcomings identified during the selection process. These shortcomings include issues related to dataset construction and usage, reporting practices, evaluation procedures, experimental design, and other factors that may affect the validity, reliability, and reproducibility of research findings. Although these issues were identified within the context of FND research, they are not specific to this domain. Since most ML/DL studies follow a common workflow involving dataset selection or acquisition, model development and training, and experimental evaluation, the methodological observations reported in this paper extend across a wide range of data-driven research areas.

4 Results

Overview of Exclusion Reasons Figure 3 presents the overall distribution of exclusion reasons identified during the full-text assessment stage. For reporting purposes, the documented exclusion reasons were grouped into four high-level categories: inconsistent and/or insufficient reporting, inconsistent and/or insufficient evaluation, conceptual misalignment, and other generic exclusion criteria. The last category encompasses exclusions originating from scope-specific inclusion and exclusion criteria defined for the underlying systematic mapping study. Since these exclusions primarily reflect the boundaries and objectives of the SMS rather than methodological shortcomings, they are not considered further in this paper. As shown in Figure 3, the most prominent source of exclusion was *inconsistent and/or insufficient reporting*, accounting

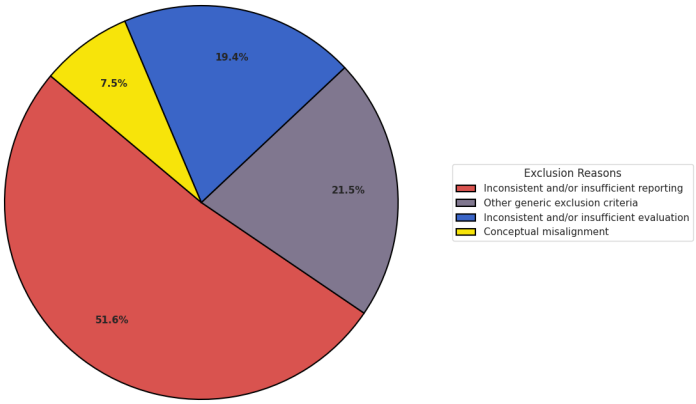

Figure 3: High-level categorization of exclusion reasons for studies removed during full-text screening. Percentages are based on exclusion-category assignments (N = 93) across 77 excluded studies.

for 51.6% of all excluded studies. In other words, more than one out of every two excluded papers exhibited reporting-related deficiencies that prevented reliable assessment of their methodology or findings. Excluding the SMS-specific criteria grouped under *other generic exclusion criteria* (21.5%), the second most common category was *inconsistent and/or insufficient evaluation* (19.4%), which includes shortcomings and errors related to experimental design, performance assessment, and result reporting. The remaining exclusions were associated with *conceptual misalignment* (7.5%), referring to studies whose interpretation of the fake news problem differed substantially from the conceptual scope adopted in this work.

The following subsections examine these three categories in detail and discuss the recurring issues identified within each group.

Inconsistent and/or Insufficient Reporting The most prevalent category of exclusion involved deficiencies in reporting practices. In many cases, the available information fell short of what is necessary to accurately understand, reproduce, compare, or verify the reported experiments. These issues affected multiple stages of the machine learning pipeline, including dataset description, labeling procedures, model specification, and result reporting.

The most frequently observed problems were related to datasets. Several studies failed to provide adequate descriptions of the datasets used for experimentation (Dorthi et al., 2022), (Kanagavalli et al., 2022), (Vadakkethil Somanathan Pillai & R, 2024), omitted references to previously established benchmark datasets (Chabukswar et al., 2023), (Merryton & Augasta, 2023), (Dorthi et al., 2022), (Kanagavalli & Baghavathi Priya, 2022), (Sujitha et al., 2025) or cited incorrect sources (Roslan et al., 2024), (Kanagavalli et al., 2022). In some cases, the reported dataset statistics were inconsistent with the referenced datasets (Jin & Wang, 2025), (Gupta et al., 2024), (Ahmed et al., 2024), (Roslan et al., 2024), (Gupta et al., 2023), (Merryton & Augasta, 2023), (Agarwal et al., 2022), (Agarwal & Samavedhi, 2022), (Kanagavalli & Baghavathi Priya, 2022), (Liguori et al., 2025), (Deepthi & Shastry, 2024). To illustrate the extent of this issue, Table 1 summarizes the characteristics reported for the PolitiFact dataset—a subset of the FakeNewsNet benchmark—including publication year, label space, reported numbers of fake and real news instances, and the corresponding dataset reference used by each study. For context, the table also includes the statistics of the two officially documented versions of the PolitiFact dataset. The table reveals substantial variation in the reported dataset statistics despite studies ostensibly using the same benchmark dataset. [1]

[1]Note: The first three rows present the original statistics reported in the FakeNewsNet papers and the official GitHub repository. The middle section lists studies reporting inconsistent statistics, while the final section lists studies reporting consistent statistics.

Table 1: Reported characteristics of datasets referred to as "PolitiFact" in the reviewed studies, including dataset statistics and cited dataset sources.

Study	**Publication Year**	**Referenced Dataset Source**	**Label Space**	**Fake**	**Real**	**Total**
FakeNewsNet Paper (Shu et al., 2018)	2018	politifact.com	"real", "fake"	432	624	1056
FakeNewsNet Paper (Shu et al., 2020)	2020	politifact.com	"real", "fake"	432	624	1056
FakeNewsNet Official GitHub Repository (Shu, 2019)	last updated in 2019	politifact.com	"real", "fake"	473	797	1270
(Khan et al., 2024)	2022	(Shu et al., 2018)	"true", "fake"	270	145	415
(Galli et al., 2022)	2022	(Shu et al., 2018)	"real", "fake"	367	434	801
(Shi et al., 2024)	2023	(Shu et al., 2020)	"real", "fake"	260	260	520
(Karami et al., 2023)	2023	(Shu et al., 2020)	"real", "fake"	319	132	451
(Liao et al., 2023)	2023	(Shu et al., 2020)	"real", "fake"	345	399	744
(Hou et al., 2024)	2024	(Shu et al., 2020)	"true", "fake"	364	457	821
(Alghamdi et al., 2024)	2024	(Shu et al., 2018), (Shu, 2019)	"true", "fake"	338	356	694
(Alghamdi et al., 2025)	2025	(Shu, 2019)	"true", "fake"	338	356	694
(Sahi et al., 2025)	2025	(Shu et al., 2020)	"real", "fake"	157	624	781
(Comito et al., 2025)	2025	(Shu et al., 2018)	"real", "fake"	448	370	818
(Gupta et al., 2025)	2025	(Shu et al., 2020)	"real", "fake"	n/a	n/a	815
(Furukawa et al., 2021)	2022	(Shu, 2019)	"real", "fake"	430	624	1054
(Faye et al., 2023)	2023	(Shu, 2019)	"true", "false"	432	624	1056
(Ayetiran & Özgöbek, 2024)	2024	(Shu et al., 2020)	"real", "fake"	432	624	1056
(Nawaz et al., 2024)	2024	(Shu et al., 2020)	"real", "fake"	474	798	1272
(Li et al., 2025)	2025	(Shu et al., 2020)	"real", "fake"	473	797	1270
(Zhou et al., 2025)	2025	(Shu et al., 2020)	"real", "fake"	473	797	1270

In addition to the dataset-related isuues, reported numbers in some studies appeared to correspond to different dataset altogether (Ateaa et al., 2023). We also identified studies with unclear dataset composition, insufficient information regarding data collection procedures, and a general lack of transparency concerning the underlying experimental data (Kazmi et al., 2025), (Ramesh et al., 2025). Such issues make it difficult to determine what data were actually used and whether the reported results can be independently reproduced.

A second recurring issue concerned the reporting of experimental results. Some studies reported only overall accuracy while omitting complementary evaluation metrics necessary for a comprehensive assessment of model performance (Mondal et al., 2023), (Cîrnu et al., 2023), (Park et al., 2023), (Sarkale et al., 2023).

Furthermore, in some cases, evaluation results were reported only graphically (e.g., bar charts) rather than as explicit numerical values (Alnabhan & Branco, 2024), limiting the precision of comparisons across studies. Other papers claimed experiments on multiple datasets but reported results for only a subset of them (K et al., 2023b).

Label-related reporting deficiencies were also common. Several studies failed to clearly describe the labeling scheme associated with the datasets they employed. For instance, datasets originally intended for multi-class classification were occasionally converted into binary classification tasks without clarifying the class-merging process (Ahamed Nishath & Murugeswari, 2025), (Merryton & Augasta, 2023) or specifying the exact version implemented (Sujitha et al., 2025). In other cases, benchmark dataset labels were incorrectly reported or described in a manner inconsistent with the original dataset documentation (Dorthi et al., 2022). Since label definitions directly determine the learning objective, inadequate reporting in this area introduces ambiguity regarding both the task formulation and the interpretation of results.

Finally, a small number of studies exhibited inconsistencies between their claimed methodology and the actual experimental setup described in the paper. For example, one study (Ahmed et al., 2025) stated that a ResNet architecture was used for image validation as part of the proposed model. However, the architecture diagram provided in the paper did not include any component corresponding to ResNet, and the dataset used for experimentation did not contain image data, making the reported use of an image-based architecture difficult to reconcile with the described experimental setting.

Beyond the above issues, we also encountered an isolated but particularly concerning situation that involved questions about the originality of the reported work (Kowsalya & Krishnan, 2023). In this instance, substantial portions of the methodology, implementation details, and all output figures appeared to closely overlap with material available in a publicly accessible source code repository SushwanthReddy, without clear attribution or sufficient evidence of independent contribution. The extent of this overlap made it difficult to verify the originality of the reported research. Although such cases were rare within the reviewed literature, they represent severe violations of scientific integrity and further underscore the necessity of transparent reporting, proper attribution, and rigorous documentation as prerequisites for reliable and reproducible machine learning research.

Inconsistent and/or Insufficient Evaluation The second major category of exclusion involved deficiencies in evaluation procedures and performance assessment. Unlike reporting-related issues, which primarily affect reproducibility and transparency, the problems identified in this category directly threaten the validity of the reported results. In several cases, the reported evaluation outcomes were mathematically inconsistent, experimentally unjustified, or incompatible with the dataset partitions described in the study.

Confusion matrix-related problems were among the most common evaluation deficiencies identified in the reviewed studies. Several studies reported precision, recall, and F1-score values that were mathematically impossible to obtain from the provided confusion matrices. For example, some papers reported F1-scores exceeding both precision and recall (Ahamed Nishath & Murugeswari, 2025), (Haritha et al., 2024), despite the fact that the F1-score is the harmonic mean of these two measures and therefore cannot be greater than both. We also identified cases where the precision, recall, and F1-score values reported in the paper were inconsistent with the confusion matrix from which they were supposedly derived (Praveen Kumar et al., 2025), (González-Celi et al., 2025), (Rini Sovia et al., 2025), (Naranjo-Espin et al., 2025). Another recurring issue involved inconsistencies between dataset partitioning strategies and the reported confusion matrices presented in the results. In several studies, the number of instances represented in the confusion matrix did not match the size of the test set implied by the reported train-validation-test split (Jin & Wang, 2025), (Alghamdi et al., 2025), (Ahmed et al., 2025), (Sharma et al., 2022). For example, studies claiming a 70-15-15 partitioning strategy sometimes reported confusion matrices whose total number of instances was substantially different from the expected test set size. We also encountered papers in which confusion matrices were presented using percentages rather than absolute counts, making independent verification of the reported metrics difficult or impossible (K et al., 2023b). Finally, confusion frequently arose from the interpretation of positive and negative classes (Oad et al., 2024), (Sharma et al., 2022). In binary classification settings, some studies provided one definition of the positive class when describing evaluation metrics but appeared to adopt a different definition when interpreting the confusion matrix and reporting

results. Such inconsistencies can lead to incorrect calculations of precision, recall, and F1-score and may substantially distort the interpretation of model performance. Beyond confusion matrix inconsistencies, we also identified a broader group of issues related to training, dataset partitioning, and evaluation protocols, where inconsistencies in experimental design and data usage undermined the validity and reproducibility of the reported results. Some studies reported conflicting information regarding dataset partitioning and experimental setup. For instance, one paper (Reddy et al., 2024) described a 60/20/20 train-validation-test split in the methodology section, while the experimental tables suggested an 80/20 train-test split. Finally, we identified cases suggesting evaluation on inappropriate data partitions. These included studies where confusion matrices appeared to be computed using the train set (Praveen Kumar et al., 2025) or the entire dataset rather than an unseen test set (Li et al., 2025), (Zhou et al., 2025), (K et al., 2023a), as well as cases where evaluation results seemed to be derived from combined validation and test partitions (Galli et al., 2022). Such practices blur the distinction between model development and independent performance assessment, increasing the risk of data leakage and overly optimistic performance estimates. Since the primary purpose of model evaluation is to assess generalization to unseen data, the use of inappropriate evaluation partitions fundamentally undermines the validity and interpretability of the reported findings.

Conceptual Misalignment The final category of exclusions arose from conceptual differences regarding the definition and operationalization of disinformation. Unlike the previous categories, these exclusions were not primarily driven by methodological deficiencies. Rather, they reflected a mismatch between the research tasks addressed by the studies and the conceptual scope adopted in our systematic literature mapping study. Because studies that appeared comparable at a methodological level were not always addressing the same phenomenon, a clearly defined and consistently applied conceptual scope is essential for ensuring meaningful interpretation and well-founded conclusions in systematic literature reviews.

Following the broader research framework adopted in the original SMS study, we use the definition proposed by the High-Level Expert Group (HLEG), which characterizes disinformation as *false, inaccurate, or misleading information that is intentionally created and disseminated to cause public harm or generate profit.* Under this definition, content types such as satire and parody are excluded because they do not necessarily intend to deceive. Similarly, phenomena such as hate speech, incitement, and other forms of harmful or illegal content are considered conceptually distinct from disinformation, even though they may co-occur in online environments.

A common source of conceptual misalignment involved studies that framed the task as propaganda detection rather than disinformation detection (Nyzova et al.), (Halloush et al., 2023), (Ebalo & Poursardar, 2024) (Horák et al., 2024), (Haidar Rasyid et al., 2023). Several datasets employed in these studies were designed for the recognition of propaganda techniques rather than the assessment of content veracity. Instead of assigning labels such as true, false, real, fake or mostly true etc, these datasets annotated manipulative attributes and rhetorical techniques, including categories such as blaming, demonization, fear-mongering, and labeling. While these tasks are closely related to information manipulation, they do not necessarily assess the factual accuracy or veracity of the underlying content. Consequently, such studies were considered outside the scope of our work.

Additional examples included studies that defined disinformation through labels related to hate speech, profanity, and unsupported claims (Putra et al., 2023), as well as a study focused on crisis or event detection rather than veracity assessment (Ganeshmoorthy et al., 2025). In both cases, since the datasets did not contain truthfulness-based labels, we considered those studies to address a different research objective and they were therefore excluded from further consideration

5 Conclusion and discussion

In this study, we examined the exclusion reasons associated with peer-reviewed primary studies identified during a systematic mapping study on FND. We investigated the methodological, reporting, and conceptual issues that led to the exclusion of a substantial portion of the candidate literature. Although the analysis was conducted within the context of FND, the identified problems are not domain-specific and are applicable to a

broad range of ML/DL studies that follow a common pipeline involving dataset selection, model development, training, and evaluation.

Our findings indicate that the most prevalent issues were related to benchmark dataset usage and evaluation procedures, particularly confusion matrix analysis and metric computation. A recurring problem concerned insufficient dataset traceability. While benchmark datasets play a critical role in enabling reproducibility and comparability, their benefits depend on accurate and transparent reporting. During the review process, we encountered numerous studies that referenced datasets through third-party hosting platforms rather than original sources. In several cases, the cited links were no longer accessible, making it impossible to determine the exact dataset version used in the experiments. One frequently encountered example was a Kaggle-hosted dataset link that is no longer available. Such situations highlight that reproducibility depends not only on reporting the dataset name but also on providing reliable and persistent references to the original dataset source and version.

The results also reveal the scale of the problem. The number of studies excluded due to methodological deficiencies, reporting inconsistencies, or evaluation-related issues was approximately 2.5 times greater than the number excluded because of scope-specific criteria defined for the original SMS. Consequently, 71% of all exclusions originated from issues that were, in principle, avoidable. This observation suggests that a substantial proportion of the literature review effort was devoted to identifying and filtering studies whose findings could not be reliably interpreted, verified, or compared. Such deficiencies introduce considerable noise into evidence synthesis studies and increase the effort required for both researchers conducting reviews and readers attempting to interpret the state of the art.

Beyond the exclusion categories discussed throughout this paper, the review process also revealed several surprising quality-related issues. For example, we encountered studies in which critical result visualizations, including confusion matrices, were presented in a form that was unreadable, limiting the transparency of the reported findings (e.g., (Ahmed et al., 2025)). Other studies contained references to non-existent tables or citations that could not be located within the manuscript (e.g., (Kanagavalli et al., 2022)). We also observed cases where experimental results were reported primarily through screenshots rather than structured scientific presentation, as well as figures that appeared incomplete or insufficiently prepared for publication (e.g., (K et al., 2023a)). Although these observations were not always directly responsible for exclusion, they raise concerns regarding quality-control processes in scientific publishing.

Perhaps the most surprising observation was that several studies exhibiting substantial reporting and methodological deficiencies had been published in well-established venues, including high-impact journals. Examples identified during the review included papers published in venues such as *Journal of Information Technology & Politics* (Zhou et al., 2025), *IEEE Access* (Oad et al., 2024), *SN Computer Science* (Gupta et al., 2023) and *PeerJ Computer Science* (K et al., 2023b). While isolated errors can occur in any publication process, the presence of fundamental deficiencies in peer-reviewed literature highlights the need for stronger evaluation criteria and more rigorous verification of ML/DL experiments.

The growing accessibility of ML/DL frameworks further increases the importance of this issue. Modern software libraries have significantly lowered the technical barriers associated with model development and experimentation. Although this democratization of technology has accelerated scientific progress, it has also increased the risk that experimental workflows requiring careful methodological consideration at every stage—from dataset acquisition and preprocessing to model training and final evaluation—may be conducted without sufficient scrutiny, particularly under the pressure to produce results quickly and keep pace with emerging trends.

Addressing these challenges requires collective effort from all stakeholders involved in the publication process, including authors, reviewers, editors, and publishers. As a practical outcome of this work, we introduce a checklist derived from recurring issues identified in the excluded studies (Appendix A). The checklist is designed to support researchers during manuscript preparation and to help reviewers systematically identify common methodological and reporting deficiencies. We hope that this resource will contribute to improving the rigor, transparency, and reproducibility of future ML/DL research, both within fake news detection and across related application domains.

References

Payal Agarwal, Sandeep Reddivari, and Kalyan Reddivari. Fake News Detection: An Investigation based on Machine Learning. In *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 61–62, San Diego, CA, USA, August 2022. IEEE. ISBN 978-1-6654-6603-5. doi: 10.1109/IRI54793.2022.00025.

Swati Agarwal and Adithya Samavedhi. Profiling Fake News: Learning the Semantics and Characterisation of Misinformation. In Bohan Li, Lin Yue, Jing Jiang, Weitong Chen, Xue Li, Guodong Long, Fei Fang, and Han Yu (eds.), *Advanced Data Mining and Applications*, volume 13087, pp. 203–216. Springer International Publishing, Cham, 2022. ISBN 978-3-030-95404-8 978-3-030-95405-5. doi: 10.1007/978-3-030-95405-5_15. Series Title: Lecture Notes in Computer Science.

S. Ahamed Nishath and R. Murugeswari. A Multichannel Convolutional Neural Networks with Bidirectional LSTM: An Investigation Into Social Network for the Identification of Fake News. In Xin-She Yang, Simon Sherratt, Nilanjan Dey, and Amit Joshi (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, volume 1416, pp. 67–81. Springer Nature Singapore, Singapore, 2025. ISBN 978-981-96-6440-5 978-981-96-6441-2. doi: 10.1007/978-981-96-6441-2_8. Series Title: Lecture Notes in Networks and Systems.

Hadeer Ahmed, Issa Traore, Sherif Saad, and Mohammad Mamun. Effect of Text Augmentation and Adversarial Training on Fake News Detection. *IEEE Transactions on Computational Social Systems*, 11(4): 4775–4789, August 2024. ISSN 2329-924X, 2373-7476. doi: 10.1109/TCSS.2023.3344597.

Khandakar Rabbi Ahmed, Md. Sayham Khan, Md Anisur Rahman Chowdhury, Shah Tawkir Nesar, Md Afjal Hosien, Md Razaul Karim, and Nasrin Sultana. Detecting Misinformation with Multimodal AI: Leveraging Vision and NLP for Fact-Checking. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pp. 1–6, Rangpur, Bangladesh, July 2025. IEEE. ISBN 979-8-3315-9694-1. doi: 10.1109/QPAIN66474.2025.11171663.

Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media. *Engineering Applications of Artificial Intelligence*, 137:109240, November 2024. ISSN 09521976. doi: 10.1016/j.engappai.2024.109240.

Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. ABERT: Adapting BERT model for efficient detection of human and AI-generated fake news. *International Journal of Information Management Data Insights*, 5 (2):100353, December 2025. ISSN 26670968. doi: 10.1016/j.jjimei.2025.100353.

Mohammad Q. Alnabhan and Paula Branco. Evaluating Deep Learning for Cross-Domains Fake News Detection. In Mohamed Mosbah, Florence Sèdes, Nadia Tawbi, Toufik Ahmed, Nora Boulahia-Cuppens, and Joaquin Garcia-Alfaro (eds.), *Foundations and Practice of Security*, volume 14552, pp. 40–51. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-57539-6 978-3-031-57540-2. doi: 10.1007/978-3-031-57540-2_4. Series Title: Lecture Notes in Computer Science.

Kylie Anglin. Addressing Threats to Validity in Supervised Machine Learning: A Framework and Best Practices for Education Researchers. *AERA Open*, 10:23328584241303495, January 2024. ISSN 2332-8584, 2332-8584. doi: 10.1177/23328584241303495.

Heba Yousef Ateaa, Ali Hussein Hasan, and Ahmed Sabeeh Ali. Detection of Fake News by Machine Learning with Linear Classification Algorithms: A Comparative Study. In Ashish Khanna, Zdzislaw Polkowski, and Oscar Castillo (eds.), *Proceedings of Data Analytics and Management*, volume 572, pp. 845–859. Springer Nature Singapore, Singapore, 2023. ISBN 978-981-19-7614-8 978-981-19-7615-5. doi: 10.1007/978-981-19-7615-5_68. Series Title: Lecture Notes in Networks and Systems.

Eniafe Festus Ayetiran and Özlem Özgöbek. An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. *Information Systems*, 123:102378, July 2024. ISSN 03064379. doi: 10.1016/j.is.2024.102378.

Arati Chabukswar, P Deepa Shenoy, and Venugopal K R. Fake News Detection Using Optimized Deep Learning Model Through Effective Feature Extraction. In *2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, pp. 118–123, Manipal, India, November 2023. IEEE. ISBN 979-8-3503-0663-7. doi: 10.1109/ICRAIS59684.2023.10367082.

Carmela Comito, Massimo Guarascio, Angelica Liguori, Giuseppe Manco, and Francesco Sergio Pisani. Beyond the Horizon: Using Mixture of Experts for Domain Agnostic Fake News Detection. In Dino Pedreschi, Anna Monreale, Riccardo Guidotti, Roberto Pellungrini, and Francesca Naretto (eds.), *Discovery Science*, volume 15244, pp. 396–410. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-78979-3 978-3-031-78980-9. doi: 10.1007/978-3-031-78980-9_25. Series Title: Lecture Notes in Computer Science.

Carmen Elena Cîrnu, Ioana-Cristina Vasiloiu, and Carmen-Ionela Rotună. Comparative analysis of the main machine learning algorithms for the automatic recognition of fake news. *Revista Română de Informatică și Automatică*, 33(1):57–66, March 2023. ISSN 12201758, 18414303. doi: 10.33436/v33i1y202305.

K Deepthi and Aditya K Shastry. A Deep Learning Methodology-Based Unsupervised Misinformation Detection (Umd) Technique for Identifying Fabricated Narratives. In *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pp. 1–6, Bengaluru, India, August 2024. IEEE. ISBN 979-8-3503-7289-2. doi: 10.1109/NMITCON62075.2024.10699051.

Kumar Dorthi, Soora Narasimha Reddy, and Muskula Shreya. Detection of Fake News: A Machine Learning Approach. In Vikrant Bhateja, Lai Khin Wee, Jerry Chun-Wei Lin, Suresh Chandra Satapathy, and T. M. Rajesh (eds.), *Data Engineering and Intelligent Computing*, volume 446, pp. 463–469. Springer Nature Singapore, Singapore, 2022. ISBN 978-981-19-1558-1 978-981-19-1559-8. doi: 10.1007/978-981-19-1559-8_48. Series Title: Lecture Notes in Networks and Systems.

Emily Ebalo and Faryaneh Poursardar. Comparative Study of ML Techniques for Misinformation Detection in War Reporting. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 8650–8652, Washington, DC, USA, December 2024. IEEE. ISBN 979-8-3503-6248-0. doi: 10.1109/BigData62323.2024.10825715.

European Commission. Directorate General for Communications Networks, Content and Technology. *A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation.* Publications Office, LU, 2018. doi: 10.2759/0156.

Géraud Faye, Wassila Ouerdane, Guillaume Gadek, Souhir Gahbiche, and Sylvain Gatepaille. A novel hybrid approach for text encoding: Cognitive Attention To Syntax model to detect online misinformation. *Data & Knowledge Engineering*, 148:102230, November 2023. ISSN 0169023X. doi: 10.1016/j.datak.2023.102230.

Ryoya Furukawa, Daiki Ito, Yuta Takata, Hiroshi Kumagai, Masaki Kamizono, Yoshiaki Shiraishi, and Masakatu Morii. Fake News Detection via Biased User Profiles in Social Networking Sites. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 136–145, ESSENDON VIC Australia, December 2021. ACM. ISBN 978-1-4503-9115-3. doi: 10.1145/3486622.3493939.

Antonio Galli, Elio Masciari, Vincenzo Moscato, and Giancarlo Sperlí. A comprehensive Benchmark for fake news detection. *Journal of Intelligent Information Systems*, 59(1):237–261, August 2022. ISSN 0925-9902, 1573-7675. doi: 10.1007/s10844-021-00646-9.

S. Ganeshmoorthy, Mouniga. K, Abinaya. C. M, Ilakkiya. G, Selvimuthu. J, and Vikash Kumar. V. S. CrisisSense: A Real-Time Framework for Event Detection and Misinformation Filtering in Social Media Streams. In *2025 International Conference on Sustainable Communication Networks and Application (ICSCN)*, pp. 607–614, Theni, India, October 2025. IEEE. ISBN 979-8-3315-9420-6. doi: 10.1109/ICSCN67106.2025.11308324.

Sebastián González-Celi, Henry N. Roa, Jorge Cruz-Silva, Edison Loza-Aguirre, Nelson Salgado-Reyes, and Javier Guaña-Moya. Ensemble Learning for Fake News Detection: Enhancing Classification Accuracy and Explainability with Structural and Statistical Metadata. In Kohei Arai (ed.), *Intelligent Systems*

and Applications, volume 1554, pp. 485–502. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-99964-2 978-3-031-99965-9. doi: 10.1007/978-3-031-99965-9_30. Series Title: Lecture Notes in Networks and Systems.

Shubham Gupta, Narendra Yadav, Suman Kundu, and Sainathreddy Sankepally. FakEDAMR: Fake News Detection Using Abstract Meaning Representation Network. In Hocine Cherifi, Luis M. Rocha, Chantal Cherifi, and Murat Donduran (eds.), *Complex Networks & Their Applications XII*, volume 1141, pp. 308–319. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-53467-6 978-3-031-53468-3. doi: 10.1007/978-3-031-53468-3_26. Series Title: Studies in Computational Intelligence.

Shubham Gupta, Abhishek Rajora, and Suman Kundu. EA^2 N: Evidence-Based AMR Attention Network for Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1742–1752, April 2025. ISSN 1041-4347, 1558-2191, 2326-3865. doi: 10.1109/TKDE.2025.3529707.

Sonam Gupta, Bhanu Verma, Pradeep Gupta, Lipika Goel, Arun Kumar Yadav, and Divakar Yadav. Identification of Fake News Using Deep Neural Network-Based Hybrid Model. *SN Computer Science*, 4(5): 679, September 2023. ISSN 2661-8907. doi: 10.1007/s42979-023-02117-0.

R. P. Haidar Rasyid, Yuliant Sibaroni, and Aditya Firman Ihsan. Classification of Disinformation Tweet on the 2024 Presidential Election in Indonesia Using Optimal Tranformer Based Model. In *2023 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 191–196, Bandung, Indonesia, August 2023. IEEE. ISBN 979-8-3503-0519-7. doi: 10.1109/ICoDSA58501.2023.10277101.

Zain Halloush, Ahmed Aleroud, and Craig Albert. Socio-Emotional Computational Analysis of Propaganda Campaigns on Social Media Users in the Middle East. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 1413–1421, Austin TX USA, April 2023. ACM. ISBN 978-1-4503-9419-2. doi: 10.1145/3543873.3587677.

D. Haritha, D. Kavitha, and M. V. Sumanth. Hybrid-Multi-channel Deep Neural Network for Fake News Detection. In Isaac Woungang, Sanjay Kumar Dhurandher, and Yumnam Jayanta Singh (eds.), *Proceedings of the NIELIT's International Conference on Communication, Electronics and Digital Technology*, volume 1022, pp. 51–68. Springer Nature Singapore, Singapore, 2024. ISBN 978-981-97-3600-3 978-981-97-3601-0. doi: 10.1007/978-981-97-3601-0_5. Series Title: Lecture Notes in Networks and Systems.

Aleš Horák, Radoslav Sabol, Ondřej Herman, and Vít Baisa. Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. *Expert Systems with Applications*, 251:124085, October 2024. ISSN 09574174. doi: 10.1016/j.eswa.2024.124085.

Ziwei Hou, Bahadorreza Ofoghi, Nayyar Zaidi, and John Yearwood. Aspect-Based Fake News Detection. In De-Nian Yang, Xing Xie, Vincent S. Tseng, Jian Pei, Jen-Wei Huang, and Jerry Chun-Wei Lin (eds.), *Advances in Knowledge Discovery and Data Mining*, volume 14650, pp. 95–107. Springer Nature Singapore, Singapore, 2024. ISBN 978-981-97-2265-5 978-981-97-2266-2. doi: 10.1007/978-981-97-2266-2_8. Series Title: Lecture Notes in Computer Science.

Jonathan Huang, Nathan A. Shlobin, Michael DeCuypere, and Sandi K. Lam. Deep Learning for Outcome Prediction in Neurosurgery: A Systematic Review of Design, Reporting, and Reproducibility. *Neurosurgery*, 90(1):16–38, January 2022. ISSN 0148-396X, 1524-4040. doi: 10.1227/NEU.0000000000001736.

Huiru Jin and Peng Wang. Fake news detection on social media using triple-attention mechanism optimized by advanced tailor optimization algorithm. *Egyptian Informatics Journal*, 32:100815, December 2025. ISSN 11108665. doi: 10.1016/j.eij.2025.100815.

Sugunedham S. K, Kadhirvel M, Ramya Kumari T, and Vanitha A. Advancing Media Integrity through AI-Powered Fake News Detection. In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–6, PUDUCHERRY, India, November 2023a. IEEE. ISBN 979-8-3503-1512-7. doi: 10.1109/ICSCAN58655.2023.10395055.

Venkatachalam K, Badriyya B. Al-onazi, Vladimir Simic, Erfan Babaee Tirkolaee, and Chiranjibe Jana. DeepFND: an ensemble-based deep learning approach for the optimization and improvement of fake news detection in digital platform. *PeerJ Computer Science*, 9:e1666, December 2023b. ISSN 2376-5992. doi: 10.7717/peerj-cs.1666.

N. Kanagavalli and S. Baghavathi Priya. Social Networks Fake Account and Fake News Identification with Reliable Deep Learning. *Intelligent Automation & Soft Computing*, 33(1):191–205, 2022. ISSN 1079-8587. doi: 10.32604/iasc.2022.022720.

N. Kanagavalli, S. Baghavathi Priya, and Jeyakumar D. Design of Hyperparameter Tuned Deep Learning based Automated Fake News Detection in Social Networking Data. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 958–963, Erode, India, March 2022. IEEE. ISBN 978-1-6654-1028-1. doi: 10.1109/ICCMC53470.2022.9753739.

Mansooreh Karami, David Mosallanezhad, Paras Sheth, and Huan Liu. Silence Speaks Volumes: Reweighting Techniques for Under-Represented Users in Fake News Detection. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1430–1437, Shanghai, China, December 2023. IEEE. ISBN 979-8-3503-8164-1. doi: 10.1109/ICDMW60847.2023.00182.

Misbah Kazmi, Sadia Nauman, Sadaf Abdul Rauf, S. Ali, Ali Daud, and Bader Alshemaimri. Fake News Detection on Kashmir Issue Using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*, 16(6), 2025. ISSN 21565570, 2158107X. doi: 10.14569/IJACSA.2025.01606106.

Fazlullah Khan, Ryan Alturki, Gautam Srivastava, Foziah Gazzawe, Syed Tauhid Ullah Shah, and Spyridon Mastorakis. Explainable Detection of Fake News on Social Media Using Pyramidal Co-Attention Network. *IEEE Transactions on Computational Social Systems*, 11(4):4574–4583, August 2024. ISSN 2329-924X, 2373-7476. doi: 10.1109/TCSS.2022.3207993.

S. Kowsalya and N. Krishnan. A Comparative Analysis of Machine Learning Approaches for Fake News Detection. In *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pp. 720–725, Trichy, India, August 2023. IEEE. ISBN 979-8-3503-2579-9. doi: 10.1109/ICAISS58487.2023.10250577.

Zhihao Lei. "Perfection or Overfitting?" Reassessing the Validity of Deep Learning Models for Respiratory Sound Classification. *Journal of Medical Systems*, 50(1):29, March 2026. ISSN 1573-689X. doi: 10.1007/s10916-026-02336-x.

Yujie Li, Yushui Xiao, Yong Huang, and Rui Ma. Analysis of social networks content to identify fake news using stacked combination of deep neural networks. *Egyptian Informatics Journal*, 30:100707, June 2025. ISSN 11108665. doi: 10.1016/j.eij.2025.100707.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. MUSER: A MUlti-Step Evidence Retrieval Enhancement Framework for Fake News Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4461–4472, Long Beach CA USA, August 2023. ACM. ISBN 979-8-4007-0103-0. doi: 10.1145/3580305.3599873.

Angelica Liguori, Francesco Sergio Pisani, Carmela Comito, Massimo Guarascio, and Giuseppe Manco. Breaking domain barriers: mixture of experts for cross-domain fake news detection. *Machine Learning*, 114(8):188, August 2025. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-025-06827-9.

Adline Rajasenah Merryton and M Gethsiyal Augasta. A Novel Framework for Fake News Detection using Double Layer BI-LSTM. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1689–1696, Tirunelveli, India, January 2023. IEEE. ISBN 978-1-6654-7467-2. doi: 10.1109/ICSSIT55814.2023.10061026.

Semanto Mondal, Prakash Srinivasan, and Md. Ekhtear Ahammed. Fake News Detection: A Comparative Study of Machine Learning Techniques. In *2023 International Conference on Computational Intelligence,*

Networks and Security (ICCINS), pp. 1–6, Mylavaram, India, December 2023. IEEE. ISBN 979-8-3503-1379-6. doi: 10.1109/ICCINS58907.2023.10450148.

Edison F. Naranjo-Espin, Johanna B. Arguero-Tello, Marcela Mosquera, and Gabriel A. Lóon-Paredes. Enhancing Fake News Detection through the Fusion of Classical and Advanced Machine Learning Models. In *2025 IEEE Ninth Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–4, Quito, Ecuador, October 2025. IEEE. ISBN 979-8-3315-5264-0. doi: 10.1109/ETCM67548.2025.11304354.

M. Zohaib Nawaz, M. Saqib Nawaz, Philippe Fournier-Viger, and Yulin He. Analysis and Classification of Fake News Using Sequential Pattern Mining. *Big Data Mining and Analytics*, 7(3):942–963, September 2024. ISSN 2096-0654, 2097-406X. doi: 10.26599/BDMA.2024.9020015.

Maryna Nyzova, Victoria Vysotska, Lyubomyr Chyrun, Zhengbing Hu, Yuriy Ushenko, and Dmytro Uhryn. Smart Tool for Text Content Analysis to Identify Key Propaganda Narratives and Disinformation in News Based on NLP and Machine Learning. *International Journal of Computer Network and Information Security*, 17(4):113.

Ammar Oad, Muhammad Hamza Farooq, Amna Zafar, Beenish Ayesha Akram, Ruogu Zhou, and Feng Dong. Fake News Classification Methodology With Enhanced BERT. *IEEE Access*, 12:164491–164502, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3491376.

Jinkyung Park, Rahul Dev Ellezhuthil, Joseph Isaac, Christoph Mergerson, Lauren Feldman, and Vivek Singh. Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling. In *Proceedings of the 15th ACM Web Science Conference 2023*, pp. 107–116, Austin TX USA, April 2023. ACM. ISBN 979-8-4007-0089-7. doi: 10.1145/3578503.3583617.

B. Praveen Kumar, K. Tamilarasi, and A. Thilagavathy. Metaheuristic-assisted deep learning model for fake news detection. *Journal of Experimental & Theoretical Artificial Intelligence*, 37(8):1481–1500, November 2025. ISSN 0952-813X, 1362-3079. doi: 10.1080/0952813X.2024.2440665.

Andhika Bayu Yudhistira Arda Putra, Yuliant Sibaroni, and Aditya Firman Ihsan. Disinformation Detection on 2024 Indonesia Presidential Election using IndoBERT. In *2023 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 350–355, Bandung, Indonesia, August 2023. IEEE. ISBN 979-8-3503-0519-7. doi: 10.1109/ICoDSA58501.2023.10277572.

Roopesh Ramesh, Sagar P Patil, and Tejasvi Bellubbi. Fake News Detection Using AI: A Hybrid Approach with ML, Deep Learning, and LLaMA Models. In *2025 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pp. 1–6, Mangalore, India, October 2025. IEEE. ISBN 979-8-3315-3898-9. doi: 10.1109/DISCOVER66922.2025.11258959.

Jaiwanth Reddy, Shikha Mundra, and Ankit Mundra. Ensembling Deep Learning Models for Fake News Classification. *Procedia Computer Science*, 235:2766–2774, 2024. ISSN 18770509. doi: 10.1016/j.procs.2024.04.261.

Rini Sovia, Dwi Andhara Valkyrie, Ruri Hartika Zain, and Firdaus. Language Processing for Detecting Fake News on Twitter Using a Long Short-Term Memory Architecture. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 9(4):935–943, August 2025. ISSN 2580-0760. doi: 10.29207/resti.v9i4.6570.

Muhammad Naim Syahmi Roslan, Masnizah Mohd, and Kiyoaki Shirai. Investigating the Performance of Machine Learning and Deep Learning Models in Fake News Detection. In *2024 16th International Conference on Knowledge and System Engineering (KSE)*, pp. 231–236, Kuala Lumpur, Malaysia, November 2024. IEEE. ISBN 979-8-3315-0940-8. doi: 10.1109/KSE63888.2024.11063621.

Aqeel Sahi, Mostfa Albdair, Mohammed Diykh, Shahab Abdulla, Hadi Alghayab, Kaled Aljebur, and Sarmad K.D. Alkhafaji. SGDM-GRU: Spectral graph deep learning based Gated Recurrent Unit model for accurate fake news detection. *Expert Systems with Applications*, 281:127572, July 2025. ISSN 09574174. doi: 10.1016/j.eswa.2025.127572.

Darshan Gangaram Sarkale, Vedant Jagdishbhai Gabani, Wandong Zhang, and Thangarajah Akilan. NLP-driven Content Classification Towards Fake News and Bully Detection. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, pp. 1–5, Mount Pleasant, MI, USA, September 2023. IEEE. ISBN 979-8-3503-2234-7. doi: 10.1109/AIBThings58340.2023.10292460.

Dilip Kumar Sharma, Priya Shrivastava, and Sonal Garg. Utilizing Word Embedding and Linguistic Features for Fake News Detection. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 844–848, New Delhi, India, March 2022. IEEE. ISBN 978-93-80544-44-1. doi: 10.23919/INDIACom54597.2022.9763294.

Jiao Shi, Xin Zhao, Nan Zhang, Yu Lei, and Lingtong Min. Rough-Fuzzy Graph Learning Domain Adaptation for Fake News Detection. *IEEE Transactions on Computational Social Systems*, 11(4):5275–5286, August 2024. ISSN 2329-924X, 2373-7476. doi: 10.1109/TCSS.2023.3312182.

Kai Shu. KaiDMML/FakeNewsNet, March 2019. URL `https://github.com/KaiDMML/FakeNewsNet`. original-date: 2017-08-03T22:56:48Z.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, 2018. Version Number: 3.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3):171–188, June 2020. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2020.0062.

M Jeevana Sujitha, Anjani Yalamanchili, Muhammed Abraar S A, Hari Haran P, B Uma Mahesh Babu, and Ennam Govinda. An Explainable AI Framework for Detecting Misinformation using Semantic NLP Features and Attention-based Models. In *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 450–455, Erode, India, August 2025. IEEE. ISBN 979-8-3315-1250-7. doi: 10.1109/ICSCDS65426.2025.11167535.

SushwanthReddy. Fake-News-Detection-using-Machine-Learning/FakeNewsDetection.ipynb at master · SushwanthReddy/Fake-News-Detection-using-Machine-Learning. URL `https://github.com/SushwanthReddy/Fake-News-Detection-using-Machine-Learning/blob/master/FakeNewsDetection.ipynb`.

Sanjaikanth E Vadakkethil Somanathan Pillai and Ashokkumar S. R. A Hierarchical Framework for Fake News Detection Using Semantic Analysis and Source Reliability. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 910–915, Greater Noida, India, March 2024. IEEE. ISBN 979-8-3503-7105-5. doi: 10.1109/ICDT61202.2024.10489413.

Fengjiao Zhou, Rui Ma, Mohamad Shaharudin Samsurijan, and Ying Yang. Recognizing fake news spreading in social networks using ensemble learning. *Journal of Information Technology & Politics*, pp. 1–19, March 2025. ISSN 1933-1681, 1933-169X. doi: 10.1080/19331681.2025.2458686.

Ana Đurić, Sara Koprivica, Danilo Nikolić, and Darko Stefanović. Assessing Reproducibility and Accessibility in Hate Speech and Fake News Detection Datasets: A Literature Review (2023-2024). In *2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–6, East Sarajevo, Bosnia and Herzegovina, March 2025. IEEE. ISBN 979-8-3315-1579-9. doi: 10.1109/INFOTEH64129.2025.10959261.

A Appendix

A.1 Evaluation Checklist for Machine Learning and Deep Learning Studies

Based on the recurring methodological, reporting, and evaluation deficiencies identified through the systematic analysis of the excluded studies, we propose a practical checklist to support the transparent, reproducible,

and methodologically sound design, implementation, and evaluation of empirical ML/DL studies. The checklist is intended to assist authors, reviewers, and editors in assessing the methodological quality, transparency, and reproducibility of empirical ML/DL research.

A.1.1 Dataset Transparency and Reporting

- ☐ Is the dataset clearly identified and referenced?
- ☐ If the dataset is referenced through a URL, is the link accessible and functional?
- ☐ If a newly created dataset is used, are dataset statistics reported for each class and for the dataset as a whole?
- ☐ If an existing dataset is used, are the reported dataset statistics consistent with the original source?
- ☐ If only a subset of the dataset was used, is the rationale for this decision clearly explained, and is the subset selection procedure sufficiently transparent to determine whether the samples were selected using an unbiased sampling strategy?
- ☐ If multiple classes are merged, are the mapping rules between original and final labels explicitly reported?
- ☐ If multiple datasets are combined, are the characteristics and contributions of each dataset reported separately?

A.1.2 Labeling and Task Definition

- ☐ Are the dataset labels clearly defined?
- ☐ Are the positive and negative classes explicitly specified?
- ☐ Is the interpretation of evaluation metrics consistent with the stated positive class definition?
- ☐ Is the prediction task clearly aligned with the labels provided by the dataset?

A.1.3 Experimental Design and Evaluation

- ☐ Is the train/validation/test partitioning strategy clearly reported?
- ☐ Is the confusion matrix consistent with the reported dataset split?
- ☐ Does the confusion matrix appear to be computed exclusively from the test set?
- ☐ Is there any indication of train–test contamination or data leakage?
- ☐ Are precision, recall, and F1-score values mathematically consistent with the reported confusion matrix?
- ☐ Are evaluation metrics reported using a consistent numerical format throughout the paper?
- ☐ Are exact numerical values provided for evaluation metrics rather than only graphical approximations?
- ☐ Are evaluation metrics reported beyond accuracy alone?

A.1.4 Model Transparency and Reproducibility

- ☐ Is the model architecture clearly described?
- ☐ Are the model components compatible with the characteristics of the dataset (e.g., image-processing modules used only when image data are present)?
- ☐ If pretrained models are used, are the model versions and corresponding references provided?
- ☐ Are the training and evaluation procedures described with sufficient detail for replication?

A.1.5 Additional Criteria for Fake News Detection Research

- ☐ Are the dataset labels clearly defined (e.g., veracity-based labels such as true/false or task-specific labels such as spam/ham)?
- ☐ Is the dataset aligned with an explicit definition of disinformation?
- ☐ Are satire, parody, political stance, propaganda, hate speech, toxicity, or other conceptually distinct phenomena clearly distinguished from the target labels?