

Machine Learning-Based Detection of Misinformation in Social Media Communication Networks

Dr. Niloofar Rahimi

Associate Professor, Center for Artificial Intelligence Research, Iran

Abstract

Social media communication networks have transformed the production, circulation, and consumption of information by enabling users to exchange messages rapidly across interconnected digital communities. Although these platforms support information accessibility and public communication, they also facilitate the circulation of inaccurate and misleading factual claims. The identification of misinformation presents significant computational challenges because message content is frequently brief, contextual information may be incomplete, and patterns of information dissemination vary across users and communication networks. Conventional approaches based exclusively on textual characteristics may overlook potentially informative relationships between messages, interaction patterns, and the temporal development of online discussions. The present paper develops a hypothetical methodological framework for examining machine learning-based misinformation detection in social media communication networks. The proposed investigation integrates textual feature analysis, network interaction characteristics, temporal propagation indicators, and supervised classification. Four computational approaches are considered: TF-IDF-based Logistic Regression, Linear Support Vector Machine, Random Forest with content and network features, and Gradient Boosting with integrated textual, structural, and temporal information.

A fictional dataset containing 12,000 social media messages associated with 600 independently verified factual claims is introduced to demonstrate the intended research design. Two illustrative analytical tables describe the hypothetical distribution of verified accurate and false claims and the comparative classification performance of the selected models. Precision, recall, and F1-score are considered the principal performance measures. All numerical values are manually constructed and do not represent collected social media records, trained classification models, or independently verified experimental outcomes. The methodological discussion examines claim verification, linguistic ambiguity, network dependency, label uncertainty, temporal data leakage, and responsible model interpretation. Particular attention is given to distinguishing misinformation from unverified rumors and avoiding unsupported assumptions concerning the intentions of individual users. The paper concludes that machine learning provides a relevant analytical foundation for misinformation detection, although trustworthy implementation requires reliable reference evidence, representative datasets, independent evaluation, and appropriate human review.

Keywords: Misinformation Detection, Social Media Networks, Machine Learning, Information Credibility, Natural Language Processing, Network Analysis, Rumor Verification, Gradient Boosting, Digital Communication, Classification Models.

1. Introduction

Social media communication networks have become important channels through which individuals, organizations, and communities exchange information. Digital platforms enable the rapid publication, redistribution, and discussion of messages across extensive networks of interconnected users. These environments support access to news, educational information, professional communication, and public discussion. However, the same communication characteristics can contribute to the circulation of inaccurate or misleading factual claims. Information can reach substantial audiences before its accuracy has been independently examined, creating difficulties for organizations responsible for evaluating information reliability. Consequently, automated methods for identifying potentially false claims have become an important subject within computational communication research.

Misinformation generally refers to false or misleading information, without necessarily establishing that the individual sharing it intended to deceive others. This distinction is important because the deliberate dissemination of false information involves additional questions concerning intention that cannot ordinarily be resolved through message classification alone. Similarly, rumors are claims whose accuracy may remain uncertain during an unfolding event. Some rumors are eventually confirmed, whereas others are disproved. Consequently, misinformation detection requires a clear definition of the classification target and an appropriate evidentiary basis for determining whether a factual claim is accurate or false.

Traditional information assessment relies on source evaluation, content examination, external evidence, and professional verification procedures. These practices can provide reliable judgments when sufficient evidence is available, but manual examination becomes difficult when extensive volumes of social media content appear within short periods. Machine learning offers computational approaches for organizing messages and identifying patterns associated with previously verified claims. Textual classifiers can examine word usage, linguistic structure, and semantic characteristics. However, message wording alone may provide insufficient information because accurate and inaccurate claims can employ similar vocabulary, emotional expressions, or persuasive communication styles.

Social media messages also circulate through interaction networks involving reposts, replies, mentions, and other forms of communication. These relationships create patterns that may provide additional information about the development and dissemination of factual claims. Network characteristics can include the timing of reposts, the distribution of interactions, and the structure of message-sharing pathways. Nevertheless, unusual popularity or rapid dissemination does not establish that information is false. Accurate information can spread rapidly, while misleading claims may remain relatively obscure. Therefore, network features should be considered supporting predictive information rather than substitutes for independent factual verification.

The present paper proposes a hypothetical comparative framework for detecting misinformation in social media communication networks using supervised machine learning. The methodology combines textual representations with early network and temporal indicators while maintaining a clear distinction between verified false claims and unresolved information. Four classification approaches are

considered to demonstrate how different feature configurations could be compared. A fictional dataset of 12,000 claim-bearing messages provides the basis for illustrative analysis. No actual platform data were collected, and no machine learning models were executed. The study instead establishes a methodological foundation for future empirical research involving responsible misinformation identification.

2. Objectives of the Study

The primary objective of this study is to examine the methodological application of machine learning for identifying misinformation within social media communication networks. The investigation focuses on textual evidence, network interactions, comparative classification performance, and evaluation reliability. Three specific objectives establish the analytical direction of the proposed research framework.

2.1 Examination of Linguistic and Contextual Misinformation Indicators

The first objective is to examine how textual characteristics of social media messages can be represented for misinformation classification. Word patterns, linguistic expressions, semantic relationships, and contextual references may contribute to predictive models trained on independently verified factual claims. The proposed framework considers the use of tokenization, term weighting, and selected language features for automated classification. Particular attention is given to distinguishing features that are statistically associated with verified false claims from characteristics that merely reflect conversational tone or writing style.

The investigation further aims to examine the limitations of interpreting linguistic characteristics without external evidence. A confident statement may be accurate, whereas an uncertain or emotionally expressive statement may also contain correct information. Similarly, quotations, irony, and the repetition of disputed claims can complicate automated analysis. Consequently, the proposed methodology emphasizes clearly defined labels and accurate identification of the claims asserted within each message. This objective seeks to prevent classifiers from treating linguistic style as direct proof of factual inaccuracy.

2.2 Comparative Evaluation of Content and Network-Based Models

The second objective focuses on comparing machine learning models using different combinations of textual, structural, and temporal information. Logistic Regression and Linear Support Vector Machine are selected as text-based classification approaches. Random Forest and Gradient Boosting are considered for models incorporating additional network interaction characteristics. The proposed comparison examines precision, recall, and F1-score under consistent evaluation conditions. Its purpose is to establish whether additional communication-network information could contribute useful predictive evidence beyond textual representations alone.

A further objective involves examining the timing and availability of network features. Social media interaction patterns develop as messages are redistributed and discussed, meaning that certain indicators become available only after a claim has circulated. A valid early-detection system must therefore restrict predictive information to a predefined observation period. The proposed framework considers features available within the first 60 minutes of claim circulation. This objective emphasizes

fair comparison and the prevention of information leakage from subsequent verification or propagation events.

2.3 Identification of Detection Reliability and Governance Constraints

The third objective is to identify methodological challenges that may affect misinformation classification in social media environments. These include inaccurate reference labels, duplicated claims, changing linguistic expressions, imbalanced class distributions, and dependencies among messages sharing the same factual assertion. Such characteristics can produce misleading performance estimates when evaluation procedures treat closely related observations as independent. The proposed investigation considers claim-level data separation, label verification, and appropriate classification metrics as potential safeguards.

The investigation also considers the responsibilities associated with applying automated misinformation detection to public communication. Incorrect classifications may create unnecessary scrutiny of accurate statements, while missed false claims may remain unidentified. Predictive systems therefore require transparent uncertainty reporting and an appropriate process for reviewing disputed decisions. This objective emphasizes that machine learning should support evidence-based information assessment rather than independently determine the credibility, intentions, or character of individual users.

3. Review of Literature

3.1 Early Research on Information Credibility and Online Rumors

Mendoza et al. (2010) examined information dissemination during an emergency and investigated differences in the propagation of rumors and confirmed information on Twitter. Castillo et al. (2011) subsequently analyzed credibility assessment using characteristics associated with social media content and communication behavior. Qazvinian et al. (2011) investigated rumor identification in microblogs using content-based and network-related information. These studies established important foundations for computational credibility analysis and demonstrated that the classification of disputed online information can benefit from multiple sources of evidence.

Gupta and Kumaraguru (2012) investigated the ranking of social media messages according to credibility during major events. Morris et al. (2012) examined how users assess the credibility of information presented through microblogging services. Their contributions highlighted the importance of message characteristics and contextual indicators, while also demonstrating that perceived credibility and verified factual accuracy are distinct concepts. Collectively, these studies support the development of classification procedures that use observable communication characteristics without treating perceived trustworthiness as a definitive measure of truth.

3.2 Network Propagation and Temporal Misinformation Patterns

Bakshy et al. (2012) investigated the role of social relationships in information diffusion and demonstrated the importance of network connections in the circulation of online content. Kwon et al. (2013) examined temporal, structural, and linguistic characteristics associated with rumor propagation. Friggeri et al. (2014) investigated rumor cascades and examined how information is redistributed through social networks. These contributions established a foundation for understanding how

communication structures and temporal processes may contribute to the identification of potentially disputed information.

Zhao et al. (2015) developed an approach for identifying emerging rumors through inquiry-related social media messages. Ma et al. (2015) examined time-varying social context information for rumor detection. Del Vicario et al. (2016) investigated the circulation of misinformation within online communities, while Zubiaga et al. (2016) examined how users respond to and redistribute rumors during developing events. These studies emphasized that network behavior contains meaningful relational information, although the identification of a disputed claim does not independently resolve whether that claim is accurate or false.

3.3 Machine Learning Classification and Evaluation Foundations

Breiman (2001) introduced Random Forest as an ensemble learning method capable of representing nonlinear relationships. Friedman (2001) developed Gradient Boosting through sequential predictive optimization. Pang et al. (2002) investigated machine learning approaches to textual classification, while Pennington et al. (2014) developed distributed word representations for natural language analysis. Kim (2014) examined convolutional neural networks for sentence classification. These contributions established methodological foundations for converting textual observations into predictive features and comparing statistical and neural learning approaches.

He and Garcia (2009) reviewed learning from imbalanced datasets, while Davis and Goadrich (2006) examined precision-recall evaluation. Saito and Rehmsmeier (2015) further emphasized the importance of precision-recall analysis when positive observations occur less frequently. Cawley and Talbot (2010) investigated bias arising from model selection. Ma et al. (2016) demonstrated recurrent neural approaches to rumor detection, and Shao et al. (2016) introduced a system for tracking online misinformation and related fact-checking activity. These contributions support careful evaluation, temporal analysis, and responsible interpretation of automated misinformation detection systems.

4. Research Methodology

4.1 Research Design and Classification Framework

The present paper adopts a hypothetical comparative quantitative research design for examining machine learning-based misinformation detection. Four model configurations are selected: TF-IDF-based Logistic Regression, TF-IDF-based Linear Support Vector Machine, Random Forest with content and network indicators, and Gradient Boosting with textual, structural, and temporal features. The intended target is binary classification of messages asserting independently verified accurate or false factual claims. Unresolved rumors, opinions, and satire without checkable factual assertions would be excluded from this binary evaluation.

The proposed analytical procedure consists of claim identification, reference verification, message preprocessing, network feature construction, supervised model development, and independent classification evaluation. Each model would be trained using only the information available within a predefined early observation period. Verification outcomes would provide retrospective target labels but would not be included as predictive features. Model parameters and classification thresholds would be selected using development information. Claim-separated chronological testing would support evaluation on assertions not represented within the training dataset.

4.2 Hypothetical Social Media Dataset and Feature Preparation

For illustrative purposes, the proposed study assumes a fictional dataset containing 12,000 social media messages associated with 600 distinct factual claims. Each claim is assumed to have 20 messages that explicitly assert its content within the first 60 minutes of circulation. The dataset contains 480 claims designated as independently verified accurate and 120 claims designated as independently verified false. These specifications are manually constructed and do not represent collected messages, real fact-checking decisions, or computationally generated observations.

The proposed preprocessing procedure would include message deduplication, claim-to-message alignment, text normalization, and removal of information unavailable at the prediction time. TF-IDF features would provide the principal linguistic representation. Additional characteristics would describe early repost activity, reply volume, interaction timing, and selected aggregate network properties. These features would avoid unnecessary personal information and would not treat user identity as proof of credibility. Messages disputing or correcting a claim would contribute contextual interaction information but would not automatically inherit that claim's false label.

4.3 Model Development and Performance Evaluation

The text-only Logistic Regression and Linear Support Vector Machine configurations would use equivalent TF-IDF representations. Random Forest would combine selected textual features with early network indicators. Gradient Boosting would additionally incorporate temporal propagation characteristics available during the first 60 minutes. Feature selection and parameter optimization would occur exclusively within development partitions. The proposed dataset would be separated chronologically at the claim level, ensuring that messages associated with the same claim remained within one partition.

The intended evaluation would report precision, recall, and F1-score for the verified-false category. Precision would indicate the proportion of messages flagged as misinformation that belonged to independently verified false claims. Recall would indicate the proportion of such messages successfully identified. F1-score would summarize their balance. A completed investigation should additionally report claim-level results, calibration, uncertainty estimates, and performance across different events. Because messages within a claim are dependent, statistical uncertainty should be evaluated using claim-level rather than naïve message-level resampling.

5. Analysis

5.1 Illustrative Distribution of Verified Information Categories

The distribution of accurate and false claims represents an important consideration when developing misinformation classification models. Social media datasets may contain unequal class frequencies, while messages associated with one widely circulated claim can contribute numerous correlated observations. Table 1 presents a hypothetical distribution of independently verified factual claims and corresponding messages. The example allocates 80% of claims to the verified accurate category and 20% to the verified false category.

These proportions are assumptions used exclusively for methodological demonstration. The illustrative distribution emphasizes the importance of separating related messages across the development and evaluation partitions. If messages repeating the same factual claim appeared in both

training and testing datasets, a model might recognize previously encountered wording rather than reliably classify unfamiliar claims. The proposed structure therefore assigns every claim to exactly one partition. The equal number of messages per claim and constant class proportions are simplifying assumptions, not realistic estimates of social media activity or misinformation prevalence.

Table 1: Hypothetical Distribution of Verified Social Media Claims and Messages

Information Category	Distinct Claims	Training Messages	Validation Messages	Test Messages	Total Messages
Independently Verified Accurate	480	6,720	1,440	1,440	9,600
Independently Verified False	120	1,680	360	360	2,400
Total	600	8,400	1,800	1,800	12,000

The fictional dataset contains 12,000 messages divided into 70% training, 15% validation, and 15% testing observations. The message labels are based conceptually on claim-level factual verification, not on sentiment, popularity, or assumptions about user intention.

5.2 Illustrative Comparison of Misinformation Detection Models

Misinformation classifiers can be compared by examining their ability to identify messages asserting verified false claims. Precision measures the reliability of positive predictions, while recall describes the proportion of false-claim messages correctly identified. Table 2 presents hypothetical classification counts for four machine learning approaches evaluated on the same fictional test set. The constructed test partition contains 360 messages asserting verified false claims and 1,440 messages asserting verified accurate claims. All model outcomes were manually selected for demonstration.

Table 2: Hypothetical Misinformation Classification Performance

Classification Model	True Positives	False Positives	Precision	Recall	F1-Score
TF-IDF Logistic Regression	216	144	0.600	0.600	0.600
TF-IDF Linear SVM	243	117	0.675	0.675	0.675
Content–Network Random Forest	270	100	0.730	0.750	0.740
Content–Temporal–Network Gradient Boosting	288	96	0.750	0.800	0.774

The hypothetical results demonstrate how different feature configurations could be compared through common evaluation measures. They do not establish that communication-network indicators reliably

distinguish false from accurate information or that a particular model should be deployed in a real moderation environment.

6. Discussion

6.1 Linguistic Characteristics and Factual Classification

Textual classification provides a computational approach for representing the linguistic characteristics of social media messages. TF-IDF features can capture patterns in word usage, while more extensive language representations may reflect semantic relationships among expressions. These characteristics can contribute to identifying messages resembling previously verified examples. However, factual accuracy is not determined solely by vocabulary, grammar, or writing style. A statement may be linguistically persuasive while inaccurate, or informally expressed while supported by reliable evidence.

The proposed framework therefore emphasizes the difference between learning associations and establishing factual truth. Models trained on verified claims may identify language patterns associated with specific information categories, but these patterns can change across topics and events. Moreover, sarcasm, quotation, and negation can complicate the interpretation of short messages. A system that labels a correction as misinformation merely because it repeats the wording of a false claim would produce a significant classification error. Reliable identification requires attention to asserted meaning and surrounding context.

6.2 Network Propagation and Misinformation Indicators

Social media network characteristics provide additional information about how factual claims circulate through connected users and discussion communities. Reposting frequency, reply patterns, and changes in interaction volume may describe the development of online information cascades. Research by Kwon et al. (2013) and Friggeri et al. (2014) demonstrates the relevance of temporal and structural characteristics to rumor propagation analysis. Nevertheless, similar communication patterns can occur around both accurate and false information.

The hypothetical comparison illustrates a possible situation in which combining network characteristics with textual features produces more favorable classification scores. However, this example cannot establish the explanatory contribution of any specific feature group. An actual study would need controlled ablation experiments comparing text-only, network-only, temporal-only, and combined configurations. Researchers should also distinguish genuine predictive information from features indirectly associated with particular topics or communities. Network activity should not be interpreted as direct evidence that an individual or group intentionally spreads misinformation.

6.3 Classification Performance and Responsible Interpretation

Classification performance must be interpreted in relation to the consequences of false-positive and false-negative predictions. Incorrectly labeling an accurate factual claim as misinformation can create unnecessary scrutiny and undermine confidence in the detection system. Failure to identify verified false information may reduce the usefulness of automated review. Consequently, precision and recall should be examined together, with decision thresholds selected according to the intended application and appropriate governance requirements.

The proposed framework treats machine learning predictions as indicators for further assessment rather than final factual judgments. Classification probabilities may assist reviewers in prioritizing claims requiring independent verification, but automated scores cannot replace supporting evidence. Genuine evaluation should examine uncertainty, disagreement among reference sources, and performance on unfamiliar topics.

The hypothetical results provide no basis for automatic removal of messages, attribution of deceptive intent, or claims concerning the credibility of individual users. Appropriate human review remains necessary when classification outcomes may influence public communication.

7. Challenges

7.1 Verification Quality and Ambiguous Factual Claims

Reliable misinformation classification depends on the availability of defensible reference labels. Social media messages may contain factual assertions that are partially accurate, incomplete, context-dependent, or impossible to verify at the time of publication. An unresolved claim should not automatically be classified as false because later evidence may establish its accuracy. Similarly, an opinion or prediction may not have a straightforward factual truth value. These distinctions complicate the construction of binary classification datasets and require careful annotation procedures.

Verification quality may also vary when sources disagree or when new evidence changes the assessment of a claim. A dataset constructed from inconsistent reference decisions can introduce contradictory training labels and weaken predictive reliability. Consequently, future research should document verification criteria, evidence sources, adjudication procedures, and the timing of factual assessments. Claims that remain uncertain should be represented separately rather than forced into a false category. The proposed framework recognizes reliable annotation as an essential requirement for meaningful classification.

7.2 Network Dependencies and Temporal Data Leakage

Social media observations are frequently connected through shared claims, repeated messages, and interaction networks. Conventional random partitioning can place closely related content in both training and testing datasets, producing optimistic performance estimates. A classifier may then recognize familiar claims instead of generalizing to previously unseen information. The proposed methodology addresses this concern through claim-level separation. However, additional dependencies may remain when similar narratives or overlapping events appear across different partitions.

Temporal data leakage introduces a related difficulty because network information changes as claims circulate. A model intended to identify misinformation shortly after publication cannot legitimately use engagement patterns that developed several hours later. Likewise, subsequent fact-checking outcomes must not be included as features during early prediction. The proposed framework restricts the available information to the first 60 minutes of circulation.

7.3 Evolving Communication Patterns and Interpretability

Social media communication practices change as platform features, user behavior, and information-sharing conventions develop. Linguistic expressions associated with one misinformation topic may provide limited predictive information for another. Network structures can also vary across communities

and communication environments. Consequently, a model trained using historical messages may become less reliable when applied to unfamiliar claims or later events. Effective evaluation requires testing across different periods and monitoring changes in classification performance.

Interpretability presents an additional challenge because machine learning predictions may depend on combinations of textual and network characteristics that are difficult to explain. A high misinformation score does not establish which factual evidence supports the classification. Similarly, engagement-based indicators may reflect popularity, coordinated discussion, or legitimate public interest. The proposed framework therefore emphasizes transparent feature documentation and the separation of prediction from verification. Future applications should provide reviewable evidence, acknowledge uncertainty, and avoid treating statistical associations as judgments about individuals.

8. Conclusion

Machine learning provides a methodological foundation for examining the identification of misinformation within social media communication networks. The present paper developed a hypothetical comparative framework integrating textual features, network interaction characteristics, and temporal propagation indicators. Logistic Regression, Linear Support Vector Machine, Random Forest, and Gradient Boosting were selected to illustrate alternative classification strategies. The proposed methodology emphasizes independently verified factual claims, claim-level data separation, and consistent evaluation procedures rather than treating message popularity or linguistic style as direct evidence of falsehood.

The illustrative tables demonstrate possible methods for reporting claim distributions and comparative misinformation classification performance. However, all numerical values were manually constructed and do not represent actual social media observations or completed computational experiments. Establishing reliable detection performance would require authentic datasets, documented verification procedures, implemented machine learning models, and independent evaluation. The hypothetical results therefore cannot demonstrate algorithmic superiority or justify operational decisions concerning real messages and users.

Overall, the study identifies verification uncertainty, network dependency, temporal information leakage, changing communication patterns, and model interpretability as important considerations in misinformation detection. Future empirical research should evaluate these factors using representative communication data and transparent experimental procedures. Machine learning may assist evidence-based information assessment when applied with suitable safeguards and professional review. Nevertheless, genuine conclusions concerning detection effectiveness must be established through independently verified evidence, reproducible analysis, and responsible interpretation of classification outcomes.

References

1. Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, 83, 51009–51067. <https://doi.org/10.1007/s11042-023-17470-8>
2. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. <https://doi.org/10.1145/3137597.3137600>
3. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). 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. <https://doi.org/10.1089/big.2020.0062>
4. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
5. Wang, W. Y. (2017). “Liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 422–426). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-2067>
6. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2931–2937). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1317>
7. Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 22–32). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1003>
8. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797–806). ACM. <https://doi.org/10.1145/3132847.3132877>
9. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80, 11765–11788. <https://doi.org/10.1007/s11042-020-10183-2>
10. Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Social Network Analysis and Mining*, 10, Article 82. <https://doi.org/10.1007/s13278-020-00696-x>
11. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), Article 109. <https://doi.org/10.1145/3395046>
12. Pal, A., Pranav, & Pradhan, M. (2023). Survey of fake news detection using machine intelligence approach. *Data & Knowledge Engineering*, 144, Article 102118. <https://doi.org/10.1016/j.datak.2022.102118>
13. Ali, I., Ayub, M. N. B., Shivakumara, P., & Noor, N. F. B. M. (2022). Fake news detection techniques on social media: A survey. *Wireless Communications and Mobile Computing*, 2022, Article 6072084. <https://doi.org/10.1155/2022/6072084>

14. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2), Article 32. <https://doi.org/10.1145/3161603>
15. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684). ACM. <https://doi.org/10.1145/1963405.1963500>
16. Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining* (pp. 1103–1108). IEEE. <https://doi.org/10.1109/ICDM.2013.61>
17. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv*. <https://arxiv.org/abs/1902.06673>
18. Ahmed, S., Hinkelmann, K., & Corradini, F. (2022). Combining machine learning with knowledge engineering to detect fake news in social networks—A survey. *arXiv*. <https://arxiv.org/abs/2201.08032>
19. Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3, 133–155. <https://doi.org/10.1016/j.aiopen.2022.09.001>
20. Touahri, I., & Mazroui, A. (2024). Survey of machine learning techniques for Arabic fake news detection. *Artificial Intelligence Review*, 57, Article 157. <https://doi.org/10.1007/s10462-024-10778-3>