Machine Learning-Based Detection of Fake News and Stock Market Manipulation: A Systematic Review

Shrey S[1], Jay A. Patel[2], Krishna J[3]
[1]Alumni of Atmiya University, Rajkot, Gujarat, India,
[2,3] Atmiya University, Rajkot,

***Abstract*—The spread of counterfeit news on the internet is increasingly becoming a significant risk to the integrity of the financial markets. Planned misinformation of equity, cryptocurrency, and commodity markets create unnatural price fluctuations and distort the trading volumes. This is a systematic review that brings together 30 peer-reviewed articles published in 2020-25 on the use of machine learning (ML) and natural language processing (NLP) in detecting fake news and identifying the presence of manipulated stocks on the stock market. In line with PRISMA standards, we answer three research questions: what are the most effective ML/NLP architectures to detect financial fake news; what market microstructure features go hand in hand with manipulation; and what knowledge gaps can be found in cross-modal and explainability-oriented methods. Transformer-based models, in particular, BERT and FinBERT, achieve better results than classical baselines of ML. The best detection performance is achieved with cross-modes that use signals, such as text, in conjunction with market data. Explainability, adversarial robustness, multilingual coverage, and real-time deployment are critical gaps that still exist.**

***Index Terms*—fake news detection, financial misinformation, machine learning, market manipulation, natural language processing, systematic review, transformer models**

I. INTRODUCTION

Connection between information and financial market is the basis of the efficient market hypothesis: The price of the asset is based on all publicly available information [1]. When information is made up selectively then this assumption is being critically undermined. Intentionally false or misleading information created with the purpose of deceiving others; fake news has ceased to be a political issue and has become a reported financial market risk with quantifiable economic impacts.

Clarke et al. [1] empirically proved that fake financial news can draw much investor interest compared to legitimate articles and has a quantifiable amount of abnormal trading volume. In January 2021, GameStop experienced a short squeeze led by organized Reddit actions that pushed the share price up as much as almost 400 dollars in several days [4]. According to a Chainalysis report, 24% of tokens issued in 2022 have pump-and-dump patterns [6]. These incidents underscore the need of automated detection systems.

Three advances have stimulated research: (i) massive pretrained language models, including BERT [7] and its financial adaptation FinBERT [8], [9]; (ii) labelled datasets like FakeNewsNet [10]; and (iii) the understanding that textual cues and market microstructure information are complementary detection evidence [2], [11]. The systematic review summarizes the state of the art in 2020-2025 and presents a concise research agenda.

The PRISMA flow diagram (Fig. 1) illustrates the study selection process across all database sources.

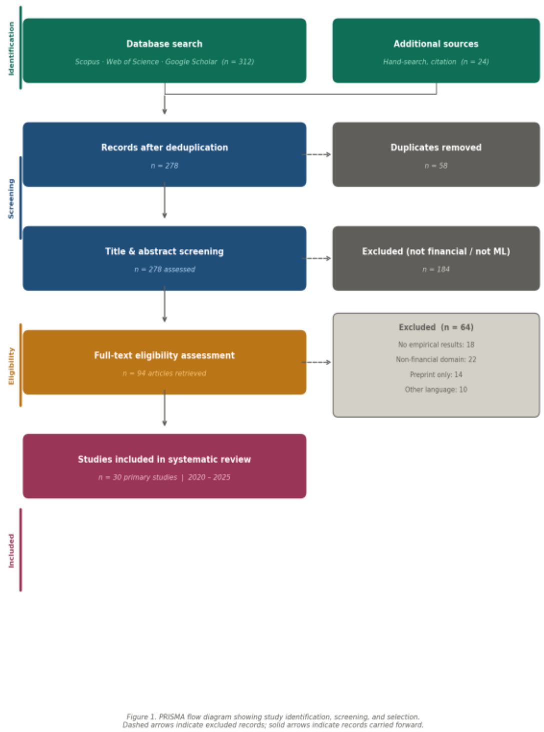

Fig. 1. PRISMA flow diagram showing study identification, screening, and selection.

II. REVIEW METHODOLOGY

A. Research Questions

There are three research questions (RQ) that are discussed in this review:

RQ1: What are the best ML/NLP architectures when it comes to the detection of fake news in financial settings?

RQ2: What are the market microstructure signs of misinformation-based manipulation?

RQ3: What are the methodological gaps of explainability, adversarial robustness, and real-time detection?

B. Search Strategy and Inclusion Criteria.

Scopus, Web of Science, and Google Scholar were searched with: fake news (or financial misinformation) + machine learning (or transformer) + stock market or market manipulation or cryptocurrency. The search was between January 2020 and April 2025. Inclusion criteria: (i) a computational system of identifying financial fake news or financial manipulation; (ii) empirical results on a given dataset; and (iii) publication in a peer-reviewed journal in English. The articles were filtered according to the fact that they were not non-financial articles, were only pre-print articles or articles that do not offer any empirical review.

The selection process based on the PRISMA (Fig. 1) allowed finding 312 records in database search and 24 in other sources. Prerequisites for full-textassessment 94 articles were retrieved after deduplication (n=278) and title/abstract screening (excluding 184). Those without a primary study were filtered off, leaving a final corpus of 30 primary studies.

III. LITERATURE SYNTHESIS

A. Publication Trends

Fig. 2 depicts the year and focus area distribution of the 30 included studies. The number of publications grew between 2020 and 2023 as a result of the synergistic stimulus of the COVID-19 infodemic, the GameStop incident, and the maturation of transformer models. The 2023 cohort contains the most studies, five fake news detection studies. There is an apparent drop in 20242025 that is due to search cutoff and not field deceleration.

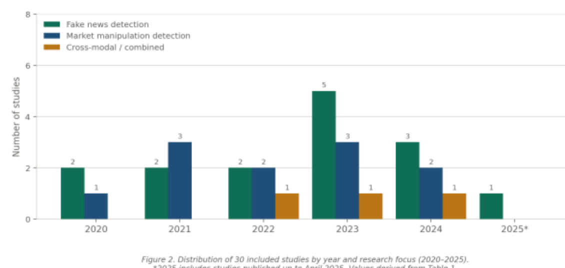

Fig. 2. Distribution of 30 included studies by year and research focus (2020–2025).

B. ML and NLP Methods (RQ1)

The frequency of ML/NLP methods is given in Fig. 3 and consists of 30 studies. In 11 studies, transformer based models, including BERT [7] and FinBERT [8], [9], [30] are the dominant paradigm. In 9 studies, we find Long Short-Term Memory (LSTM) and BiLSTM networks. Classical ML (Random Forest, XGBoost, SVM) are still applicable where their interpretability is needed, or small labelled datasets are needed. In 7 studies, statistical and econometric models are found, mainly in the financial economics stream.

Among the 30 studies, 22 use one or more neural architectures. None of them are large language model

(LLM) generation models, meaning that this frontier is still new. FinBERT has 89.7% negative sentiment detection accuracy compared to less than 60% with non-BERT algorithms [9].

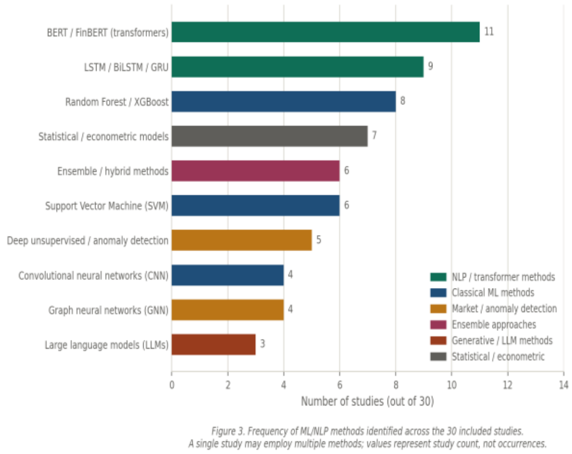

Fig. 3. Frequency of ML/NLP methods across 30 studies. A single study may employ multiple methods.

C. Market Microstructure Signals (RQ2)

Misinformation-driven manipulation accompanies three main market indicators in the studies reviewed: (i) abnormal trading volume (ATV) in the form of deviation of the rolling historical averages; (ii) abnormal price returns in the form of deviation of the asset pricing benchmark; and (iii) increasing bid-ask spreads in the form of liquidity deterioration.

Nghiem et al. [6] show that the use of social media signals together with market data is far more effective than text-only or market-only methods to detect cryptocurrency pump-and-dump actions. Anand and Pathak [4] demonstrate that intraday GameStop returns, volatility, and bid-ask spreads have short lags that are predicted by Reddit sentiment. The entire two-signal detection mechanism synthesised based on these results is shown in Fig. 5.

D. Research Gap Coverage (RQ3)

Fig. 4 plots coverage of research gaps by all 30 studies. The least studied dimension is adversarial robustness, which is studied by 2 out of 30 studies. Only 3 studies cover multilingual and non-English data set. Explainability: a fundamental regulatory provision in the EU AI Act and SEC Rule 10b-5 is only considered in 4 studies. Real-time streaming detection is found in only 5 studies.

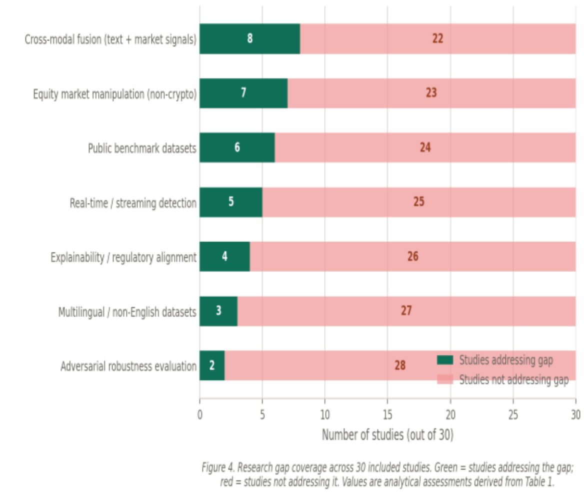

Fig. 4. Research gap coverage across 30 included studies. Green = studies addressing the gap; red = studies not addressing it.

E. Detection Mechanism

Fig. 5 shows the two-signal detection pipeline synthesised based on the literature reviewed. The perpetrator-to-market channel gets recorded by the information layer: fake stories are posted through social media and consumed by investors and converted into coordinated trading. Microstructure distortions are reflected in the market layer. The detection layer depicts NLP classifiers and market anomaly detectors processing every stream and cross-modal fusion emits a manipulation alert.

Fig. 5. Two-signal detection mechanism: fake-news-driven manipulation and the ML detection pipeline.

TABLE I SUMMARY OF 30 PRIMARY STUDIES INCLUDED IN THIS REVIEW

Ref.	Authors / Year	Venue	Method(s)	Domain	Key Focus
[1]	Clarke et al. (2021)	Inf. Sys. Res.	ML + Linguistics	Equities	Fake news & investor attention
[2]	Chung et al. (2023)	Inf. Sys. Frontiers	TRNN / LSTM	US Tech Stocks	Theory-driven disinformation
[3]	Digital Finance (2024)	Digital Finance	RF / NLP	Equities	P&D from online forums
[4]	Anand & Pathak (2022)	Econ. Letters	Text tone analysis	GameStop	Reddit & short squeeze
[5]	SEC Staff (2021)	U.S. SEC Report	Regulatory	US Equities	GameStop market structure
[6]	Nghiem et al. (2021)	Expert Syst. Appl.	ML + social signals	Cryptocurrency	P&D fraud detection
[7]	Devlin et al. (2019)	NAACL-HLT	BERT	General NLP	Foundational LM
[8]	Araci (2019)	arXiv	FinBERT	Financial text	Domain-adapted sentiment
[9]	Huang et al. (2023)	Contemp. Acctg.	FinBERT (LLM)	Financial text	LLM for financial NLP
[10]	Shu et al. (2020)	Big Data	Multi-source	General news	FakeNewsNet dataset
[11]	Khodabandehlou (2022)	Expert Syst. Appl.	Systematic review	Financial	Manipulation review
[12]	Chullamonthon (2023)	Expert Syst. Appl.	Ensemble DNN	SET Thailand	Supervised+unsupervised
[13]	Leangarun et al. (2021)	IEEE Access	Deep unsupervised	SET Thailand	Manipulation detection
[14]	La Morgia et al. (2023)	ACM Trans. Int.	Statistical + ML	Crypto	P&D analysis
[15]	Cartea et al. (2020)	Appl. Math. Finance	Math. models	Order-driven	Spoofing theory
[16]	Ardia & Bluteau (2024)	Int. Rev. Fin.	Statistical	Crypto	Twitter & P&D
[17]	Fantazzini & Xiao (2023)	Econometrics	ML + imbalanced	Crypto	P&D detection
[18]	Zhou & Zafarani (2020)	ACM Comp. Surv.	Survey	General news	FND survey
[19]	Vosoughi et al. (2018)	Science	Empirical	General news	True vs false spread
[20]	Arcuri et al. (2023)	J. Econ. & Bus.	Event study + ML	Equities	Fake news & returns
[21]	Ngo et al. (2021)	IEEE DASA	Review	Stock markets	AI manipulation review
[22]	Zulkifley et al. (2023)	CMC	Survey	Stock markets	AI manipulation survey
[23]	FinFakeBERT (2025)	Frontiers AI	BERT fine-tuning	Financial news	Domain shift in NLP
[24]	Sharma et al. (2022)	Ann. Oper. Res.	AI + ML	Supply chain	Fake news & supply
[25]	Alam et al. (2023)	Multimedia Tools	Survey ML/DL	General news	Comprehensive FND
[26]	Sultana et al. (2024)	J. AI Finance	Reg. LSTM	General news	LSTM fake news
[27]	Liu et al. (2024)	AI Finance Conf.	FMDLlama (LLM)	Financial news	LLM misinformation
[28]	Lyócsa et al. (2022)	Finance Res. Lett.	Statistical	GameStop	YOLO trading

[29]	Pedersen (2022)	J. Fin. Econ.	Theoretical	US Equities	Social networks
[30]	Liu et al. (2020)	IJCAI	FinBERT pre-train	Financial text	Financial LM

IV. CRITICAL ANALYSIS

A. Methodological Convergence

The literature reviewed demonstrates a general agreement on the use of transformer-based architectures as the paradigm of textual classification, and cross-modal fusion as the practice of the best. Out of the 30 studies, 22 use at least one neural architecture. Any new financial text classification system should be considered FinBERT-family models as the standard baseline [9].

B. Explainability Gap

A total of 4 out of 30 studies reviewed have any explainability mechanism. Accountability requirements regulatory frameworks, such as SEC Rule 10b-5, EU Market Abuse Regulation, and the EU AI Act, are not covered by most systems reviewed. The detection models should be able to give interpretable, auditable evidence to facilitate enforcement. The SHAP or LIME attribution frameworks should be incorporated in future work [11].

C. Adversarial Robustness

Adversarial robustness is only considered in 2 out of 30 studies. Since detection systems are increasingly sophisticated, there are significant incentives to change on the part of the perpetrators. According to the FinFakeBERT research [23], domain shift has been cited as a similar problem, yet a conscious adversarial attack has been all but neglected in the financial sector. Red-teaming evaluation procedures and adversarial training should be a norm.

D. Multilingual and Cross-Market coverage.

Twenty-eight studies out of 30 are English content and US or Western markets. Manipulation plans are recorded all over the world, especially in Southeast Asia [12], [13] and new cryptocurrency markets. The priorities of research are multilingual FinBERT implementations and international datasets.

E. Real-Time Detection

The majority of systems reviewed are assessed backwards on previous datasets. The limited amount of studies which touch on the temporal dynamics [2], [4] show that the time interval between the misinformation publication and the market reaction is calculated in minutes. Latency of detection ought to be established as a regular metric of evaluation with accuracy and F1-score.

V. DISCUSSION

This review contributes to theory in three main ways. To begin with, it confirms that the fake news detection and market manipulation detection literatures have a similar empirical basis: reported co-occurrence of misinformation dissemination and abnormal market behaviour [1], [4], [6], [20]. Second, the hypothesis of cross-modal complementary is validated: text and market microstructure indicators are independent and complementary pieces of evidence [2], [6], [11]. Third, the review also reveals the structural mismatch between academic optimisation goals (retrospective accuracy) and operational deployment aspects (real-time, explainable, adversarially robust detection).

To practitioners, any new financial text classification system should take the FinBERT-family transformers as the baseline to any older lexicon-based and classical ML system. To regulators, the GameStop literary evidence [4], [5], [28], [29] justifies investing in social media sentiment monitoring as an early-warning tool to complement the already established market surveillance infrastructure.

VI. FUTURE RESEARCH DIRECTIONS

This systematic review leads to 5 priority research directions:

1) Explainable architectures: This category of systems is designed to be interpretable to regulators, and includes SHAP-based attribution and audit-trail generation.

2) Adversarial training and red-teaming: regular testing on designed avoidance attempts and adversarially resistant training methods.

3) Multilingual benchmarks: labeled financial fake

news corpora in other non-English languages along with the market microstructure information.
4) Evaluation frameworks on streaming: measurement of detection latency, streaming data protocol, and real-time alert generation metrics.
5) Detection using LLM-based detection with hallucination safeguards: leveraging large language models [27] but enacting safeguards against fabrication of facts.

VII. CONCLUSION

The systematic review gathered 30 peer-reviewed articles (20202025) about ML and NLP to identify financial fake news and stock market manipulation based on these techniques. Transformer-based models - especially FinBERT adoptions - have become the new paradigm, and cross-modal integration of textual and market microstructure signals have become a proven best practice.
In real-world situations, the GameStop short squeeze [4], [5], [28], cryptocurrency pump-and-dump campaigns [6], [14], [16] and recorded instances of fake financial news producing quantifiable changes in stock prices [1], [20] have both supplied motivation and the labelled data of the field. The field is currently limited by four critical gaps: explainability, adversarial robustness, multilingual coverage, and real-time deployment. The main issue of the new generation of research in the area of financial market surveillance is the need to fill these gaps.

ACKNOWLEDGMENT

The authors thank the reviewers and editors for their constructive feedback. No external funding was received for this research.

CRediT authorship contribution statement
Shrey: Writing – original draft, Conceptualization, Validation, Software, Methodology, Formal analysis, Data curation.Krishna: Writing – review & editing, Resources, Methodology, Investigation.

REFERENCES

[1] J. Clarke, H. Chen, D. Du, and Y. J. Hu, "Fake news, investor attention, and market reaction," Inf. Syst. Res., vol. 32, no. 1, pp. 35–52, 2021.
[2] W. Chung, Y. Zhang, and J. Pan, "A theory-based deep-learning approach to detecting disinformation in financial social media," Inf. Syst. Front., vol. 25, pp. 473–492, 2023.
[3] "Detecting pump & dump stock market manipulation from online forums," Digit. Finance, 2024. doi: 10.1007/s42521-024-00121-4
[4] A. Anand and J. Pathak, "The role of Reddit in the GameStop short squeeze," Econ. Lett., vol. 211, p. 110249, 2022.
[5] U.S. Securities and Exchange Commission, "Staff report on equity and options market structure conditions in early 2021," SEC, Washington, DC, 2021.
[6] H. Nghiem, G. Muric, F. Morstatter, and E. Ferrara, "Detecting cryptocurrency pump-and-dump frauds using market and social signals," Expert Syst. Appl., vol. 182, p. 115284, 2021.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[8] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv:1908.10063, 2019.
[9] A. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," Contemp. Account. Res., vol. 40, no. 2, pp. 806–841, 2023.
[10] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information," Big Data, vol. 8, no. 3, pp. 171–188, 2020.
[11] S. Khodabandehlou and S. A. H. Golpayegani, "Market manipulation detection: A systematic literature review," Expert Syst. Appl., vol. 210, p. 118330, 2022.
[12] P. Chullamonthon and P. Tangamchit, "Ensemble of supervised and unsupervised deep neural networks for stock price manipulation detection," Expert Syst. Appl., vol. 220, p. 119698, 2023.
[13] T. Leangarun, P. Tangamchit, and S. Thajchayapong, "Stock price manipulation detection using deep unsupervised learning: The case of Thailand," IEEE Access, vol. 9, pp. 106824–106838, 2021.

[14] M. La Morgia, A. Mei, F. Sassi, and J. Stefa, "The doge of wall street: Analysis and detection of pump and dump cryptocurrency manipulations," ACM Trans. Internet Technol., vol. 23, pp. 1–28, 2023.
[15] A. Cartea, S. Jaimungal, and Y. Wang, "Spoofing and price manipulation in order-driven markets," Appl. Math. Finance, vol. 27, no. 1–2, pp. 67–98, 2020.
[16] D. Ardia and K. Bluteau, "Twitter and cryptocurrency pump-and-dumps," Int. Rev. Financ. Anal., vol. 95, 2024.
[17] D. Fantazzini and Y. Xiao, "Detecting pump-and-dumps with crypto-assets: Dealing with imbalanced datasets," Econometrics, vol. 11, no. 3, 2023.
[18] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Comput. Surv., vol. 53, no. 5, p. 3395046, 2020.
[19] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.
[20] M. C. Arcuri, G. Gandolfi, and I. Russo, "Does fake news impact stock returns? Evidence from US and EU stock markets," J. Econ. Bus., vol. 125, p. 106130, 2023.
[21] M. H. Ngo and M. Karamanoglu, "Stock market manipulation detection using artificial intelligence: A concise review," in Proc. IEEE DASA, 2021.
[22] M. A. Zulkifley, A. F. Munir, A. Sukor, M. Edil, and M. H. Mohd Shafiai, "A survey on stock market manipulation detectors using artificial intelligence," Comput. Mater. Contin., vol. 75, pp. 4395–4418, 2023.
[23] "FinFakeBERT: Financial fake news detection," Front. Artif. Intell., 2025. doi: 10.3389/frai.2025.1604272
[24] R. Sharma, V. Agarwal, A. Rasool, and S. O. Shim, "Detecting fake news and disinformation using artificial intelligence and machine learning," Ann. Oper. Res., 2022.
[25] S. Alam, N. Yao, and D. Bhatt, "A comprehensive survey on machine learning approaches for fake news detection," Multimedia Tools Appl., 2023.
[26] T. C. Sultana, F. R. Fahim, and M. M. Anwar, "A regularized LSTM method for detecting fake news articles," J. AI Finance, vol. 29, no. 4, pp. 332–350, 2024.
[27] Z. Liu, X. Zhang, K. Yang, Q. Xie, J. Huang, and S. Ananiadou, "FMDLlama: Financial misinformation detection based on large language models," in Proc. Int. Conf. AI Finance, 2024.
[28] S. Lyócsa, E. Baumöhl, and T. Výrost, "YOLO trading: Riding with the herd during the GameStop episode," Finance Res. Lett., vol. 46, p. 102359, 2022.
[29] L. H. Pedersen, "Game on: Social networks and markets," J. Financ. Econ., vol. 146, no. 3, pp. 1097–1119, 2022.
[30] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A pre-trained financial language representation model for financial text mining," in Proc. IJCAI, 2020.