IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 December 2024, accepted 23 December 2024, date of publication 16 January 2025, date of current version 29 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3530688

SURVEY

Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact Through Machine Learning

DARIUS PLIKYNAS, IEVA RIZGELIENĖ, AND GRAŽINA KORVEL, (Member, IEEE)

Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, 10257 Vilnius, Lithuania

Corresponding authors: Darius Plikynas (darius.plikynas@mif.vu.lt) and Ieva Rizgelienė (ieva.rizgeliene@mif.vu.lt)

This work was supported by the Lithuanian Government Priority Research Program "Building Societal Resilience and Crisis Management in the Context of Contemporary Geopolitical Developments" (Implemented through the Lithuania Research Council) under Grant S-VIS-23-8.

ABSTRACT In recent years, the world has witnessed a global outbreak of fake news, propaganda and disinformation (FNPD) flows on online social networks (OSN). In the context of information warfare and the capabilities of generative AI, FNPDs have proliferated. They have become a powerful and quite effective tool for influencing people's social identities, attitudes, opinions and even behavior. Ad hoc malicious social media accounts and organized networks of trolls and bots target countries, societies, social groups, political campaigns and individuals. As a result, conspiracy theories, echo chambers, filter bubbles and other processes of fragmentation and marginalization are polarizing, radicalizing, and disintegrating society in terms of coherent politics, governance, and social networks of trust and cooperation. This systematic review aims to explore advances in using machine and deep learning to detect FNPD in OSNs effectively. We present the results of a combined PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) review in three analysis domains: 1) propagators (authors, trolls, and bots), 2) textual content, 3) social impact. This systemic research framework integrates meta-analyses of three research domains, providing an overview of the wider research field and revealing important relationships between these research domains. It not only addresses the most promising ML/DL research methodologies and hybrid approaches in each domain, but also provides perspectives and insights on future research directions.

INDEX TERMS Machine learning, deep learning, propaganda and disinformation, fake news, PRISMA systematic review, authors' analysis, content analysis, social impact analysis.

I. INTRODUCTION

A. CONTEXT AND ACTUALITY

In the course of preparing this systemic overview, we have come across different interpretations of related terminology. This is why we will briefly outline here the interpretation of the key terms used in this systematic review. The broad term 'Fake News' often encompasses all three other terms (misinformation, disinformation, and propaganda). Propaganda refers to information used with a specific intent to manipulate and influence [1]. It is intentionally disseminated to promote a particular political cause or point of view for persuasion and manipulation audience's emotions, attitudes, and behaviors. Disinformation, which is deliberately false information spread to deceive, can often be considered a form of propaganda when it serves a specific political or ideological agenda [2], [3], [4], [5].

Thus, this systemic review attempts to cover the broad, controversial, and complex field of fake news (FN), propaganda (P), and disinformation (D) research in OSNs (Online Social Networks). In fact, we deliberately pay more attention to the disinformation aspect, which can be expressed $FN \cap P \supseteq D$ or FNPD for short.

The associate editor coordinating the review of this manuscript and approving it for publication was Bang L. H. Nguyen.

We interpret the term 'information war' as a strategic wider conflict where information and communication technologies are used to gain advantage over an adversary. It involves manipulating and controlling information to influence public opinion, undermine opponents, or achieve political goals. This can include spreading propaganda and disinformation, hacking, cyberattacks, and other tactics to disrupt or distort information flows. Information wars can have significant social, political, and economic consequences, impacting trust in institutions, democratic processes, and even national security [6], [7].

The other closely related term, used in the EU Research Agenda, is Foreign Information Manipulation and Interference (FIMI), which refers to deliberate and coordinated efforts by state or non-state actors to manipulate the information environment in order to achieve political, security or other strategic objectives. It is characterized as a pattern of behavior that threatens or has the potential to negatively affect values, procedures and political processes.

Social media users increasingly become the target of FNPD and FIMI activities aimed at influencing their perception of reality [8]. A crucial element is the filter bubble (content selection and personalization by algorithms) or echo chamber situation (a social and behavioral phenomenon where individuals actively seek out and interact with like-minded people, reinforcing their existing beliefs) [8], [9].

Social media and online social networks (OSN) has become an influential source of political disinformation and propaganda, often spread by malicious actors such as trolls, bots, or disguised foreign intelligence services, as was famously the case during the 2016 US election. However, what makes social media a particularly potent vector for disinformation is not only the behavior of malicious actors themselves but also OSNs, where opinion leaders and ordinary users play a crucial role in spreading and amplifying FNPD [10]. This is the main reason why in this systematic review we sought to include ML studies that analyze the authors and disseminators of FNPD.

Another important aspect of this research is the search for effective ML methods that can detect sources of FNPD at a very early stage. A major challenge in the early detection of fake news is to fully exploit the limited data observed in the initial stages of FNPD news propagation [11]. FNPD are also evolving and becoming more sophisticated, using fake accounts, AI-generated content, and bots to spread their messages at scale.

Operations of influence today target people within a society, influencing their beliefs as well as their behavior and eroding trust in government and public institutions. Adversaries of democracies now seek to control and exploit the trending mechanism on social media to inflict damage, discredit public and private institutions, and sow domestic discord [2]. Socio-political cleavages are key to increasing the likelihood of domestic political instability, including atrocities. These include significant social and political polarization, anti-democratic or weakened democratic regimes, and severe governance or security crises.

This problem has arisen due to the emergence of several concomitant phenomena, such as 1) the digitization of human life and the ease of news dissemination through OSN applications; 2) the availability of ''big data'' that allows the customization of news feeds and the creation of polarized so-called ''filter bubbles'' and ''echo chambers''; and 3) the rapid progress of generative machine learning (ML) and deep learning (DL) algorithms in the creation of realistic-looking yet fake digital content (such as text, images and videos) [12].

While fact-checking websites such as Snopes, PolitiFact, and major companies such as Google, Facebook, and Twitter have taken initial steps to address FNPD, much more remains to be done [13]. As an interdisciplinary topic, different facets of fake news have been studied by communities as diverse as machine learning, databases, journalism, social science, psychology, cognitive science, political science, and many more. In this systemic review, we focus on studies that use ML and DL approaches, addressing FNPD analysis in three research domains:

domain#1: authors/spreaders,
domain#2: textual content,
domain#3: social impact.

Such a broad systemic review framework is, to our knowledge, unique. It provides a much broader and more comprehensive view of this field of research.

Content analysis of FNPD news is, of course, a basic element in this field of research. However, it is important to note that an increasing number of studies are also integrating author and content analysis. This improves not only the detection of FNPDs, but also the estimation of the social impact. This is a less explored area of flagman research in terms of ML/DL deployment and is more challenging as FNPD social impact metrics is not well defined yet. However, we see it as a much-needed niche of research, ranging from the study of media networks, clustering, development of echo chambers and filter bubbles to the study of social impact dynamics in terms of online social network support, civic engagement, personal relationships, trust, and cooperation, etc. [14], [15], [16], [17], [18], [19], [20], [21].

After this brief introduction, which has touched more on the relevance of the topic, the following two subsections outline the work of previous systematic reviews and the structure and scope of this study.

B. PREVIOUS SYSTEMATIC REVIEWS

Below we briefly present some key previous literature reviews in the three FNPD research domains: (i) authors/disseminators analysis, (ii) content analysis, and (iii) social impact analysis.

1) FNPD AUTHORS/DISSEMINATORS (DOMAIN#1)

In the context of FNPD, authors' and disseminators' data analysis can be approached in various ways, such as detecting

the primary source of FNPD and measuring its trustworthiness, analyzing news dissemination patterns, or identifying malicious accounts on social media. For instance, study [20] delves into multiple methods for identifying sources within OSNs from different perspectives, evaluating key factors for source detection. Meanwhile, [22] focuses on user trust in social media by examining critical factors such as profile information and user actions. An extensive analysis of anomaly detection in OSNs is provided in [23], which investigates various deep-learning methods to identify unusual behavior and structural transformations. Additionally, comprehensive research on malicious social bot detection [24] outlines various network-based detection approaches, highlighting their strengths and weaknesses.

The foundational theories of fake news in OSNs and various perspectives on its attribution and analysis are explored in [25]. This study examines fake news types, attributes, and dissemination mechanisms, emphasizing the role of creators, including both genuine users and bots. It evaluates different detection techniques, such as linguistic and semantic analysis, along with machine learning and deep learning approaches, highlighting the importance of interdisciplinary collaboration. Similarly, [1] analyze the impact of social bots in propagating false information on OSNs, discussing their tactics and identification approaches, and underscoring the challenges posed by fake news in rapidly growing online social networks.

Existing systematic reviews emphasize the importance of integrating data from authors, creators, and dissemination patterns for more effective FNPD detection, primarily by analyzing the features used. However, an in-depth analysis of the specific characteristics of authors and disseminators is often lacking. This paper addresses this gap by providing a comprehensive analysis of FNPD detection with an explicit focus on the role of authors and disseminators. It explores how authors' content and disseminators' engagement patterns can be used to strengthen FNPD detection. Such analysis is valuable for developing more accurate and reliable detection models that can better capture the nuanced behaviors and complex interactions involved in the dissemination of FNPD.

2) FNPD CONTENT RESEARCH (DOMAIN#2)

Many survey papers in the literature cover various aspects of content analysis and classification using ML and DL methods with a focus on FNPD. Studies have focused on three main aspects: data collection, feature extraction, and classification algorithms. Several main trends can be identified. The review examines the classification algorithms, noting that previous work has dealt only with classical machine learning methods [26], [27] or classical methods with a limited focus on deep learning [28]. This was due to the significant number of studies that used classical machine learning for automatic fake news detection, and only a few studies that used deep learning for automatic feature classification in fake news detectors. Later, as deep learning became more popular, reports on the use of deep learning appeared [29], [30]. The paper [31] classifies and evaluates the results of ML and DL algorithms, highlighting their high accuracy rate in detecting fake news, which is a valuable benchmark for researchers. Several studies have looked at specialized network types to improve detection accuracy. A notable example is the comprehensive review by [32], which focuses on the use of graph neural networks (GNNs) for this purpose. The authors classify different GNN architectures and explore their effectiveness in interpreting the complex and often hidden patterns in data. Methods for combining human expertise with machines learning systems are also reviewed [33]. The review [34] extends to multimodal deep learning approaches that use both textual and non-textual data (e.g., images and videos) to improve the accuracy and reliability of fake news detection systems.

In the context of feature extraction, aspects of how natural language processing (NLP) methods could be used to evaluate information from OSNs are discussed. The analysis of textual representations from linguistic context, psycholinguistic factors, and syntactic and semantic analysis is provided [35], [36]. NLP techniques such as data preprocessing, data vectorization, and feature extraction are discussed in detail, and their advantages and disadvantages are described [29]. An attempt is made to systematically enumerate the main algorithms for each step involved in an NLP system [37]. It is worth noting that most papers focus on NLP as a subset of FNPD detection, and only a few consider its interaction with ML/DL techniques, noting that the effectiveness of fake news detection systems depends on the careful selection and use of the most powerful content-based models and features [29], [38]. A comprehensive taxonomy of machine learning and deep learning models and features specifically used in content-based fake news detection is presented in [38].

An attempt is also made to review the selected literature that focuses on current datasets for training and testing fake news discrimination training [29], [31], [37]. In addition to listing the datasets, their characteristics are also discussed, which may influence the choice of machine learning models and features [39]. The effectiveness of these models and features is carefully compared across studies to identify the most effective approaches [38]. Our systemic review paper partially covers these papers [29], [38]. We have also reviewed the different learning algorithms with respect to the text preparation phase. However, we also include a combined approach, investigating whether authors integrate disseminator and social impact information into FNPD content detection and how these approaches differ from those that only analyze content.

This article also discusses the challenges and developments in real-time fake news detection, which is particularly relevant for social media platforms where news spreads rapidly. As part of this discussion, it provides a novel perspective by examining the intersection of logical fallacies and FNPD

detection. Logical fallacies are a critical aspect of manipulative techniques that undermine logical reasoning and exploit audience vulnerabilities. While logical fallacies are often used to create misleading narratives and have some overlap with propaganda techniques, they are often treated as a separate research area from FNPD detection. This paper focuses on exploring the intersection between these two areas, specifically examining how logical fallacies are addressed in the context of FNPD detection.

It is important to note that although large language models (LLMs) such as GPT-3 or BERT, traditionally used for end-to-end tasks, are very promising for various NLP tasks due to their deep contextual understanding, they are deliberately not included in this review. Only embedding methods derived from these models are considered, focusing on how they can be integrated into different classification frameworks without directly using the full models themselves. This limitation is due to the desire to explore how narrower and more computationally efficient ML/DL methods can be used effectively to combat FNPD.

3) FNPD SOCIAL IMPACT (DOMAIN#3)

Several previous systemic reviews have analyzed the social implications of the FNPD. For example, Ahmed et al. [27] and Ahsan et al. [40] proposed that the integration of machine learning and knowledge engineering can be helpful in detecting the impact of fake news on different domains and society in general. In this work, they summarize and present the efforts and achievements in combating the spread of rumor information. Choraś et al. [35] and Varlamis et al. [41], were concerned with the directions of application of intelligent systems in the detection of misinformation sources or use Graph Convolutional Networks (GCNs) for the task of detecting fake news, fake accounts, and rumors spreading in OSNs. Figueira et al. [42] and Kumar and Shah [43] focus on content analysis, network propagation, fact-checking, fake news analysis and emerging detection systems in their surveys and discuss the reasons behind successful deception. They present various aspects of fake information, namely the actors involved in spreading fake information, the rationale behind successfully deceiving readers, quantifying the impact of fake information, measuring its characteristics across different dimensions, and finally algorithms developed to detect fake information.

Abbas [44] provides an overview of the state of the art in different applications of OSN analysis using deep learning techniques. He considers applications such as opinion analysis, sentiment analysis, text classification, recommender systems, structural analysis, anomaly detection, and fake news detection. He compares different schemes based on the focus and characteristics of the papers. Similarly, Chaabene et al. [23] provided an overview of several methods that aim to solve the problem of detecting abnormal behavior in social media. They distinguished three diverse types of anomalous behavior: structural methods based on the analysis of graphs of OSNs, behavioral methods based on the extraction and analysis of user activities, and hybrid methods that combine the two types of methods mentioned above. Aïmeur et al. [1] aim to provide a comprehensive and systematic review of fake news research as well as a fundamental review of existing approaches used to detect and prevent fake news from spreading via OSNs. Rum et al. [45] examined computing methods and approaches employed by the existing works for identifying political polarization in social media. Mahmoudi et al. [46] identify terminology, examine the effects of echo chambers, analyze approaches to echo chamber mechanisms, assess modeling and detection techniques, and evaluate metrics used to specify echo chambers in online OSNs.

C. THE SCOPE

The structure of this systematic review article is designed to cover and explore in parallel three adjacent domains of FNPD research (domain#1: authors/disseminators, domain#2: textual content, domain#3: social impact) that make use of state-of-the-art ML/DL methods. In this paper, these domains are examined in parallel, from the introduction to the conclusions, where we highlight their interrelationships and interdependencies [47], see Fig. 1. This figure shows the structure of the full systematic review article: main sections, sub-sections and appendixes (A, B, C, D, E and F), which contain systematic visualizations (charts and tables).

The research scope of this study is centered on the following five key research questions:

RQ 1: What machine learning approaches have been used in recent studies to define FNPD disseminators in OSNs?

RQ 2: What machine learning advances are being used to analyze the textual content of FNPD?

RQ 3: What machine learning and related approaches are used to model the social impact of FNPD?

RQ 4: How well integrated is the combined analysis of FNPD disseminators, content and social impact?

RQ 5: Is it possible in the future to have a fundamental FNPD model trained on large FNPD datasets covering a wide range of research fields and scientific disciplines, using combined and hybrid ML/DL approaches?

Key contributions outlined in this systemic review are as follows:

(i) The article provides a systemic review of research in the analysis of fake news, propaganda, and disinformation (FNPD), using a combined approach that includes the study of authors, content, and social impact, while applying advanced machine learning (ML) and deep learning (DL) methodologies.

(ii) Integration of three systemic reviews instead of one gave advantages such as a) in-depth analysis of the different approaches used in each research domain, b) specialized criteria-based meta-analyses in each research domain, c) the convenience for the reader of having three thematic related systemic reviews in one place.

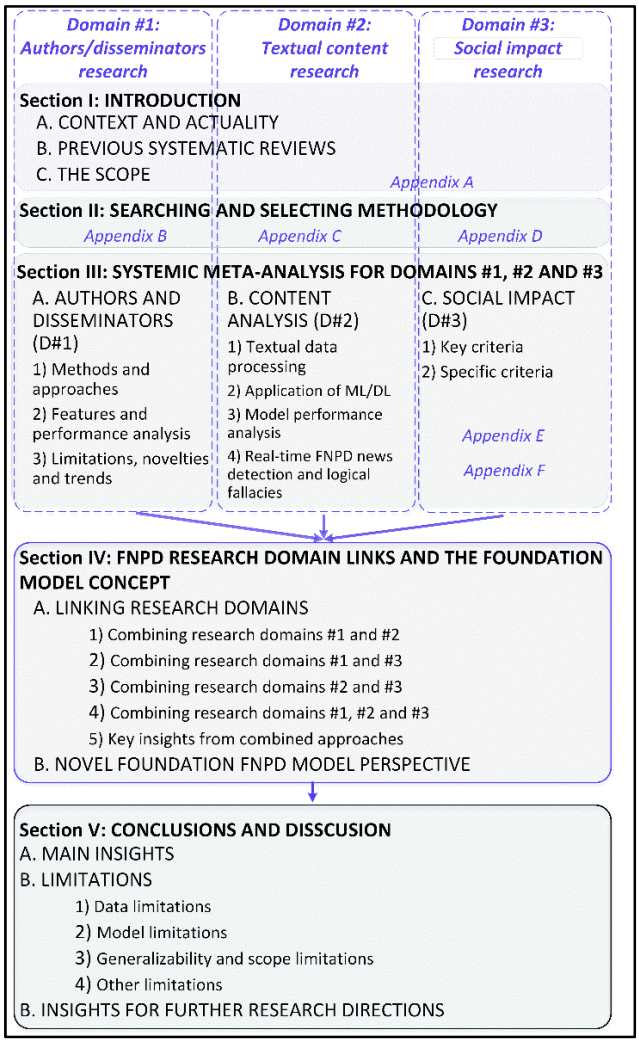


FIGURE 1. Structure of the systematic review according to the three study domains (1, 2 and 3) and sections.

(iii) In general, our systematic review indicates that the most effective FNPD detection methods are those that integrate data from multiple sources, including authors, content, and OSN analysis. These combined approaches outperformed methods that rely on a single data source. This trend underscores the importance of multifaceted data analysis in improving FNPD detection and is expected to continue to shape research directions in this area.

(iv) The fight against FNPD on social media platforms is being tackled on multiple fronts, focusing on early detection techniques, broad system development (e.g., FakeNewsTracker), and external influences (e.g., Russian trolls), shifting towards more technical and AI-driven solutions (e.g., geometric deep learning, social bot detection, network-based patterns analyses, and employing deep sociological insights.

(v) To operationalize how and which author actions disseminate FNPD content and which content elements influence social behavior in OSNs, we propose to use the production rules approach (linking antecedents (IF) with consequents (THEN)) to provide a structured framework for representing knowledge and reasoning in FNPD research field. One of the advantages of this approach is that FNPD can then be modeled as a generative FNPD process that can respond flexibly to the rapidly evolving nature of FNPD by expanding the set of production rules rather than creating fragmented solutions.

(vi) In the field of social impact research, there are important areas of analysis (niches) that have not yet received sufficient research attention, such as social impact research via social behavioral patterns analysis, radicalization and polarization research, reasons for successful deception, social impact modeling, echo chamber polarization effects, cognitive warfare.

(vii) We found that the analysis of FNPD authors and disseminators is closely related to social impact modeling, particularly in the use of similar user characteristics, where measuring author/user credibility can be the key metric in FNPD author and disseminator analysis and social impact modeling. Both fields make extensive use of graph-based models, which are recognized as one of the most advanced methods for modeling network propagation. A common trend in these fields is a preference for combined approaches.

A brief summary of the content of the following systematic review is given below. The second section describes selection process of relevant articles using the systemic methodology in three research domains (domain#1: authors/disseminators, domain#2: content research, and domain#3: social impact research). The third section presents the main results of a meta-analysis of these research domains using a set of qualitative and quantitative criteria. The fourth section outlines FNPD research domain links and the foundation model concept. The fifth section presents the conclusions and discussion in terms of the main findings, relationships between the three research domains, limitations, and further research perspectives.

II. SEARCHING AND SELECTING METHODOLOGY

Methodology plays a crucial role in conducting a systematic literature review. For this study, we have chosen to follow the approach outlined in the PRISMA Statement [45], [46], which is a widely accepted checklist used by researchers worldwide to guide and inform the development of systemic literature reviews [46]. Thus, all three domain-specific searching and selection processes employed the PRISMA systemic review framework.

We established four criteria for inclusion: the article must (1) be published in a peer-reviewed academic research journal, (2) be written in English, (3) be published between January 2018 and April 2024, and (4) have its full text available.

Regarding the field of this research and best practices in the field, four databases were selected for this research including Semantic Scholar, Google Scholar, Crossref, and Scopus databases. All searches and records we performed

with Publish or Perish software program that retrieves and analyzes academic citations from external data sources.

The selection processes of relevant articles using the systemic PRISMA methodology for domain#1 (authors /disseminators), domain#2 (content research), and domain#3 (social impact research) are given in Appendixes B, C, and D, respectively. The flow charts there highlight six steps, based on standard selection criteria, to find articles of interest. In the last step of the selection process, we selected the most cited recent papers from each field in equal proportions to ensure equal coverage of all research domains. Therefore, for each research area, 30 articles were selected for detailed meta-analysis.

Our systematic review spans January 2018 to April 2024. The research articles we found, using the well-known PRISMA systematic article selection and analysis methodology, provide objective representation of the research and experimental results, innovations, datasets and new frameworks that characterize this period. Thus, our systematic analysis was limited to these articles, which may have missed a number of important others that may have significant value to the field. We admit that such selection biases are possible due to systematic error of selection methodology. The selection process was not manual (until step 5, see Appendixes B, C and D) and not based on our preferences or likes and dislikes. However, in order to increase the inclusion of more academically recognized new articles in the meta-analysis phase, we sorted them (in the fifth step, see Appendixes B, C and D) and selected the most cited ones according to their average annual citation rate (total number of citations/number of years published).

Fig. 2 illustrates the number of publications per year, revealing a stable increase in publications in all research domains. The fact that the number of publications on FNPD authorship, content and social impact studies has increased by a factor of around five between 2018 and 2023 is a clear indication not only of the growing relevance of this area of research, but also of the increasing need for systemic analyses that reveal and contextualize the multitude of methods and models used.

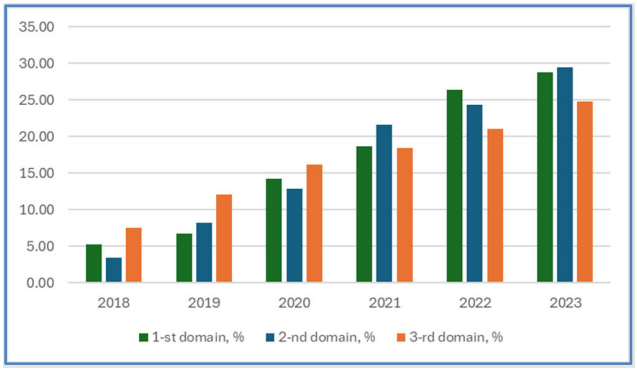


FIGURE 2. Publication statistics (number of articles found) between 2018 and 2023 in all three FNPD research domains.

All three research areas have their own scope of research, which consists of a domain-specific research context, modelling approaches, measurement metrics and outcomes. Therefore, before going into a detailed analysis of the individual domains, it is useful for the reader to be aware of the overall conceptual map of the research field presented in Fig. 3. In the following meta-analysis section, we will look at this research field scope in more detail.

In essence, this systematic review consists of three main research processes: the selection of articles, the meta-analysis of their content and the results interpretation (synthesis). All of these key processes are associated with high levels of bias. To reduce the risk of bias in the study, the aim was, following the PRISMA methodology, to ensure a continuous monitoring of those processes.

Admittedly, the first bias refers to any potential bias that may arise during the selection process of studies. The second bias refers to any bias that may arise in the meta-analysis process for each study. The third bias concerns the interpretation and synthesis of the results of the meta-analysis, which is particularly important as it influences the overall quality of this paper. This bias relates to cases where systematic reviewers may be biased in their interpretation of the results of the studies according to their own perspectives.

In this section, we present the methodology used to search for the most relevant articles in the three research domains that used DL and ML approaches for (FNPD) analysis. To the best of our knowledge, the closest match to the proposed systemic review structure can be found in the paper written by Varlamis et al. [41], where the authors review the application of GCN (Graphical Convolutional Networks) in the task of detecting fake news in these three directions: (i) the fake news content, (ii) the sources that generate fake news, and (iii) the networks that amplify the spread of fake news. However, their research is limited to GCN approaches, which our study is not.

Below we explain the need to systematically cover the three main thematic domains related to FNPD research in a single study:

(i) computer science researchers specialized in ML/DL lack an understanding of the broader social scope of this research subject, including FNPD authors' analysis, social impact modelling, etc.

(ii) it offers wider opportunities for synergies and fusion of FNPD research across several related fields, in order to obtain more efficient and faster FNPD detection methods and models.

(iii) it provides researchers with the convenience of finding three thematically related systematic reviews in one place (combined research approach),

(iv) it provides greater opportunities to explore the linkages between all three research areas, not only from the perspective of combining ML/DL approaches, but also from the perspective of the overall needs of the social problem-solving context and of the ultimate beneficiaries of such research.

There are a number of other systematic reviews, mentioned in the introduction, that look at specialized technical solutions for individual FNPD research domains, but there are hardly any that look at the linkages between research domains, and at the solutions for integrating them. Therefore, the systemic review presented here analyses three related areas of FNPD research, specifically looking for links between the research of FNPD authors (sources), the FNPD content they produce on social networks, and the impact of this content on societies (e.g., formation of echo chambers, fragmentation, radicalization, etc.). This opens up new possibilities for linking the causes of FNPDs with their consequences. It should be emphasised that, as our review has shown, combining these areas produces more robust and reliable results.

In this regard, this systematic review highlights and seeks to address one of the main shortcomings of current FNPD research, i.e., the lack of integration between above mentioned domains of research. We therefore aim to explore the wider possibilities for bridging the gaps in their integration in this systematic review. The need for such integrated research is clearly expressed in various government programs, political agendas, regional strategic and research initiatives, FNPD prevention programs, national security programs, regulations on transparent and credible journalism (see Appendix A).

It is important to underline that (a) each research domain individually reaches its own ceiling of potential, which is lower than that of the combined approaches, and (b) without integration with the other domains, each domain individually is not able to provide integrated real-time practical solutions to the containment of the proliferation of FNPD flows. In fact, as FNPD flows become faster, more complex, AI-generated, covert, multimodal, multi-networked, systemic and organized, all the domains of research mentioned in our work need to be effectively combined.

Thus, this triple systemic review reveals relationships between research domains #1, #2 and #3 in terms of datasets, methodological approaches, results obtained, etc. However, a combined systemic review also has its drawbacks, such as overlapping some searching results, a need for separate meta-analyses in each domain, and a higher volume of work.

From a broader perspective, in order to operationalize how and which author actions generate and disseminate FNPD content, and in turn which content elements influence social behavior in OSNs, we need a versatile and flexible approach to monitor, model and influence different FNPD creation, dissemination and influence scenarios. Therefore, we propose to use the well-known AI production rules approach. Production rules provide a structured framework for representing knowledge and reasoning in AI. Each rule consists of an antecedent (IF): a condition or set of conditions that must be met for the rule to be activated, and a consequent (THEN): an action or conclusion to be taken when the conditions are met. Uncovering and algorithmization of these production rule sets is a major challenge with many plausible effects in the field of FNPD research, see Fig. 4.

One of the advantages of the proposed approach is that the FNPD research process can then obtain an integral and generative form, which, depending on the production rules sets' IF/THEN conditions, could generate a wide range of possible FNPD tackling scenarios with different social impact outcomes. We suggest that finding two sets of production rules, namely #1 $\rightarrow$ #2 and #2 $\rightarrow$ #3, would allow the construction of the generative approach for automatic FNPD detection and monitoring to take countermeasures, see Fig. 4.

Another advantage of this proposed approach, according to the authors, would be the automated ability to combine, through production rules, a wide range of ML/DL and other methods and models used in FNPD research, and to use them selectively or in combination, depending on the available datasets and the end goal, to achieve the best results.

In this way, it would provide modularity and flexibility to respond to the rapidly evolving nature of FNPD by expanding the set of production rules with constantly updated new datasets, ML/DL methods and models, rather than creating fragmented solutions.

In essence, such production rules approach provides a framework for foundational FNPD model construction, i.e., an AI model trained on extensive FNPD datasets that can be applied to different FNPD domains. In other words, such foundational model refers to a large-scale machine learning model that is pre-trained on a broad and diverse FNPD dataset, allowing it to be adapted and fine-tuned for a variety of downstream tasks and applications.

Such a foundation model takes advantage of transfer learning, where knowledge gained from one FNPD task can be applied to another, making it highly efficient and versatile as it can be adapted to new FNPD tasks with relatively little additional data. The scale of such a foundational model would allow it to capture complex patterns and relationships in the data. It can also use a transformer architecture, which is particularly effective for tasks involving sequences of data in FNPD NLP tasks.

We elaborate on our proposal for using such an approach in the development of a foundational FNPD model in the Conclusions and Discussion section of this paper.

III. SYSTEMIC META-ANALYSIS FOR DOMAINS #1, #2 AND #3

The systematic meta-analysis of the literature was carried out for all three FNPD research areas (#1, #2, and #3), with a separate subsection for each area in turn. Each FNPD area was analyzed within the framework of the selected criteria. The main results are presented in this section, and the following summary insights are presented in the Conclusions and Discussion section.

A. AUTHORS AND DISSEMINATORS (DOMAIN #1)

Authors and disseminators play an essential role in disseminating FNPD on social media. The process starts with the author publishing FNPD, which then becomes visible to disseminators. These disseminators engage with the content

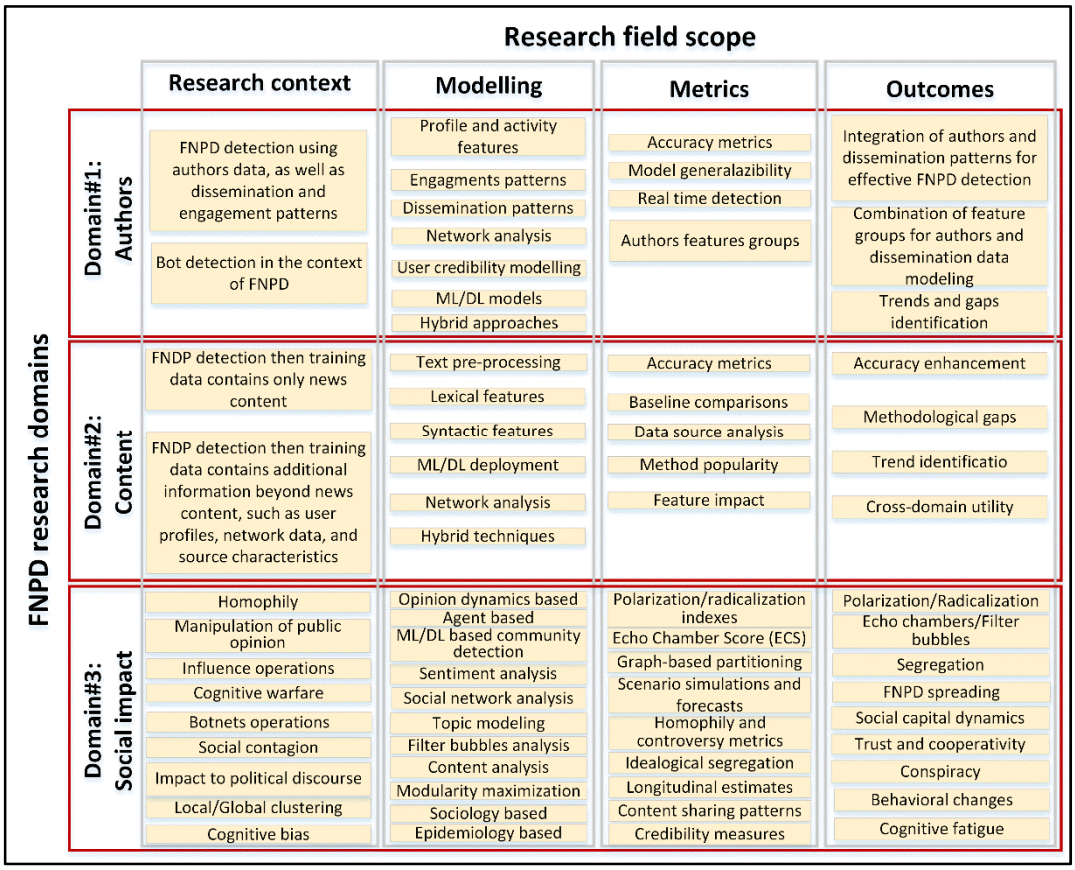


FIGURE 3. A conceptual map of the research field scope (context, modelling approaches, metrics used and outcomes) in all three research domains.

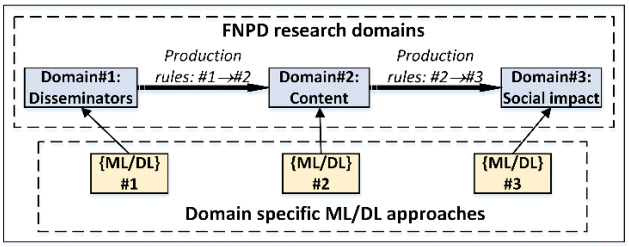


FIGURE 4. Relationships between research domains #1, #2 and #3 in terms of domain-specific ML/DL approaches and production rules.

through various interactions, including likes, shares, comments, emotional reactions, clicks, and views. The more engagements the content receives, the more visible it becomes on social media platforms, leading to broader dissemination. This review primarily focuses on FNPD detection using data from authors and disseminators, as well as engagement and dissemination patterns. Additionally, the review analyzes bot detection methods on social media, considering their significant role in the dissemination process of FNPD.

For the final full-text meta-analysis, 30 articles published between 2018 and 2024 were selected: 9 focused on bot detection, while 21 addressed FNPD detection. Fig. 5 presents the main methods used in the first research domain. The methods employed various ML and DL models, revealing three primary approaches: traditional ML-based, DL-based, and hybrid models.

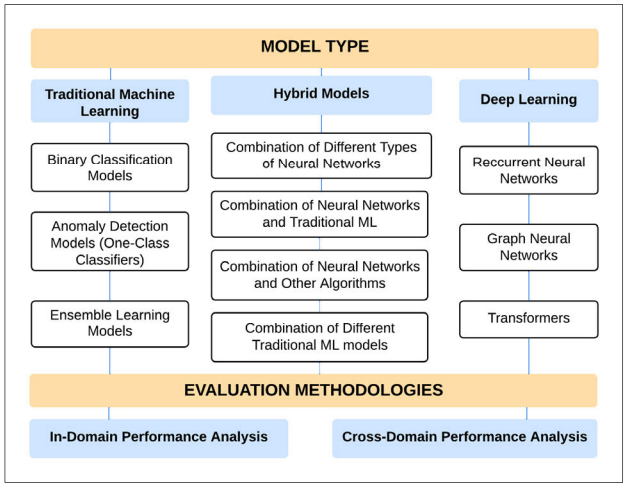


FIGURE 5. Hierarchical categorization of methods for the domain#1 (authors/disseminators).

A total of 9 studies used only traditional ML models, which are more commonly applied to bot detection tasks. One-class

classifiers were used exclusively for bot detection, where the problem was framed as anomaly detection. An ensemble model was also used exclusively for bot detection. Ensemble methods were classified as traditional ML rather than hybrid approaches because they involve independent classifiers that vote on a result, whereas hybrid methods combine multiple models to produce a single prediction.

Since traditional ML was mainly used in bot detection, hybrid models were primarily employed for FNPD detection. However, not all approaches focused solely on FNPD classification; some models aimed to identify specific components involved in the FNPD dissemination process, such as detecting conspiracy propagators, specific accounts, and sources. In general, hybrid models were the most commonly used in the first research domain, with almost half (14 out of 30) of the studies utilizing hybrid approaches. The most popular combinations involved different neural network combinations and traditional ML models integrated with neural networks. Some methods integrated other algorithms with neural networks, such as generative models, transfer learning, and rule-based models. DL-based models using specific types of neural networks were found in 7 studies; these methods were almost equally common in both bot detection (3 studies) and FNPD detection (4 studies).

As the content and dissemination strategies of FNPD in social media rapidly change, it is essential to investigate the generalizability of models. Consequently, evaluation methodologies were also examined. Two different performance analysis types were identified: in-domain and cross-domain. In-domain performance analysis evaluates how well a model performs on data from the same distribution as the training data. In contrast, cross-domain performance analysis assesses how well a model generalizes and performs on data from a different distribution or domain than the training data. Generally, in-domain performance analysis was the more common method for model evaluation, with 23 out of 30 studies using this approach. Cross-domain performance analysis, however, was primarily used in bot detection tasks; 5 out of 9 studies on bot detection employed this method, and only one FNPD identification study used such an approach. Moreover, cross-domain performance analysis was mainly applied in traditional ML models (4 studies). However, some DL-based approaches (2 studies) also used this evaluation methodology, while none of the hybrid methods employed it.

1) METHODS AND APPROACHES

As mentioned earlier and illustrated in Fig. 5, the main methods used in the first research domain are categorized into three primary model types: traditional ML, DL-based models, and hybrid approaches. This section provides an in-depth analysis of the methods employed in the first research domain. The first three subsections (a, b, c) offer a detailed examination of the modeling techniques used in traditional ML, DL-based models, and hybrid approaches, while the fourth subsection (d) discusses real-time detection systems, focusing on methods designed for early FNPD detection using limited early-stage data.

a: TRADITIONAL MACHINE LEARNING

Various ML models have been used for authors' and disseminators' data modeling, including binary classifiers, one-class classifiers, and ensemble methods, see Fig. 5. Traditional machine learning was predominantly used for bot detection tasks, with 6 out of 9 approaches exclusively employing this method. In contrast, for the FNPD detection, only 3 out of 21 approaches relied solely on traditional ML.

The random forest model outperformed other traditional ML models across various modeling strategies, including detecting social bots using minimal account metadata [48], training multiple classifiers with diverse features like profile data, user activity, and engagement information [49], detecting bots in low-resource languages [50], and leveraging features effective for identifying bot accounts to detect fake human-created accounts on social media [51]. It also delivered top performance in FNPD detection, utilizing feature vectors derived from news dissemination patterns at different network levels, such as network analysis [52] and both micro and macro dissemination patterns [53].

However, for certain specific tasks, other traditional ML models demonstrated the best performance. For example, logistic regression was most effective at identifying potential misinformation target topics and detecting fake news [54], SVM excelled at bot detection using profile data and tweet text analysis [55], and Bagging-TPMiner, a one-class classifier, successfully identified bots as anomalies even when their behavior had not been seen before [56].

It is worth noting that cross-domain performance analysis for traditional ML methods has only been applied to bot detection task, where a collection of labeled datasets, rather than a single dataset, was used to evaluate model performance [48], [49].

b: DEEP LEARNING

Deep learning-based approaches were less common in the first research domain and primarily relied on textual information. However, some methods incorporated additional data. For example, an LSTM model combined textual information with user metadata for bot detection [57] and FNPD detection [61]. Other methods relied solely on textual data, such as the BiLSTM model used for bot detection based exclusively on tweet content [58], a pre-trained BERT model for sentiment classification derived from tweets [59], and the DEFEND system, which utilized textual data from news articles and user comments, applying a GRU model for FNPD detection [60].

Most DL-based models used textual content, whereas graph-based models did not rely on textual information. Instead, GNNs have been used for fake news detection by incorporating news propagation patterns with user profile features [62] or by using propagation patterns combined with

user profile and activity information through geometric deep learning [63].

Among the evaluation methodologies, only two DL-based approaches employed cross-domain performance analysis: the bot detection model that relied solely on textual information [58] and the propagation-based fake news detection approach [62].

c: HYBRID MODELS

Hybrid approaches were commonly used in the first research area, focusing on FNPD detection with various modeling methods and tasks. Some approaches targeted specific components of the FNPD dissemination process, such as identifying conspiracy theory disseminators using CNN layers for content-based embeddings combined with dense layers for psycholinguistic features [64]. Another method combined FastText and MUSE models to detect malicious social media accounts spreading political disinformation in a low-resource language (Tagalog), leveraging aligned multilingual embeddings for transfer learning [65]. Additionally, a boundary-based community detection approach integrated SVM, k-means clustering, and the Leader Rank algorithm to identify propagandistic communities and core propagandistic nodes on social networks [66].

Some approaches have focused on detecting fake news sources. For instance, a trust management system was developed to evaluate news sources and detect fake news, using XGBoost to classify articles based on textual content and source URLs, with claims verified against the FEVER dataset and aggregated to assess overall reliability [67]. Another approach combined traditional ML models with node2vec embeddings from the social graph, employing a light gradient boosting machine to select key features and achieving optimal classification results with an LDA model for identifying users as fake or reliable news sources [68].

Several studies focused on FNPD detection by leveraging news dissemination patterns, utilizing hybrid approaches with GNN components. For example, the FAKEDETECTOR model [69] combined information from textual news articles, news creator profiles, and subject descriptions with latent features derived from GRU layers to create a graph-based model. The UFPD framework [70] integrated BERT for encoding textual content, GraphSAGE for aggregating information from neighboring nodes in the news propagation graph, and GCN for encoding the graph structure, incorporating both content and social context. The FANG model [71] combined GraphSAGE with Bi-LSTM to capture the structural relationships between news articles, sources, and users, while modeling the temporal dynamics of user engagements.

Models integrating different techniques for capturing relationships within news dissemination patterns extended beyond graph neural networks. The FakeNewsTracker framework [72] combined an LSTM model with an autoencoder, where the autoencoder captured linguistic features from news articles and the LSTM modeled temporal patterns of social media engagements. The TriFN framework [73] modeled publisher-news, user-news, and user-user relations using matrices and user credibility scores, integrating these into a semi-supervised linear classifier for fake news prediction. The FR-Detect framework [74] evaluated news content and publisher credibility, using CNN to extract linguistic features and the CreditRank algorithm to assess publisher-related attributes like Credibility and Influence, enhancing early detection. CNN was also used for text modeling in [75], where features extracted by CNN were processed by Bi-LSTM to capture context and dependencies. Another approach [76] combined CNN with a User Response Generator to analyze textual content and generate user responses. Lastly, RNNs and CNNs were combined in [77] to model news propagation paths, with GRU units processing user dissemination sequences and 1-D convolution generating feature vectors.

d: REAL-TIME DETECTION SYSTEMS

Several studies have explored the challenge of early FNPD detection. Some research emphasized rapid identification, with methods capable of detecting fake news within minutes of dissemination by analyzing user profile characteristics [77], while others demonstrated high accuracy within a few hours. For example, one study achieved 90% accuracy just two hours post-dissemination, with performance peaking after seven hours [63]. The impact of delay time on model performance was also examined, showing that extending detection windows up to 48 hours significantly improved F1 scores [73]. To address early detection without user comments, a user response generator was proposed [73]. Models trained on initial social engagements, such as retweets, demonstrated accuracy levels of 0.7 to 0.8 early in the news cycle [62], with engagement-only models reaching 80% accuracy, which increased to 0.933 when additional data were incorporated [52]. Systems were developed to monitor and analyze news in real time, integrating text analysis, fact-checking, and contextual evaluation for swift feedback [67], while other frameworks utilized content and publisher features for prompt fake news detection [74]. Additionally, an early detection model was devised to identify emerging fake news topics [54].

2) FEATURES AND PERFORMANCE ANALYSIS

To understand how specific features of authors and disseminators can be leveraged to develop more accurate FNPD detection models, we conducted a detailed analysis of the features used for modeling author and disseminator data, as well as a performance evaluation across different datasets. This analysis is organized into four subsections: a) examines how user credibility is modeled; b) explores key feature groups used for modeling authors and disseminators; c) discusses the combinations of multiple feature groups; d) evaluates the performance of models across different datasets. The section

concludes with the main highlights, summarizing the key insights from the meta-analysis of the first research domain.

a: MODELING THE CREDIBILITY OF AUTHORS AND DISSEMINATORS

Within FNPD identification on social media, user credibility metrics are crucial for measuring the reliability of authors and disseminators. The basic hypothesis is that trustworthy users with a consistent history of disseminating credible information positively contribute to the integrity of the information they share, like, or comment on. Conversely, users with a history of engaging with FNPD tend to amplify the dissemination of non-credible content. Some studies [63], [73] found that users tend to group themselves based on their trustworthiness when sharing news, creating clusters that share reliable or unreliable information. Additionally, research [52], [54] indicated that the extent users engage with specific topics through commenting could signal their likelihood of disseminating false information. Regarding textual information, studies [69], [71] discovered that analyzing the words and phrases in users' profiles could help identify their tendency to share specific narratives.

Furthermore, research [53], [68] explored the connections between user interactions and the credibility of the news they share, suggesting that observing users' behavior can provide clues about the content's reliability. User credibility modeling has been significantly enhanced by evaluating features such as activity history, follower credibility, and engagement levels [74]. This approach has improved the detection of fake news through a detailed analysis of authors' behavior and influence. Similarly, the ConspiDetector model [64] integrated psycholinguistic and profile characteristics to differentiate between users likely to propagate conspiracy theories and those who oppose them.

b: FEATURES GROUPS

Various features have been used to model authors' and disseminators' data, drawing on a wide range of characteristics from user profiles, network analysis, engagement, and dissemination patterns. The most common features were derived from user engagement data, although characteristics from user profiles and activity metrics were also frequently utilized. In general, six different feature groups were identified in the first research domain:

i) USER PROFILE AND ACCOUNT CHARACTERISTICS

Static attributes of a user's social media profile. (18 studies used these features)

ii) USER ACTIVITY METRICS

Measures authors' and disseminators' actions on social media platforms, providing metrics that capture specific user activity frequencies, such as the total number of statuses, favorites, replies, retweets, shares, etc. (19 studies)

iii) ENGAGEMENT INFORMATION

Analyzes disseminators' interactions with content on social media platforms, encompassing actions such as likes, emotional reactions, comments, shares, clicks, etc. (22 studies)

iv) NETWORK ANALYSIS

Measures overall network metrics and structures, such as network density, depth, centrality, and connectivity, including metrics like cascade patterns, relationship densities, and triad formations (4 studies).

v) DISSEMINATION PATTERNS

Focuses on the flow and dissemination of content, analyzing how information disseminates through disseminators' engagements. This includes tracking propagation paths, identifying user engagement sequences, and examining the temporal dynamics of content dissemination using metrics like adjacency matrices, propagation paths, etc. (7 studies)

vi) CREDIBILITY METRICS

Measures users' trustworthiness by evaluating their engagement with credible versus non-credible content. (10 studies)

c: COMBINATIONS OF DIFFERENT FEATURES

Different features combinations were used to model authors' and disseminators' data. Fig. 6 illustrates all combinations of features found in the first research domain used to model authors and disseminate data.

In general, the most used feature group was engagement information, which includes all the information related to interactions with published content. The engagements were used in 22 out of 30 studies. Some studies relied only on engagement information [59], [60], [75], [76], while others combined it with other feature groups. Overall, most engagement combinations included profile characteristics by adding other features such as dissemination patterns [71], network analysis [66], activity metrics [49], [50], [56], [58]; activity and credibility metrics [64], [74], credibility metrics [69]; activity and dissemination patterns [77]; activity, dissemination patterns, and credibility metrics [68], activity, network analysis, dissemination patterns, and credibility metrics [62]. Others did not include in the engagement combinations profile characteristics and added other metrics such as activity information [70], credibility metrics [62], [73], credibility metrics, network and dissemination patterns [52], [53], and activity and credibility metrics [60]. Some approaches did not use engagement information. All these models included activity information, and most combined it with user profile characteristics [48], [51], [55], [61], [65], while others relied only on activity information [58], [67] or combined it with profile information and dissemination patterns [62].

d: PERFORMANCE ANALYSIS

In the first research domain, it was challenging to analyze and compare the performance of models due to the diverse

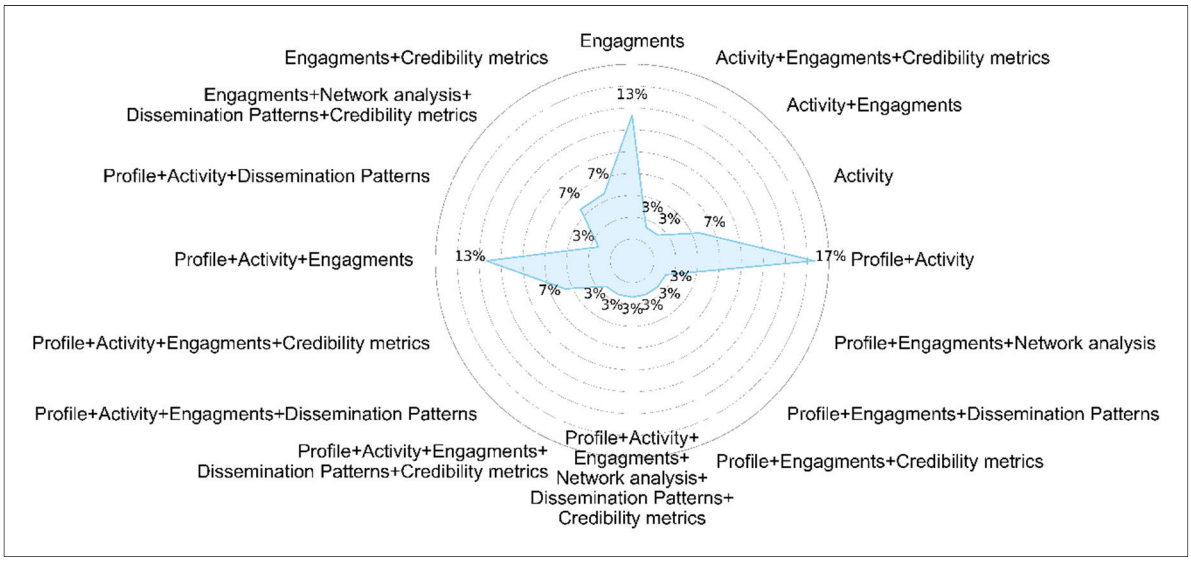


FIGURE 6. Hierarchical categorization of methods for the domain#1 (authors/disseminators).

datasets and tasks involved. To address this issue, the focus was placed solely on approaches with binary outputs for FNPD or bot detection, excluding studies targeting specific, excluding studies targeting specific components like propaganda communities [66] or critical propagators [64]. Research-based on unique datasets collected by authors [51], [54], [61], [63], [65], [67], [69], [71], [75] or combinations of different datasets [48], [49], were also excluded, resulting in the evaluation of 17 studies: 11 for FNPD [52, 53, 60,62,68, 70, 72-74, 77, 76] and 6 for bot detection [50], [55], [56], [57], [58], [59]. For the FNPD detection performance analysis, the data sources used include the FakeNewsNet data repository with BuzzFeed, GossipCop, and Politifact datasets [52], [53], [60], [62], [63], [70], [72], [73], [74], Weibo [77,76], Twitter15 [77], Twitter16 [77], and CredBank [68] datasets. Cresci [55], [56], [57], [58], [59] and Twibot-20 [50] datasets were used for bot detection.

Table 1 summarizes the performance of the best approaches for each dataset, illustrating that hybrid models consistently achieve superior results across various datasets. Engagement features were crucial for FNPD and bot detection tasks and were frequently included in the best-performing models.

The best-performing models for the FakeNewsNet data repository, which includes news articles from various fact-checking websites and additional social context information, varied in their approaches.

For the BuzzFeed dataset, the top approach utilized a hybrid model combining traditional ML with relationship modeling [73]. The GossipCop dataset achieved the best results with a hybrid model that combined GNN and BERT [70]. The Politifact dataset's highest performance came from a traditional ML approach that emphasized engagements, network analysis, dissemination patterns, and credibility metrics [52].

The Weibo dataset, comprising user-generated content from Sina Weibo social network, achieved the best performance with a hybrid model combining GRU and CNN [77]. The same research also applied this hybrid model to the Twitter15 and Twitter16 datasets, which consist of labeled tweets with user information and interactions. However, the highest results were achieved with the Weibo dataset. The CredBank dataset, a large-scale collection of annotated tweets surrounding real-world events, achieved the best performance with a hybrid model that combined GNN and traditional ML [68].

For bot detection, the cresci dataset, annotated with human and bot accounts, demonstrated the effectiveness of a model combining bert with mlp [59]. meanwhile, the twibot-20 dataset, a large-scale twitter dataset for bot detection, achieved the best performance using a tradi.

Main Highlights:

- Hybrid models that combine various ML and DL methods are the most effective for FNPD detection, as they integrate diverse data types, including textual content, dissemination patterns, user profiles, engagement data, and social context. By leveraging multiple data sources, these models provide a comprehensive understanding of how authors and disseminators contribute to the FNPD dissemination process.
- Modeling authors and disseminators in the context of FNPD relies on diverse features; however, engagement features are the most frequently used, highlighting the crucial role of social interactions in the FNPD dissemination.
- Relying on static features, such as user metadata, often fails to capture the dynamic nature of social media interactions. This limits the effectiveness of traditional ML approaches, especially for evolving bot detection strategies.

TABLE 1. Performance of top models across different datasets in the first research domain.

Dataset	Article	Model	Authors and disseminators features	Task	Accuracy	F1
FakeNewsNet BuzzFeed	[73]	Hybrid (Traditional ML + relationship modelling)	Engagements, Credibility metrics	FNPD detection	0.865	0.884
FakeNewsNet GossipCop	[70]	Hybrid (GNN +BERT)	Activity, Engagements	FNPD detection	0.972	0.972
FakeNewsNet PolitiFact	[52]	Traditional ML	Engagements, Network analysis, Dissemination patterns, Credibility metrics	FNPD detection	0.929	0.939
Weibo	[77]	Hybrid (GRU+CNN)	Profile, Activity, Engagements, Dissemination patterns	FNPD detection	0.921	0.92
Twitter 15	[77]	Hybrid (GRU+CNN)	Profile, Activity, Engagements, Dissemination patterns	FNPD detection	0.842	0.843
Twitter 16	[77]	Hybrid (GRU+CNN)	Profile, Activity, Engagements, Dissemination patterns	FNPD detection	0.863	0.859
CREDBANK	[68]	Hybrid (GNN +Traditional ML)	Profile, Activity, Engagements, Dissemination patterns, Credibility metrics	FNPD detection	0.99	0.98
Cresci	[59]	Hybrid (BERT + MLP)	Engagements	Bot detection	0.979	0.962
Twibot-20	[50]	Traditional ML	Profile, Activity, Engagements	Bot detection	0.914	0.91

- Measuring user credibility based on engagement history is essential, as trustworthy users promote credible content, while others amplify FNPD. Analyzing these metrics can help more accurately identify the dissemination of FNPD.
- Real-time detection systems primarily rely on early social engagement features for timely FNPD identification. However, the limited availability of dissemination data in the early stages restricts the effectiveness of the development of these systems.

3) LIMITATIONS, NOVELTIES AND TRENDS

a: LIMITATIONS

Several limitations have been identified in the first research domain. Reliance solely on user metadata has been found to be insufficient for detecting bots, emphasizing the need for content analysis and the inclusion of network-based features, which are crucial for identifying coordinated botnets [50], [60]. Additionally, profile characteristics alone are ineffective in predicting whether a user is a disseminator of non-credible content, highlighting the need for more dynamic and context-specific features to improve detection accuracy [64]. Data limitations present inherent challenges in these tasks.

The laborious and expensive manual annotation process restricts the volume of annotated data, making it difficult to detect new types of bots and unreliable news over time, necessitating periodic updates with new annotations [50]. The rapid obsolescence of data, particularly hyperlinks and social media traces, along with model sensitivity to training and testing datasets, poses additional challenges [49], [71]. FNPD studies also highlight issues such as model complexity, resource requirements, and difficulties in model interpretation and generalization [63], [67].

Reliance on a single platform, predominantly Twitter, introduces bias, as it may not accurately reflect broader social media behavior [64]. The dynamic nature of social media further complicates generalization and real-time detection across platforms, making it challenging to identify unreliable news from official sources [54], [67], [74], [75]. Dataset constraints imposed by the limitations of the Twitter API standard and recent changes in Twitter's API usage policies restricting access to the Twitter Academic API further limit data collection availability [67], [80]. Additionally, language-specific challenges such as jargon, minimal word usage, and non-standard language constructs [59], [66] and the limited amount of data in low-resource languages [65] further complicate the detection process.

b: NOVELTIES

The most frequently mentioned novelty in the first research domain is advanced neural network architectures. Studies [50], [57], [60], [64], [67], [74], [75], [77], highlighted creating hybrid models for combining news content and user interactions as a novel aspect. Some approaches were characterized using advanced deep learning techniques focusing on in-depth semantic analysis and detection based on news propagation methods, notably in [53], [59], [62], [63], [76]. Graph neural networks were the main novelty in studies focusing on network-based detection and social context modelling, explicitly mentioned in [52], [71], and [78]. Some studies emphasized novelties as a focus on the specific modelling direction. For example, studies [54], [56], [66], [68], [70], [74] considered analyzing user behavior and social dynamics as their main novelty, incorporating features such as user behavior, polarization or specific communities' detection, the integration of psycho-linguistic features for conspiracy

propagators identification as a novel aspect is mentioned in [64].

Real-time detection strategies were also highlighted as novel in studies [48], [55], [58], [61]. The adaptability of detection models to different scenarios and platforms was highlighted as a novel aspect in [51] and [72]. The adaptation for specific language, addressing a gap in existing methods that are primarily based on English-language models, is a novel aspect highlighted in [50]. The creation of novel dataset, specifically designed for propaganda detection task as a novelty is highlighted in [66]. Explainability and transparency of AI in detection processes was a novel focus in [49], [67], [69], and [73]. These studies emphasized user-friendly AI techniques, indicating a trend towards more accessible and explainable AI solutions.

c: TEMPORAL PATTERN

Studies were searched from 2018 to April 2024. In general, the evolution of research shows a trend towards more sophisticated and diverse methods, reflecting the increasing complexity of the challenge posed by FNPD dissemination on social media. Here is a summary over time:

2018-2019: Emphasis on traditional ML models and the early development of hybrid models, focusing on authors' and disseminators' profile data and textual content analysis.

2020-2021: Diversification and advancement in DL techniques, hybrid models, and model generalizability through cross-domain performance analysis. Focus on integrating network metrics and dissemination pattern analysis to improve detection methodologies.

2022- April 2024: Shift towards practical applications, real-time detection systems, and advanced DL approaches integrating more sophisticated features such as psycholinguistic metrics and multilingual methods. Enhanced focus on user credibility evaluation and engagement patterns.

B. CONTENT ANALYSIS (DOMAIN #2)

For the final full-text meta-analysis, 30 articles published between January 2018 and April 2024 were selected. Existing approaches to FNPD content analysis are analyzed in terms of 1) textual data processing and 2) application of machine learning (ML) and deep learning (DL) algorithms. An overview of these approaches is provided in two scenarios based on the composition of the training data. The first scenario involves training models using only news content and is the focus of 20 of the 30 papers reviewed. This approach is referred to in the literature as content based. The second scenario, explored in the remaining 10 papers, incorporates social context information into the training data beyond news content. These papers examine user profile features [61], [63], [79], [80], [81], post features such as likes, retweets, and replies to tweets [61], [63], [69], [79], social network features and aspects of dissemination, including propagation patterns and the dynamics of news sharing [53], [63], [79], [80], [82], [83], [84], [85], demonstrating a diverse approach to this feature extraction. Based on an analysis of the literature, it is evident that social context features have a common goal: to improve the accuracy of FNPD detection. The effectiveness is confirmed in each paper mentioned by comparing them with several existing baselines.

1) TEXTUAL DATA PROCESSING

A machine learning system begins by converting information from textual data into numerical representations. This subsection examines how numerical representations are constructed for FNPD detection. For text representation construction, we review techniques ranging from hand-crafted, rule-based to sophisticated deep learning models that understand language in a human-like manner by learning distributed representations. In addition, we explore the incorporation of sentiment analysis to answer the question of whether understanding emotional tone is important for FNPD detection. A summary of the techniques used to handle text data is given in Table 2.

TABLE 2. Summary of textual data processing techniques across studies.

Technique	Training data consists of news content only	Training data includes social context information for news content
Hand-crafted features	Syntax-based features [99,103,98], lexical features [89,86,103,98, 87,88]	Syntax-based features [79,61,80], lexical features [79, 80], user profile and post features [79,63,61,80]
Traditional vectorization techniques	TF-IDF [88,100,98,81], BOW [88, 98]	
Static word embeddings	Word2Vec [92,93,86,94, 102,81], GloVe [92,95,96,81], FastText [97,81]	Word2Vec [81,85], GloVe [81,85], FastText [84], MITTENS [84]
Contextual word embeddings	BERT [104,93,81], Xlnet [93], RoBERTa [93,81], DistilBERT [101], BART [81]	
Graph-based features	Sequential, syntactic, and semantic features [90,91]	Temporal feature representations [82], propagation tree features [83,63,53,85], authors-created hybrid model for representations learning [69]
Sentiment analysis	Polarity [86], sensitivity [86], subjectivity [99].	Subjectivity [79], polarity [83]

a: IN-DEPTH ANALYSIS OF TEXTUAL FEATURES

As shown in Table 2, each paper makes a unique contribution to the field of FNPD detection, using different aspects of techniques and features. Hand-crafted features include syntax-based and semantic-based analysis. In the context of lexical cases, researchers use N-gram-based features [86], [87], [88]. Linguistic Inquiry and Word Count (LIWC) emerges as a tool that demonstrates its utility in extracting content from news sources [79], [80], [86], [89]. Another alternative tool is RST [80]. The features described

in [90] and [91] are graph-based. They are extracted from text to create graphs of word dependencies, which are processed by graph neural networks for propaganda detection. The paper [90] focuses on integrating external news sources to provide additional context, while the paper [91] emphasizes the hierarchical integration of features within and between different types of graphs.

Social context analysis attempts to analyze the behavior of the users involved in sharing the news and extract various features from the network to determine the veracity of the news. This analysis is approached using simple metrics such as the number of likes, shares, and comments [61], [79] and more complex graph-based methods [53], [63], [69], [82], [83]. Graphs are constructed to represent user interactions, with nodes representing users and edges representing interactions [69]. Graphs are also used for propagation tree construction [53], [63], [81], [82], [83], focusing on analyzing how news spreads across the network. The main difference between [82] and the remaining work using graph-based features is its strong focus on temporal dynamics and how interactions evolve over time. The authors of [69] deploy a new Hybrid Feature Learning Unit (HFLU) for learning the explicit and latent feature representations of news articles, creators, and subjects, respectively. In the process, specific words are extracted from the text based on their frequency and relationship to whether the news is fake or real. Work [63] emphasizes propagation patterns and social network structures. Castillo emerges as a tool for features from a social context [82].

Static word embeddings are the most cited embedding technique, and they are used in both scenarios, depending on whether the training data consists only of news content or includes social context information. Word2Vec [84], [86], [92], [93], [94] and GloVe [84], [92], [95], [96] are the most preferred static word embeddings. In addition, studies have used FastText [84], [97] and MITTENS [84]. MITTENS is an extension of GloVe for learning domain-specific representations.

The overlap of papers using multiple techniques indicates a comprehensive comparison of different methods across different categories [81], [87], [93], [98] or a detailed examination of techniques from the same category [84], [93]. For example, [84] constructed eight document embeddings using five word embeddings (Word2Vec CBOW (Continuous Bag of Words), Word2Vec SG (Skip-Gram), FastText CBOW, FastText SG, and GloVe) and three transformer embeddings (BERT, RoBERTa, and BART).

b: QUANTITATIVE COMPARISON OF THE TEXT DATA PROCESSING TECHNIQUES

To enhance the clarity of the summary results (see Table 2), we have included a quantitative comparison of the text data processing techniques in Fig. 7. This comparison provides a visual representation of the differences between the text data processing techniques of the two categories of research - those that focus solely on news content and those that incorporate social context information - and helps the audience better understand the results.

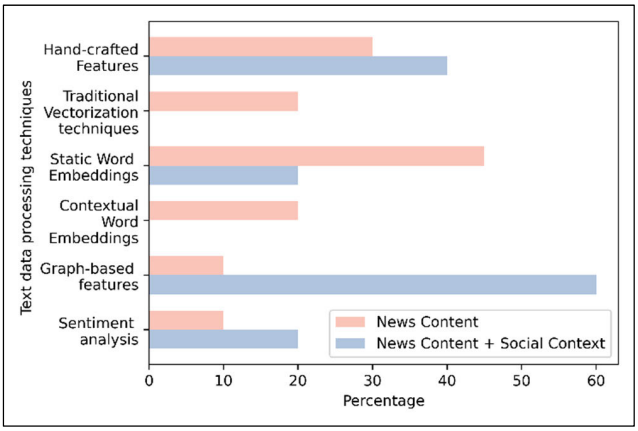


FIGURE 7. Quantitative comparison between textual features extraction methodologies, depending on whether the training data consists only of the news content or social context information (user profile, user posts, or social network information).

A review of the data presented in Fig. 7 reveals the following trends in the popularity of text-processing techniques:

- Graph-based features have the highest percentage when social context is included, indicating their effectiveness in handling complex social data.
- Static word embeddings are mainly used in studies that focus on news content, revealing a potential limitation in their ability to fully leverage additional social context.
- Hand-crafted features show a balanced usage between news only and news with social context.

c: TEMPORAL PATTERN

We include the time analysis of text processing techniques to ensure the correct interpretation of the results. The observations are as follows:

- Interest in hand-crafted news content features is evident between 2019 and 2023, with the greatest interest observed in 2021. Studies that include social context information in the training data focused on 2019 and 2021.
- Traditional vectorization techniques remain popular throughout the period, reflecting their relevance in research.
- Static word embeddings were explored between 2020 and 2024, with an emphasis on incorporating multiple embeddings in the later years, highlighting the evolving experimentation in embedding methods.
- In recent years, a notable shift has been towards more sophisticated techniques, such as contextual word embeddings and graph-based features. Research on contextual word embeddings shows an increase around 2020, reaching its peak in 2023. In the case of graph-based features.

- The range of studies in sentiment analysis covers the years 2019 to 2022. The lack of recent publications and the limited number of studies suggest that sentiment analysis may not be the primary focus of current research.

d: TEXT PREPROCESSING TECHNIQUES

Of the 30 articles analyzed, 13 mentioned text preprocessing. The following preprocessing techniques were mentioned: tokenization [84], [88], [95], [99], [100], [101], [102], lowercasing [84], [88], [97], [102], remove stop words [81], [85], [88], [90], [93], [94], [95], [97], [100], [101], [102], stemming [88], [93], [100], lemmatization [81], [84], IP (Internet Protocol) address and URL removal [89], [94], [95], punctuation and ASCII character removal [81], [84], [85], [94], [97], [99], [100], [101], [102], language filtering [99], contraction expansion [84].

2) APPLICATION OF ML AND DL TECHNIQUES

This subsection examines the ML and DL methods used in research focused on analyzing and classifying FNPD, both in the context of training data consisting only of news content and in the context of training data that includes new content with social context information. A summary of the ML and DL models is presented in Table 3.

TABLE 3. Summary of the ML and DL models used across the studies.

Model	Training data consists of news content only	Training data includes social context information for news content
Traditional ML algorithms	J48 [100], SMO [100], DT [100], SVM [103, 98], FFN [90,91], NB [81], Perceptron [81], MLP [81]	FFN [82], LR [80,61, 53], NB [80,61,53,79], DT [80,61,53], XGB [80, 79], AdaBoost [80], GradBoost [80], KNN [79, 61], RF [79, 53], SVM [79, 61, 53]
Recurrent Neural Networks	BiLSTM [96,81], LSTM [94,81], GRU [81], BiGRU [81]	LSTM [61]
Convolutional Neural Networks	CNN [104]	GCNN [63]
Capsule Neural Networks	Basic CapsNet [87]	
Hybrid approaches	CNN-RNN [95], CNN-LSTM [93,97], Conv-BiGRU-Attention-CapsNet [101] RNN-LSTM [102]	LSTM-FC-Node2Vec [83], CNN-3BiLSTM [85] 3BiLSTM [85]
Ensemble approaches	LR+RF+KNN [89], LR+ SVM+CART [89], LSTM+depth LSTM+ LIWC CNN+ N-gram CNN [86], RF+LR+DT [88]	LSTM+CNN [84]

a: IN-DEPTH ANALYSIS OF MODEL ARCHITECTURES

Let us take a closer look at the architecture of the models used. In traditional machine learning, different algorithms have been used to detect FNPD. The most popular traditional ML algorithms are decision tree (DT) [53], [61], [80], [100], support vector machine (SVM) [53], [61], [79], [98], [103], naive Bayes (NB) [53], [61], [79], [80], [81], logistic regression (LR) [53], [61], [80]. Less common techniques include sequential minimal optimization (SMO) [100], perceptron [81], J48 [100], multilayer perceptron (MLP) [81], XGBoost (XGB), [79], [80], AdaBoost [80], gradient boosting (GradBoost) [80], k-nearest neighbors (KNN) [61], [100], random forest (RF) [53], [79]. Although a feedforward neural network (FFN) is a simple form of neural network, this paper categorizes it under traditional ML techniques. The authors applied FFN to the aggregated node embeddings to predict the veracity of the news [82], [90], [91]. Within Recurrent neural networks (RNN), Long Short-Term Memory (LSTM) networks are used for both types of data: training data consisting only of news content or including social context information [61], [81], [94]. Gated recurrent unit (GRU) and bidirectional variants (BiLSTM and BiGRU), which process data in both forward and backward directions, are preferred for non-social contexts [81], [96]. The convolutional neural network (CNN) is primarily used without social context [104], while the graph convolutional neural network (GCNN) is specifically used when social interaction data is involved [63]. The GCNN model applies convolutional operations tailored to the graph structure of the data, which includes social networks and propagation patterns.

Ensemble approaches combine models from different families to take advantage of their different strengths, such as combining tree-based methods (RF, DT), statistical methods (LR), and advanced neural networks (LSTM, CNN). The most used algorithms in ensembles are LSTM [81], [86] and LR [88], [89]. Analysis of the hybrid approaches used reveals a complex combination of neural network architectures designed to leverage the unique strengths of each component to solve specific FNPD detection problems. LSTM and its bidirectional variant are often used in hybrid models [83], [85], [93], [97], [102]. CNN is the second most used architecture [85], [93], [95], [97]. In [101], advanced techniques such as attention and capsule networks are used.

b: QUANTITATIVE COMPARISON OF THE ALGORITHMS

The quantitative comparison between ML and DL techniques is shown in Fig. 8, where the algorithms are analyzed in two categories - those that focus solely on news content and those that incorporate social context information.

The findings related to the use of ML and DL models for FNPD detection are the following:

- Traditional machine learning algorithms are widely preferred in both categories of data - training data with only news content and training data with social context. In particular, their usage is significantly higher for social context-enriched datasets, indicating their adaptability to such data.
- Hybrid approaches are widely used in both categories, reflecting their effectiveness in combining multiple

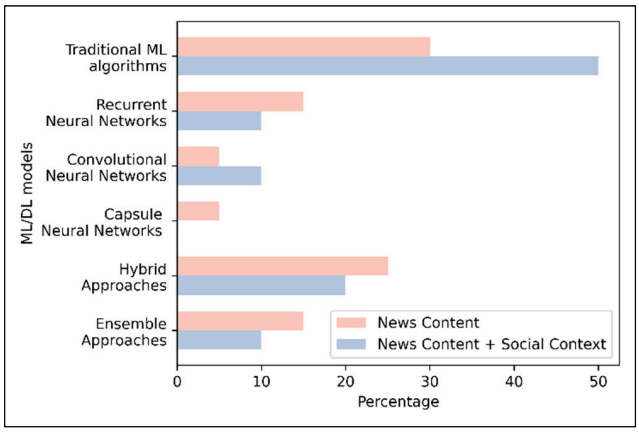


FIGURE 8. Quantitative comparison between ML and DL techniques, depending on whether the training data consists only of the news content or social context information (user profile, user posts, or social network information).

techniques for better results. Similarly, ensemble methods follow this trend, reflecting their utility in improving predictive accuracy through model combination.

- RNNs show a balanced usage in both categories, with a slightly higher percentage for news content only, while CNNs are also moderately used in both categories, but with a slightly higher percentage for news content with social context.

It is important to note that some studies (such as [61], [81]) use multiple algorithms, as they are listed under different categories. This was done to determine which model is better suited to a particular type of data or problem and to learn how different algorithms can be combined to create a more accurate model, which is particularly useful when developing hybrid or ensemble methods.

c: TEMPORAL PATTERN

To further our understanding of the results, a temporal analysis was performed to illustrate the evolution of these methods over time. The main findings of this analysis are as follows:

- Traditional ML techniques have remained consistently popular over the years, with a noticeable interest in Feedforward Neural Networks (FFN) in recent studies (2023-2024), highlighting a growing interest in these models.
- The RNN was first introduced in 2021. This coincides with the beginning of the use of CNN.
- Ensemble methods have been used steadily since 2021, with their use increasing significantly in 2023, indicating their growing importance in achieving robust performance.
- From 2021, hybrid approaches have been observed.

3) TEMPORAL ANALYSIS SHOWS THAT THERE IS A STRONG BIAS TOWARD REAL-TIME FNPD NEWS DETECTION AND LOGICAL FALLACIES

Several approaches and methodologies have been proposed in recent studies to effectively address the challenges of fake news detection in real-time scenarios. Article [63] explores how the performance of a fake news detection model changes over time, especially when applied to new data in real-time. By investigating the performance of the model on URL- and cascade-wise settings, the study aims to emulate real-world scenarios where the model trained on historical data is applied to new tweets in real-time. The proposed framework in the paper [61] focuses on detecting fake news on the Facebook platform, specifically on users' home pages. The experimental results show that the framework was successfully deployed in a Chrome environment, analyzing user information and shared posts to detect fake news in real-time. The paper [82] introduces a new way to study how news spreads in social media. The framework uses a temporal graph attention network (TGAT) to capture the structures, content, and temporal information of news propagation. Experiments on real-world data show that TGNF outperforms other methods for detecting fake news. In addition, the paper [80] uses social media data to detect fake news. This method improves early detection by using different signals from users, posts, and networks. It is effective for real-time scenarios.traditional machine learning methods in the early years, followed by a shift toward more sophisticated neural networks, including RNNs and CNNs. On the other hand, traditional ML techniques have not yet been abandoned, suggesting that they still have value in certain combinations with more advanced methods, i.e. ensemble approaches. The increase in ensemble approaches in recent years reflects a growing recognition of the benefits of ensemble methods in improving the accuracy of predictive models. A more recent trend to combine multiple approaches to exploit their collective strengths also indicates the emergence of hybrid approaches. While an ensemble approach combines different models to form a consensus, a hybrid approach combines different techniques or algorithms to exploit their advantages within a single framework.

4) MODEL PERFORMANCE ANALYSIS

We perform a comparative analysis of text classification models presented in 30 peer-reviewed scientific articles. The main goal of this analysis is to find out which datasets are used and to identify the models that perform best on each dataset. This comparative framework demonstrates the performance of different models and serves as a reference for future FNPD classification studies.

The majority of the papers analyzed obtained their data from pre-labeled databases, with a particular emphasis on ISOT [87], [89], [94], [95], [100], [101], BuzzFace [84], [98], [100], [103], BuzzFeed [79], [81], [96], [100], FNC (Fake News Corpus) [81], [85], [88], [93], and Kaggle databases [81], [85], [89], [92], [94], [98], [103], [104]. The datasets PolitiFact and GossipCop, which are currently included in the FakeNewsNet database [53], [69], [80], [81], [82], [83], [96], [97], and SemEval datasets (for SemEval-2020 Task 1 [90], [91] and the SemEval-2017 Task 8 [96]) are also considered. Some datasets, such as

QProp [90], [91], WELFake [97], [98], Fa-KES [95], [101], LIAR [81], [87], and TSHP-17 [81], [91], are mentioned twice. There are 9 datasets mentioned only once in the analyzed papers. These include Weibo [82], FakeNewsPrediction [97], CAAprop [90], TI-CNN [94], SMS Spam [94], Snopes [96], NewsTrust [96], Reuters [98], Fakeddid [85]. Twitter has become a commonly used platform for collecting fake news datasets directly from the platform's API (Application Programming Interface) or established datasets derived from this source [63], [84], [86]. In some cases, the authors [61], [102] manually gathered user posts to create datasets. One dataset [102] included Turkish fake and real news tweets. The distribution of the datasets used is shown in Fig. 9.

After analyzing the datasets, we started to evaluate the performance of the classification models applied to these datasets. Models were evaluated based on their metrics, specifically accuracy (Acc), precision (Prec), recall (Rec), and F1 measure (F1). The performance analysis results are presented in Table 4, which lists all the datasets and the model that performed best according to the identified metrics.

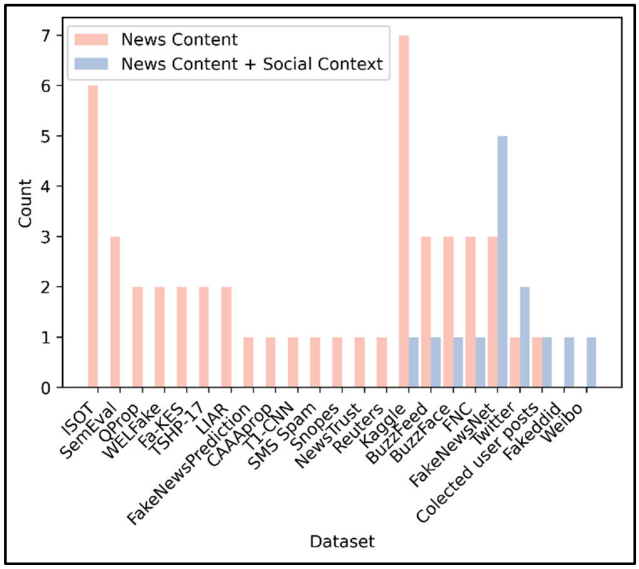


FIGURE 9. Dissemination of the datasets used, depending on whether the training data consists only of the news content or includes social context information (user profile, user posts, or social network information).

In contrast to real-time systems, the authors of [90] and [91] recognize that logical fallacies are an integral part of propaganda strategies. The G-HFIN framework proposed in [90] incorporates logical fallacies into a graph-based model using semantic, syntactic, and sequential features. Although the work focuses on integrating textual features to improve the detection of propaganda techniques without focusing on logical fallacies, it highlights the importance of logical fallacies and the need for further research.

Table 4 demonstrates that the combination of Word2Vec and LSTM consistently achieves high accuracy (≥0.98) across all textual datasets, demonstrating its robustness. Similarly, FastText combined with Hybrid (CNN+LSTM) performs consistently across all datasets and achieves near-perfect accuracy (0.99).

Conversely, the effectiveness of the models is significantly influenced by the dataset's characteristics. For example, on the ISOT dataset, all models that use the ISOT dataset achieve high accuracy, often above 99%, due to its structured nature, balanced classes, and relatively low noise. This includes approaches such as capsule neural networks [87] or hybrid CNN-RNN models [95].

More complex datasets, such as LIAR or Fa-KES, which contain noisy data, challenge the generalization capabilities of the models, resulting in lower performance metrics. Social context-based datasets such as BuzzFace, Twitter15, and Twitter16 achieved moderate accuracy (0.78-0.8), demonstrating the challenge of effectively integrating textual and social features.

The variability of the model's performance within the complexity of the data set is illustrated in Fig. 10.

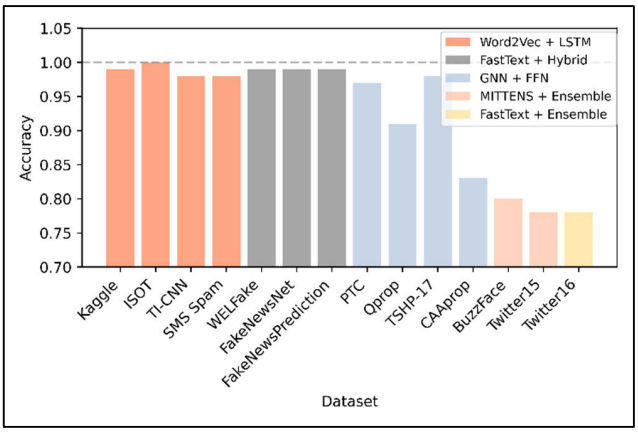


FIGURE 10. Model accuracy on different datasets.

The GNN + FFN model shows variability in performance, with robust results on structured datasets such as PTC and TSHP-17, but challenges on datasets such as Qprop and CAAprop (see Fig. 10). The results of Truică et al. [84] show moderate performance on social graph focused datasets such as BuzzFace, Twitter15, and Twitter16. These datasets, which integrate social network information alongside textual features, present unique challenges due to their multimodal nature and inherent complexity.

C. SOCIAL IMPACT (DOMAIN #3)

The following meta-analysis of 30 selected articles provides an overview of FNPD social impact research. After careful consideration, we have identified several key and specific meta-analysis criteria, see Fig. 11. The first subsection presents the former, the second the latter.

First, some statistics. Selected articles were cited on average of 69 times, average publication date 2021, average use of the term 'social' 82 times. Datasets used: 55.56% Twitter, 5.56% BuzzFeed, 5.56% NELA-GT-19 and Fakeddit source,

TABLE 4. Comparison of FNPD classification model performance on different data sets.

Dataset	Authors, year	Data type	Input Feature	Classification model	Results (%)			
					Acc	Prec	Rec	F1
ISOT	Mallik, 2024 [94]	textual content	Word2Vec	LSTM	1	1	1	1
Kaggle	Mallik, 2024 [94]	textual content	Word2Vec	LSTM	0.99	0.99	0.99	0.99
WELFake	Hashmi, 2024 [97]	textual content	FastText	Hybrid model (CNN+LSTM)	0.99	0.99	0.99	0.99
FakeNewsNet	Hashmi, 2024 [97]	textual content	FastText	Hybrid model (CNN+LSTM)	0.99	0.99	0.99	0.99
FakeNews Prediction	Hashmi, 2024 [97]	textual content	FastText	Hybrid model (CNN+LSTM)	0.99	0.99	0.99	0.99
TI-CNN	Mallik, 2024 [94]	textual content	Word2Vec	LSTM	0.98	0.98	0.98	0.98
SMS Spam	Mallik, 2024 [94]	textual content	Word2Vec	LSTM	0.98	0.98	0.98	0.98
TSHP-17	Liu et al., 2024b [91]	textual content	GNN	FFN	0.98	0.98	0.98	0.98
FNC-1	Umer et al., 2020 [93]	textual content	Word2vec	Hybrid model (CNN+LSTM)	0.98	-	-	0.98
PTC	Liu et al., 2024b [91]	textual content	GNN	FFN	0.97	0.97	0.97	0.97
Weibo	Song et al., 2021 [81]	topological structure+textual content+temporal information	GNN	FFN	0.97	0.92	0.92	0.92
Qprop	Liu et al., 2024b [91]	textual content	GNN	FFN	0.91	0.91	0.91	0.91
LIAR	Truica et al., 2023a [81]	textual content	GLOVE	LSTM	0.83	0.83	0.99	-
CAAprop	Liu et al., 2024a [90]	textual content	GNN	FFN	0.83	0.83	0.83	0.83
Fa-KES	Nadeem, 2023 [101]	textual content	Distil-BERT	Hybrid model (Conv+BiGRU+ Attention+CapsNet)	0.61	0.6	0.61	0.61
BuzzFace	Truicaa, 2024 [84]	textual content + social branches	MITTENS	Ensemble model (LSTM for textual content; STM+CNN for social branches)	0.8	0.75	0.8	-
BuzzFeed	Truica et al., 2023a [81]	textual content	BART	Perceptron	0.79	0.75	0.79	-
Twitter15	Truicaa, 2024 [84]	textual content + social branches	MITTENS	Ensemble model (LSTM for textual content; STM+CNN for social branches)	0.78	0.79	0.78	-
Twitter16	Truicaa, 2024 [84]	textual content + social branches	FastText	Ensemble model (LSTM+CNN for textual content and social branches)	0.78	0.79	0.78	-
Fakeddid	Truica et al., 2023b [85]	textual content+ users graph	Word2Vec	3biLSTM	0.76	0.77	0.76	0.76

5.56% Weibo source, 11.12% PolitiFact, GossipCop, 5.56% Socialsitu source and 5.56% multi-platform source. Network and behavior analysis of FNPD propagators is present in 72%, articles with real-time FNPD detection and social impact modeling 55.56%, geospatial data is used in 22%, analysis of propaganda techniques are detected in 22%, sentiment analysis is used in 56%, FNPD distribution pattern analysis is performed in 67%.

1) KEY CRITERIA

The main novelties of the selected articles can be summarized from different perspectives as follows below. The key analysis criteria by author that have been analyzed in this subsection are presented in Appendix E. The cross-tabulation there shows more detailed linkages and some aggregated estimates.

a: NOVELTY

i) GRAPH-BASED LEARNING AND PROPAGATION PATTERNS

Deep learning tailored for graph-structured data, such as the novel geometric deep learning approach, the ''Dynamic GCN'' for dynamic rumor representation, and the ''Propagation2Vec'' for utilizing partial propagation networks, highlights the emphasis on capturing the dynamics and patterns of information spread in networked structures [63], [104], [105].

ii) CONTENT AND USER INTERACTION FUSION

There is an evolving focus on combining content analysis with user interactions and behaviors. The ''DeepFakE'' model integrates news content with echo chambers' existence. The ''DSS'' approach analyzes propagation tree and stance network features. Furthermore, the Graph-aware Co-Attention Networks (GCAN) aims to validate tweet veracity based on its retweeters' sequence [83], [106], [107].

iii) BOT DETECTION AND INFLUENCE

Research has shifted from just identifying bots to understanding their behaviors and impact. The introduction of an adaptive deep Q-learning model for bot detection and the identification of bots that interact more with humans shows the sophistication in tackling bot-driven disinformation. The concept of ''generalized harmonic influence centrality'' quantifies the influence of these bots on networked opinions [108], [109].

iv) ROLE IDENTIFICATION AND INFILTRATION

There is a focus on understanding user roles and the hidden manipulators within online social networks. This is seen in

the novel approach to classifying Twitter users based on their roles and the investigation into human-controlled sock-puppets, particularly "infiltrators," who blend into genuine online communities [110], [111].

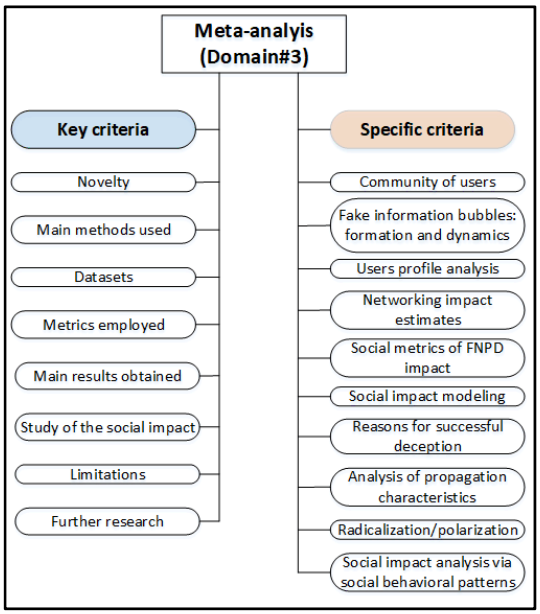


FIGURE 11. Key and specific criteria used in the meta-analysis (domain#3).

v) HOLISTIC APPROACHES AND COMPREHENSIVE DATA

The introduction of comprehensive data repositories like "FakeNewsNet" and systems like "FakeNewsTracker" highlight the shift towards creating holistic solutions and benchmarking platforms to combat misinformation on social media [18], [72].

vi) ADVANCED ANALYTICAL FRAMEWORKS

Several novel frameworks have been proposed to detect and understand disinformation. The combination of actor-network theory with deep learning, the use of social situation analytics for trend identification, and the study focusing on activities of Russian trolls during the U.S. Presidential election display the intersection of sociological, political, and computational methods in addressing the issue [112], [113], [114].

vii) DISINFORMATION THROUGH NETWORK EFFECTS

Social media platforms are designed in a way that can unintentionally amplify false information. Closed networks of echo chambers, AI-based information filtering/profiling, and the way users interact online contribute to this. This narrative focuses on the mechanisms within social media that make it fertile ground for FNPD [115], [116], [117], [118], [119].

viii) DETECTING DISINFORMATION CAMPAIGNS

This narrative highlights the ongoing effort to detect and counter coordinated disinformation campaigns. Researchers are developing new tools to identify and track the spread of coordinated disinformation. These tools analyze online interactions and user behavior to pinpoint suspicious activity [120], [121], [122].

ix) MEASURING ECHO CHAMBERS' POLARIZATION

Another rapidly growing concern is the rise of echo chambers and how they contribute to social polarization. Researchers are proposing new metrics to quantify this phenomenon, aiming to understand how social media shapes ideological divides. This narrative focuses on the impact of social media on social and political discourse [123], [124].

x) COGNITIVE WARFARE

According to recent research on information manipulation and interference, echo chambers have become crucial weapons in the arsenal of Cognitive Warfare for amplifying the effect of psychological techniques aimed at altering information and narratives to influence public perception and shape opinions [125], [126].

The fight against FNPD on social media is constantly developing. Lately it is tackled from multiple fronts, employing advanced computational techniques, rigorous data collection, and in-depth sociological insights. Fig. 12 shows the temporal dynamics of novelty in the domain#3.

b: MAIN METHODS USED

Focusing on the social impact aspects of the methods used, here is a consolidated view from different perspectives:

i) NETWORK & GRAPH-BASED TECHNIQUES

Geometric deep learning (GDL - graph-structured data for recognizing inter-relational dynamics [63], label propagation (method used to infer the ideological leanings of users within a network, demonstrating how beliefs or labels may spread in OSNs) [113], graph convolutional networks (GCN) with attention mechanisms (captures evolving rumor propagation patterns in social structures, emphasizing the temporal dynamics [105]. *DSS model* (incorporates dynamic, static, and structural analysis to understand how information or content traverses through OSNs [83], network-based pattern-driven model (focuses on extracting features from patterns of fake news dissemination on social platforms) [52].

ii) SOCIAL CONTEXT & INTERACTION ANALYSIS

Coupled matrix-tensor factorization (captures relationships between news content and its social context, such as echo chambers and user profiles) [106], deep Q-network architecture (DQL) (by treating each social attribute of a user as a state, this method conceptualizes the dynamics of social behaviors and interactions) [109], GCAN model (integrates word embeddings, neural networks, and a dual co-attention mechanism to analyze correlations between source content, retweet propagation, and user interactions) [107], Two-Pronged approach (divides the social user circle based on

content dissemination and contextual information, portraying how users are influenced by and engage with different content types) [112].

iii) BOT & USER ROLE ANALYSIS

Advanced machine learning techniques for bot detection (underscores the non-human entities that might manipulate social dynamics online) [113], using a statistical physics model (to identify bots and measure their influence on shifting opinions within OSN [108], hierarchical self-attention neural network (delineates how different user roles might influence or be influenced in social contexts) [110], supervised machine learning guided by journalistic investigations (by integrating journalistic insights, this method underscores the human-social perspective in validating and understanding online content) [111].

iv) COMMUNITY DETECTION AND DYNAMICS MODELLING

Agent-based simulation (simulates the behavior of individual users within a social network to understand how information spreads) [115], physics-informed neural networks (modeling of complex social systems) [118], system dynamics modeling (explores how interconnected parts of a system influence each other over time) [120], [121], community detection algorithms (identify groups of users within a network who are more likely to interact with each other) [117], user embedding models (analyzing how users with similar ideologies connect) [123], network distance measures (these techniques measure how ''far apart'' users are within a network, potentially indicating how likely they are to be exposed to opposing viewpoints) [119], scaling law analysis (explores how different aspects of a system change in relation to each other) [124].

Brief Summary of Key Methodic Insights and Innovation Trends: There is a clear temporal shift from traditional machine learning methods in 2018 towards more complex deep learning and neural architectures in subsequent years. The year 2020 saw a diverse range of techniques being employed, while 2021 strongly leaned towards graph networks and attention mechanisms. By 2022 and 2023, there is an observable trend towards integrating multiple complex techniques, like transformer architectures, attention mechanisms, agent-based simulations, and system dynamics to address fake news detection, user interaction analysis, and social networks development. This suggests a rising trend in favor of deep learning approaches, with traditional machine learning and specific analytical methods still being quite prominent in the research. Attention mechanisms and graph-based techniques, while less frequent than deep learning, have a notable presence, indicating their increasing importance in the realm of fake news detection and analysis of social media data.

c: DATASETS

We looked at the datasets used from a few perspectives, which are listed below.

i)SOURCES OF DATASETS

Twitter-Based Datasets (44.4%): specific news stories (5.6%), election-related tweets (5.6%), generic Twitter datasets (33.3%); Fact-Checking Websites (22.2%): BuzzFeed & PolitiFact (5.6%), PolitiFact & GossipCop (16.7%). Mixed or Multi-Modal Datasets (5.6%). Datasets with Unspecified Origins (16.7%): Unspecified Real-world Datasets (11.1%), social Networks & Geo-Political Issues (5.6%). Specific or Unique Datasets (11.1%): Kyrgyzstan-focused (5.6%), Socialsitu Metadata (5.6%). Self-Collected Datasets (5.6%).

ii) VERIFICATION MECHANISMS

Fact-checking organizations (like Snopes, PolitiFact, Buzzfeed) (5.6%); U.S. Congress investigation for troll identification (5.6%); Whistleblower insights (5.6%).

iii) SIZE OF DATASETS (WHERE SPECIFIED)

largest 18,58,575 entries [112], smallest 30,000 tweets [114].

Brief Summary of Key Data Used: Twitter emerges as the most popular platform for sourcing datasets, being used in nearly half (44.4%) of the studies. Articles also prominently utilize fact-checking platforms such as PolitiFact and GossipCop, featuring in over a fifth (22.2%) of the studies. A minority of studies (16.7%) use unspecified real-world datasets. Some datasets have been specifically curated or tailored for specific research purposes, such as Socialsitu metadata (total 11.1%). Verification mechanisms for data authenticity and accuracy include external fact-checking organizations, governmental investigations, and whistleblower insights.

d: METRICS EMPLOYED

We examined the metrics employed from several different perspectives, as outlined below.

i) POPULAR METRICS USED

Accuracy [129, 107-109, 111, 115], precision [106], [109], [113], recall [106], [113], [127], F1-Score [104], [127], ROC AUC [63].

ii) NETWORK ANALYSIS & MODELING

User segregation [115], metrics of systems dynamics [116], community detection algorithms [117], community detection algorithms [123], network distance measures [119], opinion dynamics [124], consensus metrics [125].

iii) AUXILIARY/ADDITIONAL METRICS & METHODS

Descriptive statistics [113], early detection rates [83], [129], linguistic features and social engagements [72], shift in equilibrium opinions [108], statistical indicators (Lorentz curve and Gini coefficient) [112].

Articles with ambiguous or not explicitly mentioned metrics: [18], [52], [72], [104], [105], [106], [111], [114].

Brief Summary of Key Metrics Used: Accuracy emerges as the most popular metric used across the studies, being

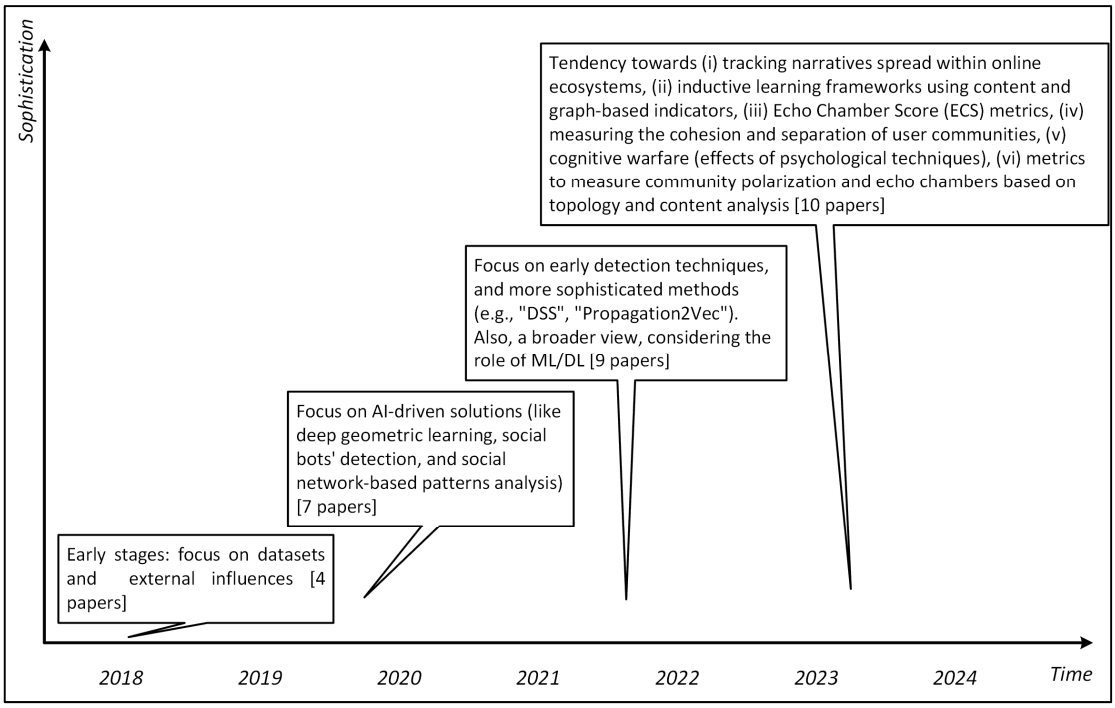


FIGURE 12. Key innovation trends for the period 2018-2024.

explicitly mentioned in a third of the articles. *Precision and Recall* are also prominent metrics, together appearing in a third of the articles. There is an interest in using additional descriptive and statistical metrics to provide a comprehensive understanding of the datasets, as seen in articles like [112], [113]. A notable portion of the articles (44.4%) do not provide explicit details on the metrics employed, instead hinting at the use of common or state-of-the-art measures for evaluation or focusing on the overarching goals of the research rather than metric specifics.

e: MAIN RESULTS OBTAINED

Considering the primary findings derived from the articles within this research domain, here is a summarized view from different perspectives.

i) FAKE NEWS DETECTION EFFICIENCY

High accuracy in fake news detection was observed in multiple models. Article [63] achieved a 92.7% ROC AUC, the DeepFake model in [106] obtained validation accuracies of 85.86% and 88.64% on two different datasets, and the model in [128] demonstrated improved accuracy and early detection capabilities compared to existing methods. In article [127] model distinguished between real and fake news with 90% accuracy, and the DSS model [83] surpassed state-of-the-art methods by up to 8.2%. Meanwhile, the network-based pattern-driven approach [52] was robust against manipulations and effective even with limited network data, and Propagation2Vec from [104] outperformed other models by up to 5.55% in F1-score. GCAN, from [95], significantly outperformed existing methods, and the model in [110] boosted its accuracy when combined with a transfer learning scheme.

ii) BOT DETECTION AND INFLUENCE

Significant information about bots emerged from the articles. Article [113] found that 4.9% of liberal users and 6.2% of conservative users were bots. Article [18] observed that bot users are more involved in spreading fake news, while Article 96's Ising model algorithm efficiently identified bots. Article [111] unveiled that as bot detection methods improve, disinformation agents are now more focused on using sock-puppets, especially infiltrators. Article [114] highlighted the critical role of bots in influencing online public opinion and spreading false narratives.

iii) INSIGHTS ON CONTENT AND DISSEMINATION

Article [113] provided a breakdown of Russian troll content, highlighting that it had a conservative, pro-Trump agenda. It also noted that conservatives retweeted Russian trolls at a rate 36 times higher than liberals, with most troll content originating from the Southern states. Article [90] uncovered that while disinformation arises across various platforms, it spreads more predominantly on its original platform. The research also discerned four distinctive disinformation propagation trends.

iv) MODEL ARCHITECTURE AND METHODS

Several articles introduced unique model architectures and methods. Article [109] deep Q-learning algorithm integrated

with various social attributes demonstrated improved precision over other algorithms. The Dynamic GCN from [99] outperformed other leading methods in rumor detection. In Article [72], FakeNewsTracker was effective in using linguistic and social engagement features for fake news detection. Article [95] GCAN highlighted suspicious retweeters and specific tweet segments, adding a layer of explainability to the model. Lastly [108] emphasized the use of the Ising model from statistical physics for bot detection.

v) ECHO CHAMBERS AND POLARIZATION RESEARCH

Ideological segregation in social networks increases the spread of false information by creating local infrastructures that align with biased partisans [115]. Confirmation bias, sharing of posts, and algorithmic ranking are critical variables driving this process [116]. Coevolving dynamics of opinions and network structures can lead to stable bipolarized community structures, with phase transitions across different polarization phases [124]. An inductive learning framework identified how echo chambers foster polarization and dysfunctional political discourse [117]. Complementing topology-based metrics with semantic analysis of viewpoints and beliefs is essential to fully capture community closeness and prevailing beliefs [125]. Studies propose methodologies for identifying narratives, estimating underlying dynamics, and quantifying polarization levels in social networks, considering opinion variations, community assortativity, and the interplay between opinions and network structures [119]. Overall, the research underscores the complex interplay between network structures, algorithmic mechanisms, confirmation biases, and the dissemination of misinformation, leading to echo chambers and exacerbating polarization in online communities.

vi) ALGORITHMIC MECHANISMS AND COUNTERMEASURES

Computational techniques and frameworks for identifying coordinated manipulation campaigns and disinformation operations are a major focus. For instance, an inductive learning framework determines content- and graph-based indicators of coordinated manipulation, encodes abstract signatures using graph learning, and evaluates generalization capacity across operations of influence [121]. Systems for identifying prevalent narratives and aiding fact-checkers in addressing misinformation more quickly are also highlighted [122]. Social media algorithms monitor user behavior, interests, and actions to recommend relevant content, refining suggestions by adapting and learning from user interactions [115], [121], [126]. Policymakers are encouraged to adopt a portfolio approach, pursuing a diversified mixture of counter-disinformation measures, including fact-checking, foreign sanctions, algorithmic adjustments, and counter-messaging campaigns [116], [117]. The research emphasizes the importance of developing computational techniques, analyzing underlying dynamics, and proposing policy interventions to combat the harmful effects of propaganda and misinformation on social media platforms.

Brief Summary of Main Results: Multiple models demonstrated high accuracy in detecting fake news, with some achieving over 90% accuracy. Studies revealed that bots play a crucial role in spreading fake news and influencing public opinion. Research uncovered distinct disinformation propagation trends across platforms, with content typically spreading more on its original platform. Novel approaches like deep Q-learning algorithms, Dynamic GCN, and FakeNewsTracker showed improved precision in fake news detection. These models often incorporated linguistic features and social engagement data for better accuracy. Studies highlighted how ideological segregation in social networks facilitates the spread of false information. Research emphasized the complex interplay between network structures, algorithmic mechanisms, and confirmation biases in exacerbating polarization. Researchers developed computational techniques for identifying coordinated manipulation campaigns. Policymakers were encouraged to adopt a diversified approach to counter disinformation, including fact-checking, algorithmic adjustments, and counter-messaging campaigns.

f: STUDY OF THE SOCIAL IMPACT

Here is an overview of the main social impact assessments from different perspectives in this domain of research.

i) FAKE NEWS IMPACT ON POLITICAL EVENTS AND DEMOCRACY

The substantial societal consequences of fake news during political events like the US 2016 elections and Brexit are highlighted, with a specific emphasis on their potential threat to democracies [63]. The dissemination of misinformation can heavily influence democratic discussions, leading to societal confusion and potential instability [113]. The spread of fake news on platforms, particularly during major events like the 2016 U.S. Presidential Election, carries notable societal ramifications, including financial, political, and emotional [52], [83], [106]. Instances like the anti-vaccine misinformation during the COVID-19 pandemic underscore the importance of addressing the challenge of fake news [104], [112], [128].

ii) SOCIAL BOTS AND THEIR INFLUENCE

The ability of social bots to spread misleading information, manipulate public sentiment, and compromise the integrity of networks makes their detection vital [109]. The presence of politically motivated bots on OSNs poses a considerable threat to democratic processes [108]. The acceleration of information spread, both factual and fictitious, by social bots emphasizes the need for thorough research to mitigate potential threats [114].

iii) REAL-WORLD CONSEQUENCES OF MISINFORMATION

The broad challenges posed by fake news include the potential to shift genuine news dynamics, influence public perceptions, and even affect tangible events such as elections [18]. Events such as the ''Pizzagate'' tweets during the

US elections provide tangible evidence of the consequences of misinformation [72]. The proliferation of false news can potentially benefit certain factions unjustly, whether in political, economic, or psychological domains [107].

iv) INFILTRATION AND MANIPULATION BY DIGITAL AGENTS

The changing landscape where disinformation agents shift towards meticulously designed infiltrators that have the potential to genuinely sway beliefs and viewpoints highlights a significant threat to authentic discourse [110], [111]. Recognizing the roles of various bots and entities offers deeper insights into the dynamics of misinformation spread on digital platforms [110].

v) IMPACT FRAMEWORKS AND COUNTERMEASURES

Several studies highlight the detrimental effects of echo chambers and polarization fostered by the spread of disinformation on social media platforms [117], [123], [125]. They emphasize how echo chambers can make political discourse dysfunctional and exacerbate polarization in open societies, contributing to the identification of problematic interaction patterns [116], [119]. They develop a comprehensive frameworks to accurately simulate information and counter-propaganda spread, evaluating performance on real-world data and providing insights into factors influencing information warfare. They also suggest countermeasures to combat disinformation include legislation to hold social media platforms liable for illegal content, mandatory licensing, and the establishment of independent statutory authorities to adjudicate minimum epistemic and moral standards countermeasures like legislation holding social media platforms liable for illegal content, mandatory licensing, and independent authorities to adjudicate minimum standards [126]. Overall, these studies emphasize the social impacts of echo chambers, polarization, and rapid misinformation spread, while proposing detection methods, metrics, frameworks, and policy interventions to address these issues and their consequences for political discourse and societal well-being.

Brief Summary of Social Impact Studies: Fake news has been shown to have substantial societal effects, particularly during major political events. Social bots play a crucial role in spreading misleading information and manipulating public sentiment, posing considerable risks to democratic processes. The research also highlights the growing threat of sophisticated digital agents, such as infiltrators, that can genuinely sway beliefs and viewpoints. To combat these issues, studies propose various countermeasures, including legislation to hold social media platforms accountable, mandatory licensing, and the establishment of independent authorities to adjudicate minimum standards. Overall, the research emphasizes the urgent need to address the challenges posed by misinformation and its impact on political discourse and societal well-being.

The development of trends in the analysis of the social impact of FNPD has only recently gained momentum, see Fig. 12. The spread of reactions on social networks is obviously a very significant part of the most associated studies. However, researchers make the core assumption (unfortunately not always correct) that people's reactions on social networks are a direct reflection of their attitudes and behavior. In particular, there is a large gap between people's reactions in OSNs and their actual behavior.

In Appendix E (Table 6), the bold total numbers per article indicate the dominance of such sub-criteria: popular metrics used, community detection and dynamics modelling, articles with ambiguous or not explicitly mentioned metrics, efficiency of fake news detection, popular methodological metrics, ambiguous or not explicitly mentioned metrics, efficiency measures for fake news detection, and impact of FNPD on political events and democracy. We also see that papers [83], [106], [113], [117] cover the key criteria well.

2) SPECIFIC CRITERIA

Below we present results of the meta-analysis according to the specific criteria chosen mainly to find out the extent to which the social impact aspects of the FNPD were explored in the selected articles, see Appendix F (Table 7). The main novelties can be summarized from different perspectives as follows below. The specific analysis criteria by author that have been analyzed in this subsection are presented in Appendix F. The cross-tabulation there shows more detailed linkages and some aggregated estimates.

a: COMMUNITY OF USERS

Around 70% of the articles used community analysis in their research. In this regard, presented meta analysis revealed a few key aspects like i) construction of user community matrix (using user relationships in the dataset; these matrices help identify echo chambers), ii) analysis of social contexts (such as posts, likes, shares, replies, and user interactions with news articles), iii) centrality measures (like clustering coefficient, betweenness centrality, and closeness centrality to understand user interactions and information propagation), iv) reliability determination characteristics of users to determine their reliability in sharing news), v) dynamic analysis (a dynamic graph convolutional network-based model to understand evolving patterns of rumor propagation), vi) sentiment and structural analysis (a model for determining news article veracity through the analysis of propagation tree and stance network features).

This domain of research is dominated by the above types of analysis. More recent approaches like [123] leverages an embedding space to measure the cohesion and separation of user communities, providing insights into the echo chamber effect. It presents EchoGAE, a self-supervised user embedding model that captures ideological similarities among users and generates accurate embeddings to facilitate measuring distances between users. The article [125] discusses a solution to analyze community members' opinions on a topic by discriminating different opinions of the same user on different aspects of the topic through Aspect-Based Sentiment Analysis (ABSA). It employs consensus metrics in Group

Decision-Making (GDM) to measure community polarization and echo chambers based on topology and content analysis. The measure in the article [119] is based on the generalized Euclidean distance, which estimates the distance between two vectors on a network representing people's opinions.

b: FAKE INFORMATION BUBBLES (FORMATION AND DYNAMICS)

Around 67% of the articles incorporated information bubbles and echo-chambers analysis. For instance, in [106], the analysis of fake information bubbles involves considering the content of news articles and the existence of echo chambers in the online social network. A tensor representing social context is used to combine news, user, and community information, and matrix-tensor factorization is employed to represent news content and social impact. Article [18] emphasizes the significance of user networks on social media for fake news detection. It underscores the value of extracting network-based features to represent echo chamber cycles. Article [112] focuses on the formation of a social user circle (group) based on the content sequence and social contextual information of users associated with disinformation.

c: USER PROFILE ANALYSIS

Around 53% of the articles used users profile analysis. For instance, in [113], Twitter users are labeled as liberal or conservative based on their tweet behavior, specifically focusing on the number of tweets with links to liberal or conservative sources. Article [106] utilizes a news-user engagement matrix to represent user responses to news articles in terms of sharing, indicating an analysis of user interactions. Article [128] emphasizes the importance of determining the credibility of social media users for FNPD detection. It mentions the adoption of a zero-shot learning approach to build the user credibility module. Article [127] introduces user profile classification, with a system based on deep neural networks proposed for classifying user-related social features. In [104], user-based features, including whether the user is verified, number of followers, friends, lists, favorites, tweets, mentions, and more, are studied as part of the analysis. Article [95] explores the idea that user characteristics in real news propagations differ from those in fake news propagations, suggesting an analysis of user profiles to differentiate between real and fake news. Article [96] focuses on the detection of bots and their opinions, indicating an analysis of user profiles to identify automated accounts. The article [123] proposes EchoGAE, a self-supervised graph autoencoder-based user embedding model. It leverages users' posts and the interaction graph to embed users in a manner that reflects their ideological similarity.

d: FNPD DISSEMINATION ANALYSIS

Around 63% of articles included analysis of FNPD network estimates. For instance, in [113], the analysis includes estimates in terms of retweeting FNPD tweets and web sources, focusing on the impact of FNPD dissemination. Article [106] discusses impact estimates in terms of echo chambers, aiming to understand how information spreads within specific online communities. Article [83] proposes to analyze the propagation tree and stance network features for fake news detection, emphasizing the importance of impact analysis in detecting fake news. In [104], a hierarchical attention mechanism is proposed to encode propagation networks, which assigns varying levels of importance to nodes/cascades in propagation networks, indicating an impact analysis. The study [116] presents a rumor propagation model based on epidemiological models to address the spread of false news on social networking sites. The article [122] uses the large-language model MPNet and DP-Means clustering to analyze the spread of narratives originating from unreliable news websites. In [119], the measure is based on the generalized Euclidean distance, which estimates the distance between two vectors on a network, representing people's opinions, and analyzes the spread of polarization in real-world networks.

e: SOCIAL METRICS OF FNPD IMPACT

Around 50% of the articles used social impact metrics in one way or another. For instance, [63] discusses the manifestation of polarization in OSN and assigns credibility scores to users based on their interactions with true and fake news. It highlights the formation of distinct communities among credible and non-credible users. Article [113] explores the sharing of misinformation by Russian trolls on Twitter, focusing on the differences between conservatives and liberals in terms of content production and user engagement. Articles [72] and [109] involve social context analysis, particularly comparing SAF/A and SAF models in utilizing social context for fake news detection. They compare SAF/A (utilizing social context only) and SAF (exploiting both news article contents and social engagements) models for analyzing fake news. Article [83] presents a model with dynamic, static, and structural analysis components, emphasizing the use of a recurrent neural network for dynamic analysis of propagation patterns. Article [52] considers user susceptibility and influence as attributes for fake news detection, identifying patterns related to spreading behavior. Article [94] proposes a hierarchical attention mechanism for encoding propagation networks, allowing varying levels of importance assignment to nodes/cascades in the networks. The study [117] uses greedy modularity maximization and HITS metric to identify echo chambers in social networks, focusing on interaction patterns rather than content to detect problematic actors spreading disinformation. Introduces Echo Chamber Score (ECS), a novel metric for quantifying echo chambers and polarization in social media networks. The other study [123] introduces Echo Chamber Score (ECS), a novel metric for quantifying echo chambers and polarization in social media networks.

f: SOCIAL IMPACT MODELING

Around 50% of the articles incorporated social impact modeling. For instance, articles [63], [106], [113] introduce different techniques for impact modeling (using graph convolutional networks, machine learning, and presenting the DeepFake model). Article [99] discusses impact modeling in the context of retweets, exploring the influence and spread of information through retweet interactions. Article [52] delves into impact modeling at multiple network levels, including ego, triad, and community levels. Articles [104], [107] propose hierarchical attention mechanisms and dual co-attention mechanisms, respectively, for modeling the importance and influence of nodes or cascades in propagation networks. Article [104] also focuses on early propagation networks. Article [112] mentions identifying typical disinformation propagation trends based on propagation patterns and peak times, indicating a form of impact modeling. These articles delve into various aspects of social impact assessment, including metrics such as ROC AUC (Area Under Curve), echo-chamber estimation, graph snapshots, and spreading patterns. The analysis in these articles focuses on evaluating the effectiveness and consequences of information dissemination and its influence on society. The article [120] discusses the use of system dynamics as the main technique for designing a simulation model to analyze social media disinformation as a strategy for diplomacy. The research [115] employs agent-based simulation to argue that network segregation disproportionately aids the diffusion of implausible messages, favoring false news over true news. The article [121] introduces an inductive learning framework that determines content- and graph-based indicators to encode abstract signatures of coordinated manipulation, evaluating generalization across operations from different countries. The study [118] proposes a novel approach by integrating a modified logistic differential equation with a Physics-Informed Neural Network (PINN) to model population dynamics in the context of disinformation spread. This study [124] discovers a universal scaling law for opinion distributions in social systems, offering a framework to predict coevolving network dynamics and quantify different polarizing phases.

g: REASONS FOR SUCCESSFUL DECEPTION

This aspect of the study was used in about 40% of the articles. For example, articles [63], [72], [104], [110], [111] provide explanations for reasons contributing to the successful deception analysis. These articles discuss a range of factors or strategies that lead to successful deception. This article [115] argues that network segregation aids the diffusion of implausible messages, favoring false news over true news, which is a key reason for the success of deceptive information. The research [117] identifies problematic interaction patterns in social networks that facilitate the spread of disinformation by problematic actors, highlighting interaction patterns as a reason for successful deception. This study [125] discusses how echo chambers amplify the effect of psychological techniques aimed at altering information and narratives, which are crucial for successful deception in cognitive warfare.

h: ANALYSIS OF PROPAGATION CHARACTERISTICS

This aspect of the study was used in about 67% of the articles. Articles [18], [52], [63], [72], [83], [104], [112], [127] discuss propagation characteristics analysis. These articles explore various aspects of how information spreads, including temporal distribution patterns and propagation features analysis. The research [122] introduces a system to automatically identify and track narratives spread within online ecosystems, analyzing the propagation characteristics of narratives originating from unreliable news websites. The analysis in these articles focuses on understanding how false information disseminates within networks and across time.

i: RADICALIZATION/POLARIZATION.

This aspect of the study was used in only 33% of the articles. Article [63] explicitly mentions a 'bipolar analysis' focusing on the actions and reactions of liberals and conservatives on Twitter, suggesting that it explores political polarization on social media platforms. Article [111] also addresses radicalization or polarization, but no specific details are given. So, this is a huge niche to explore. t uses a self-supervised graph autoencoder-based user embedding model to capture ideological similarities among users and measure polarization. Recent studies analyzing radicalization and polarization from different perspectives include articles [119], [123], [124], [125].

j: SOCIAL IMPACT ANALYSIS VIA SOCIAL BEHAVIORAL PATTERNS

This aspect of the study was used in *only 37%* of the articles. Articles [111], [114] explore how individual and collective behaviors contribute to the social impact of false information. Article [108] addresses both OSN activities and social behavioral patterns analysis. Therefore, this is another niche for exploration. The following articles [115], [116], [117], [120], [121] study social impact analysis or social behavioral patterns.

In summary, a cross-examination of the meta-analyzed results in Appendix F (Table 7) provide the reader with the following information:

- which research approaches or criteria are most frequently used in the selected articles (see totals in the last column),
- which articles use specific research criteria and how many of these criteria are covered in their approaches (see totals in the last row).

Appendix F (Table 7) also provides a summary of the types of analysis (criteria) used in the selected articles. For example, articles [63], [108], [112], [125] used most analysis criteria. It is important to note that the highest number of articles met Authors and social (networking) impact multimodal analysis criterion (77%). This shows that the first and third domains of research are very closely interlinked

when studying the social dimension of FNPD. Other most prominent criteria in descending order are Community of users (70%), Fake information bubbles (67%), Authors and content multimodal analysis (67%), and Analysis of propagation characteristics (67%).

Another particularly important insight for further research comes from the fact that the last two criteria - radicalization/polarization (33%) and social impact analysis via social behavior patterns (37%) - are the least researched in the field. This is an astonishing revelation, as these two criteria should be the most important in FNPD social impact research. We imply that the ML/DL research community has walked around them, as they require modelling of complex social behavior patterns, multi-aspect clustering analysis and complicated social impact metrics. Although such additional research composition is not yet well established in the ML/DL field, it is likely that in a few years prospective research efforts will challenge this new frontier.

3) EMERGING RESEARCH TRENDS

In the authors' opinion, the presented meta-analysis helps to identify the most promising and fast-growing niches in the field of social impact research in FNPD, such as, cognitive warfare, echo chambers, polarization, opinion and influence dynamics. Let us briefly to comment just a few of them.

Cognitive warfare is distinguished by several crucial characteristics. It aims to influence entire populations, not just specific groups. Rather than simply disseminating false information on particular topics, it seeks to alter a population's behavior by fundamentally changing their thought processes. This approach relies on advanced psychological manipulation techniques and aims to undermine institutions, particularly governments. Often, this destabilization begins by targeting epistemic institutions like media outlets and academic institutions [126].

A key aspect of cognitive warfare is its exploitation of modern communication channels, especially social media platforms, which have become integral to how people consume information. The strategy often begins by exacerbating existing societal divisions and hindering cooperation within the target population. This is achieved by highlighting pre-existing differences and promoting extreme viewpoints across the political spectrum. Computational propaganda plays a significant role in these efforts. In short, researchers have identified cognitive warfare as either a non-kinetic component of traditional warfare (as seen in Russia's 2022 invasion of Ukraine) or as a form of conflict that falls short of outright war. Notably, it's often classified as a type of covert operation, which can be deployed during both wartime and peacetime scenarios [125], [126].

Another very closely related and promising research niche concerns coordinated operations of influence, which is an aspect of cognitive warfare aimed at influencing target audiences behavior during specific public events, such as presidential elections, parliamentary elections, referendums or other major political events [121].

It is also important to mention other highly relevant research niche in the field of social impact concerns echo chambers and polarization, which arise organically on OSNs as people gravitate towards groups that align with their interests and views. This tendency is reinforced by recommendation algorithms that cater to users' existing beliefs, biases and emotional responses. Recent studies on information manipulation highlight echo chambers as a powerful tool in cognitive warfare campaigns. By amplifying certain narratives and psychological techniques, echo chambers can significantly shape public perception and opinion. Given their potential impact, researchers are increasingly focused on developing methods to identify nascent echo chambers and track how they evolve over time. The goal is to better understand these digital enclaves and their role in spreading information - or misinformation - across social networks [106], [117], [119], [125]. Through this meta-analysis, the authors identify the three most advanced and promising areas of research on the social impact of FNPD: echo chamber research, influence operations, opinion dynamics, and cognitive warfare, see Fig. 13. From the overall research context, they are the most distinctive in terms of the diversity of contemporary methodological approaches and practical in preventing the most threatening societal impacts of the FNPD. The social impact of these areas is very high, as it has a direct impact on polarization and radicalization, segregation, distrust, behavioral changes, cognitive fatigue, etc.

Numerous studies have employed agent-based approaches to examine interactions between users in social networks and assess the extent of polarization. These methods frequently map out user interactions, providing valuable insights into the formation and strength of echo chambers [125]. For instance, researchers in [129] developed an agent-based framework incorporating Zaller's model of public opinion to investigate how individuals process and convert political information into their own views. Another study analyzes agents' behavioral patterns to determine their role in echo chamber formation [130]. Additionally, some research delves into how information consumption affects the proliferation of echo chambers. In [131], for example, the authors introduce an agent-based model of opinion dynamics that explores the relationship between news consumption habits and the spread of echo chambers within social networks.

There are other under-explored but promising areas of research. For example, various methodologies have been developed to assess opinion divergence in social networks [125]. The model presented in [132] enables the measurement of trust and distrust levels between agents, while the strategy outlined in [133] aims to reduce polarization by introducing new connections within a network. A content-based learning parameter linked to the network structure was introduced in [134] to quantify echo chambers in political discourse. Contemporary approaches also utilize influence diffusion models to gauge the echo chamber effect during information propagation [135].

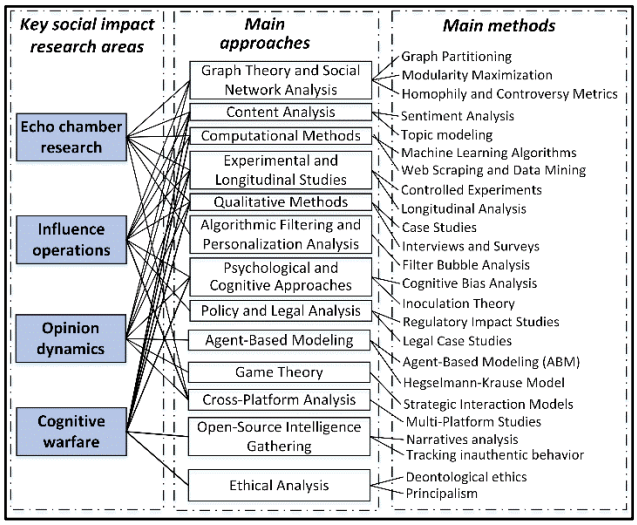


FIGURE 13. Relationships between main social impact research methodological approaches.

IV. FNPD RESEARCH DOMAIN LINKS AND THE FOUNDATION MODEL CONCEPT

In Part A of this chapter, we discuss observations and insights into relationships across all three research domains (authors/disseminators, content and social impact). In Part B, we present one of the main insights of this work - the prospect of developing a foundation FNPD model.

A. LINKING RESEARCH DOMAINS

A detailed examination of the progress in identifying FNPD from January 2018 to April 2024, highlighting the strategies used, data collected and innovative developments, is presented in Table 5.

Overall, we found that combined approaches effectively integrate and leverage the strengths of author-based, content-based, and dissemination-based methods. The results of our systematic review indicate that these integrated approaches consistently outperform stand-alone methods. We anticipate that this trend will continue to grow in future research.

The categorization is as follows: 11 articles covered domains [#1, #2], 9 articles addressed domains [#1, #3], 12 articles focused on domains [#2, #3], and 12 articles encompassed all three domains [#1, #2, #3]. In total, 44 out of the 90 reviewed studies adopted combined approaches, integrating multiple domains.

This distribution reflects an emerging trend toward more comprehensive research methodologies and underscores the strong interconnections among the three research areas. notably, a particularly strong relationship was observed between the first and third research domains.

1) COMBINING TEXTUAL CONTENT AND AUTHOR/DISSEMINATOR ANALYSIS (RESEARCH DOMAINS #1 AND #2)

Studies that integrate author and content analysis emphasize several key features, particularly focusing on user profile characteristics. For instance, [61] explores elements such as user behavior on social media, Twitter API data, and specific user profile details. Similarly, [64] introduces the ConspiDetector model, which combines psycho-linguistic features derived from tweet content and user profile characteristics. Another approach presented in [65] uses text embeddings and transfer learning to effectively incorporate both user profile information and textual content. Additionally, [74] assesses news content and publisher credibility within the FR-Detect framework, which integrates linguistic and publisher-related features such as Credibility, Influence, Sociality, Validity, and Lifetime.

Source characteristics are another crucial aspect. Studies like [61], [67], [69], [79] investigate the credibility of news sources, political bias, and the relationship between news articles and topics. These works introduce concepts such as domain localization and analyze news content features, including headlines and body text, to understand their impact on the FNPD detection process.

Research emphasizes the equal importance of both news content and author information in FNPD detection. For example, [60] demonstrated the critical role of both news content and user comments. This is further supported by [51], which found that models achieve optimal performance when trained using a combination of user profile and content features. Likewise, [76] showed that integrating a user response generator into the TCNN-URG model significantly enhanced the accuracy of fake news detection.

In the area of bot detection, three studies utilized textual content information to identify bot accounts disseminating FNPD on social media. These approaches incorporated various types of textual data, such as author profiles, article and topic descriptions [69], textual similarity analysis using the Levenshtein distance metric [55], and sentiment analysis based on polarity scores [59].

2) COMBINING AUTHOR/DISSEMINATOR DATA AND SOCIAL IMPACT (NETWORKING) ANALYSIS (RESEARCH DOMAINS #1 AND #3)

While textual content is considered the most important aspect of FNPD detection, a notable body of research also emphasizes the importance of combining author and disseminator analysis with social context information. Some studies have compared the predictive power of user profile information with timeline tweet features [62] and social graph data [68], [78] in the context of FNPD detection. Although some research shows that user profile features can outperform social context information [68], [78], results from different model training strategies suggest that while user profile information generally offers stronger predictive capabilities than social engagement features, combining these features together yields the highest performance [62], [68], [78].

Other approaches have focused specifically on different components of social context, including social networks [98, 18], echo chambers [72], the dissemination of disinformation

TABLE 5. Summary of the meta-analysis results for selected criteria.

Estimates	Author's/disseminators' research	Content research	Social impact research
Real-time detection (%)	30%	5 %	37%
Main three datasets	1.Datasets collected by authors (36.6%) 2.FakeNewsNet (30%) 3.Cresci (16.6%)	1. Kagle fake news (27%) 2. FakeNewsNet (27%) 3. ISOT (20%)	1. Twitter-Based (44.4%) 2. Fact-Checking Websites (22.2%) 3. Specific or Unique Datasets (11.1%)
Five main novelties	1. Advanced Neural Network Integration. 2. Comprehensive Multi-source data analysis. 3. Real-time detection and processing. 4. Explainable AI and transparency. 5. Graph-based and network analysis techniques for bot detection.	1. Introduction of hybrid approaches. 2. Increase in ensemble methods 3. Adoption of contextual word embeddings. 4. The emergence of models based on heterogeneous data (user information, OSN structure, news propagation, and content). 5. Real-time FNDP detection.	1. Advances in graph-based learning of propagation patterns. 2. Content and user interaction fusion. 3. Bot detection and OSN influence. 4. Role Identification and Infiltration. 5. Measuring echo chambers' polarization and cognitive warfare research.
Temporal pattern for the period 2018-2024	2018-2019: Focus on traditional ML models and early hybrid models, emphasizing profile data and textual content analysis. 2020-2021: Advancements in DL techniques and hybrid models, with an emphasis on model generalizability and integrating network metrics and dissemination patterns. 2022-April 2024: Shift to practical applications and real-time detection systems, using advanced DL approaches with psycholinguistic metrics, multilingual methods, user credibility evaluation, and engagement patterns.	2018-2019: The proliferation of conventional ML models and vectorization techniques. Use of GNN for feature extraction. 2020-2021: Introduction of RNN and CNN. First appearance of hybrid approaches and ensemble methods. 2022: Continued interest in propagation tree features and hybrid approaches. 2023-2024: Further development and increased use of hybrid and ensemble approaches, incorporating advanced techniques such as transformer architectures and attention mechanisms.	2018: a focus on broad system development; 2019-2020: shift towards more technical and AI-driven solutions; 2021-2022: a focus on early detection techniques, and more sophisticated methods. 2023-2024: tendency towards (i) tracking narratives spread within online ecosystems, (ii) inductive learning frameworks using content and graph-based indicators, (iii) Echo Chamber Score (ECS) metrics, (iv) cognitive warfare, (v) metrics to measure polarization of echo chambers.
Main methods used	Advances DL models (hybrid approaches); Graph-based methods; User behavior and social dynamics analysis; Real time detection; Model adaptability and explainability.	Hand-crafted features, static word embedding, graph-based features, traditional ML algorithms, hybrid approaches, and ensemble approaches.	OSN analysis, graph-based ML techniques, social context & interaction analysis, bot & user role analysis, cognitive bias analysis, homophily and controversy measures, agent-based modeling, narrative analysis, tracking inauthentic behavior, filter bubble analysis, modularity maximization, longitudinal analysis multiplatform studies, strategic interaction models

[90], diffusion patterns [118], and polarization [124]. These studies have integrated user profile characteristics with social context analysis to gain deeper insights into the dissemination and social impact of FNPD. For example, [98] discusses the integration of author analysis with social network (influencer) analysis and introduces a top-k influential user algorithm based on tweets and user interactions. In [72], the SAF/A approach investigates the influence of echo chambers by collecting information about users' followers. Study [118] combines population dynamics analysis with a physics-informed neural network (PINN) to model the dissemination of disinformation, while [124] identifies a universal scaling law for opinion distributions in social systems, facilitating the analysis and prediction of different polarizing phases in social dynamics.

3) COMBINING TEXTUAL CONTENT AND SOCIAL IMPACT (NETWORKING) ANALYSIS (RESEARCH DOMAINS #2 AND #3)

Textual content has been effectively combined with various social context features, including network characteristics [52], [53], [82], [84], social context integration techniques [18], [63], [113], [127], diffusion and propagation analysis [95], [108], [122], and methods to understand community polarization and echo chambers [119], [125].

Network features were often integrated with textual content to capture user interaction information. For instance, [84] combined user interactions, social graphs, and user relationships by creating embeddings that were used alongside textual content embeddings in an ensemble approach. Moreover, [53] examined hierarchical propagation networks emerging from news dissemination, analyzing micro- and macro-level networks from structural, temporal, and linguistic perspectives. Meanwhile, [52] analyzed social interactions and textual information from retweets using GNNs and co-attention mechanisms.

From the perspective of combining social context with textual content, more complex models have been integrated. These include the use of tensor-based representations with Transformer architectures [63], [113] and the aggregation of node-level attributes in graphs while preserving structural properties [18].

FNPD diffusion analysis has been explored from various perspectives, such as trend analysis using social analytics, which considers content sequence and social contextual information [108]; feature representation of social user circles [95]; and the identification and tracking of narrative dissemination within online ecosystems [122].

The aspect of polarization was integrated with textual content using methods like Aspect-Based Sentiment Analysis (ABSA) and Group Decision-Making (GDM) consensus metrics to measure community polarization and echo chambers by analyzing social media posts that reflect community opinions [125]. Additionally, [119] introduced measures of ideological polarization, examining factors such as opinion extremity and the organization of echo chambers within social networks.

4) COMBINING AUTHOR/DISSEMINATOR DATA, TEXTUAL CONTENT AND SOCIAL IMPACT (NETWORKING) ANALYSIS (RESEARCH DOMAINS #1, #2 AND #3)

Recent advancements in FNPD detection have increasingly focused on integrating a variety of features, including textual content, user metadata, social engagement, and network dynamics, to develop comprehensive models that address the complexity and multifaceted nature of FNPD dissemination on social media.

The integration of metadata, profile, and dissemination features has been analyzed in [50], [57], [63], and [66]. For instance, [57] examined the impact of incorporating tweet metadata alongside tweet text in an LSTM model, highlighting its benefits. Studies [63], [66] emphasized the importance of user profile and dissemination features for effective classification in propagation-based fake news detection tasks. Additionally, [50] demonstrated that combining content-based features with profile, activity, and engagement information is effective for bot detection in low-resource languages.

Integrating social engagement features has proven to be highly effective in enhancing model performance. For example, a tri-relationship framework developed in [60] emphasized the importance of combining news content with social engagement elements, significantly improving model accuracy. Similarly, [70] confirmed that merging user engagement data with textual content leads to substantial gains in model accuracy, demonstrating the superiority of hybrid models over those relying on a single data type. Additionally, some models provided deeper insights into social interactions. For instance, [71] employed social context graphs to map the relational dynamics between users, news, and sources, while [53] demonstrated that leveraging both article content and social interactions allows micro-level features to outperform macro-level ones. Moreover, differences in user behavior between controversial and non-controversial topics were investigated, highlighting the importance of user reactions and polarization features in addressing social media propaganda [54].

The comprehensive analysis of FNPD presented in papers [63], [79], [80], [85] utilizes a multifaceted approach that combines insights from authors, content, and social context. These studies emphasize critical characteristics of news sources, such as political bias, credibility, and trustworthiness. By examining user engagement patterns—such as likes, shares, and comments [79]—and user profile characteristics along with activity metrics [63], a deeper understanding of user interaction is achieved, and the interplay between network dynamics and content effects is explored. Additionally, user behavior analysis, post characteristics, and network dynamics are employed to address the scarcity of labeled data, introducing the concept of weak social monitoring and underscoring the importance of network-based features like friendship and diffusion networks [80]. Furthermore, [85] proposes the integration of social context embeddings with textual content features, using advanced neural network techniques to enhance the detection and analysis of fake news.

5) KEY INSIGHTS FROM COMBINED APPROACHES

To summarize the key insights on combined approaches, the main conclusions are as follows:

a: TEXTUAL CONTENT ANALYSIS FOR BOT DETECTION

In bot detection for FNPD, analyzing textual content can significantly improve detection accuracy. Leveraging sentiment features or text similarity measures enhances the effectiveness of identifying bots.

b: COMBINATION OF DIVERSE TEXTUAL INFORMATION

Integrating various sources of textual information, such as creator profiles, articles, and subject descriptions, improves the predictive capability of FNPD models.

c: USER PROFILES AND SOCIAL CONTEXT

User profile features are useful for understanding the behavior of individual FNPD disseminators. When combined with social context elements - such as social graphs and engagement features - models can capture broader patterns of

interaction and better understand how FNPD disseminates through networks.

d: INFLUENTIAL DISSEMINATOR IDENTIFICATION

Combining user and social context analysis enables the development of algorithms to identify key FNPD disseminators and other influential components within networks.

e: ECHO CHAMBERS AND NEWS CREDIBILITY

Analyzing social context is crucial for understanding the dynamics of echo chambers and assessing the credibility of news. Combined methods benefit from this analysis to build more robust FNPD detection frameworks.

f: COMPREHENSIVE ANALYSIS OF DISSEMINATION PATTERNS

Integrating textual content sequence analysis with social context data provides a comprehensive perspective on FNPD dissemination. This highlights the importance of combining author characteristics, social network influences, and textual content information.

g: MICRO AND MACRO LEVELS ANALYSIS

Models that integrate detailed, multifaceted analyses - spanning individual behaviors (micro-level) and network-wide dynamics (macro-level) -are more effective.

h: HYBRID MODEL FOR BROADER LEARNING

Models that combine diverse data types and analytical techniques outperform those relying on a single data source or method.

B. NOVEL FOUNDATION FNPD MODEL PERSPECTIVE

We introduce the novel idea of a universal (foundational) tool trained on big FNPD data to address a wide range of FNPD problems, from narrow FNPD classification tasks to complex and multidimensional social impact modeling and assessment tasks. Foundation models, also known as large AI models, are versatile machine learning systems trained on vast datasets that can be applied to a wide range of tasks.

This novelty is not about optimizing a specific ML/DL model or method for the specific dataset and task, but about a universal set of integrated ML/DL tools, pre-trained on FNPD big data sets, capable of automatically selecting optimal research approaches in the FNPD research solution space, combining ensembles of FNPD research domains, ML/DL models, methods, and datasets.

Admittedly, there are a number of automated machine learning models and methods for the selection and/or optimization of methods that fall under the umbrella of Automated Machine Learning (AutoML), the main idea of which is to automate the end-to-end process of applying machine learning to real-world problems, including tasks such as pre-processing, feature engineering, model selection, hyperparameter tuning, and model estimation. These methods cover areas such as Bayesian optimization, reinforcement learning, evolutionary algorithms, and gradient-based methods. Many commercial and open-source AutoML systems such as Auto-Weka, Auto-Sklearn, TPOT, Google Cloud AutoML, etc. implement these techniques [23], [24].

Thus, AutoML provides a set of methods to automate the various steps of the machine learning process, with the aim of making machine learning more accessible and creating models that can match or outperform manually tuned solutions. However, as one of the main insights in our systematic review, we foresee the forthcoming next step. That is, FNPD foundation model (FM) development, which would differ from traditional AutoML by 1) ability to represent different phenomena, using combined or hybrid ML/DL, authors and disseminators network analysis techniques, social impact modeling approaches, etc. 2) integrating FNPD multidisciplinary big data (set of datasets), 3) covering and integrating multimodal solutions (text, audio and video) recognition, classification, ranking, network propagation analysis, social impact prognoses, etc.), 4) being universally applicable for a wide range of FNPD research domains, 5) providing more comprehensive and integrated solutions in the area of FNPD research, 6) fine-tuning with domain-specific or task-specific training data to enhance their performance for particular applications.

The later process allows to leverage the generalized knowledge of the foundation model while tailoring it to the specific needs. By utilizing pre-trained FNPD FM (foundation models) as a starting point, developers can apply transfer learning to create more specialized downstream applications. This approach significantly reduces the time and resources required to develop AI solutions for specific tasks.

Thus, foundation model development is critically needed in FNPD research due to several key factors (see Fig. 14):

1) EVOLVING NATURE OF THE FNPD THREAT

The landscape of propaganda and disinformation is rapidly evolving, particularly with the advent of advanced technologies like large language models (LLMs) and artificial intelligence (AI)).

2) LIMITATIONS OF CURRENT APPROACHES

existing research often relies on outdated assumptions or narrow perspectives. For instance, many studies focus primarily on social media platforms as the source of disinformation, ignoring the broader ecosystem that includes state actors, legacy media, and other influential entities. Furthermore, current approaches often lack historical context and fail to account for long-standing issues of inequality, power dynamics, and cultural differences in information consumption; developing more robust models would allow researchers to address these limitations and provide a more nuanced understanding of the problem.

3) NEED FOR INTERDISCIPLINARY INTEGRATION

FNPD research spans multiple disciplines, including political science, communication studies, psychology, and computer

science; however, there is often a lack of integration between these fields, leading to fragmented approaches and inconsistent terminology; FNPD FM development could help bridge these disciplinary gaps, creating a more cohesive framework for understanding and addressing disinformation across various contexts.

4) CHALLENGES IN MEASUREMENT AND EVALUATION

One of the significant obstacles in disinformation research is the difficulty in measuring the impact and spread of false information. Current methods often struggle to accurately quantify the effects of FNPD campaigns or the effectiveness of countermeasures. Developing more sophisticated models could improve our ability to measure these phenomena and evaluate the efficacy of interventions.

5) ADAPTING TO GLOBAL CONTEXTS

Much of the existing research on FNPD is centered on Western, particularly U.S.-centric, perspectives. However, the nature and impact of FNPD can vary significantly across different cultural, political, and social contexts. FNPD FM development is needed to create more adaptable frameworks that can account for these global variations and provide insights applicable to diverse settings.

6) ADDRESSING ETHICAL CONCERNS

As research in this field progresses, there are growing ethical concerns about privacy, data collection, and the potential misuse of research findings. Developing FNPD FM that incorporate ethical considerations from the ground up is crucial for ensuring that research in this area remains responsible and beneficial to society.

7) ENHANCING AUTOMATED DETECTION AND MITIGATION

While progress has been made in automated detection of FNPD, current models still face significant challenges. FNPD FM development could lead to more accurate and efficient automated systems for identifying and mitigating the spread of false information, which is crucial given the volume and speed at which FNPD can propagate online.

8) IMPROVING COLLABORATION BETWEEN ACADEMIA AND INDUSTRY

There is a notable gap between academic research and industry practices in addressing FNPD. Developing shared models and frameworks could facilitate better collaboration between these sectors, leading to more effective and practical solutions.

In summary, the complex and rapidly evolving nature of FNPD in the age of cognitive warfare requires a focus on the development of FNPD FM that would substantially reduce the limitations and shortcomings of the current rather narrow and specialized approaches. This would be possible by integrating a wide range of related research disciplines (including AI, network analysis, cognitive and behavioral science, and other social sciences), a multitude of applicable methods, models, measurement metrics, validation approaches, etc. Such a FNPD FM would consider the global context of this research area, which would include the large number of datasets available on FNPD. This would lead to the development of a universal, much more comprehensive, and effective framework for combating FNPD, enabling timely and effective identification of the authors, content, and dissemination features of FNPD information, as well as comprehensive social impact analysis. Such developments are essential to enhance academic knowledge and underpin policy decisions and technological solutions in the ongoing fight against FNPD.

In Fig. 14, we propose production rules as a well-known structured framework for AI knowledge representation and reasoning. This approach provides a versatile method to model and influence various FNPD creation, dissemination, and influence scenarios. Uncovering these rule sets presents a significant challenge in FNPD research, with the potential for wide-ranging effects. By identifying two production rule sets (#1 → #2 and #2 → #3) researchers could develop a generative approach for automatic FNPD detection and monitoring, enabling the creation of diverse FNPD tackling scenarios with varying social impact outcomes.

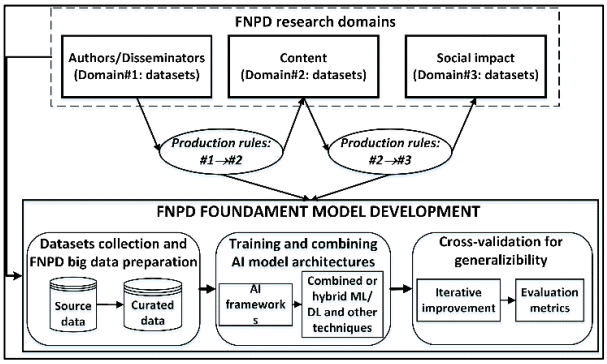


FIGURE 14. FNPD domains, production rules, and FNPD foundation model.

In essence, these production rules represent the relationships that bind the FNPD research areas together. For example, the set of production rules #1 → #2 would formalise the relationships between FNPD content creators/disseminators (#1) and the information content they generate (#2), while the set #2 → #3 would formalise the relationships between the information content generated (#2) and the social impact of that content (#3).

In this way, the formalisation of the logic of relationships between FNPD research domains, based on empirical data, would allow the combination of different ML/DL models into ensembles and hybrid architectures. Such architectures could address universal challenges ranging from the collection, preparation, and analysis of various FNPD data, to the generation of concrete results not only for real-time FNPDs identification, but also for the prediction of societal impacts and the recommendation of countermeasures.

It is likely that such an FNPD FM, by learning and using the 'human in the loop' principle, could eventually

become a proactive system that not only reacts post factum to the 'innovations' of FNPD misinformation creators, but also anticipates new developments in the FNPD field and recommends or takes proactive steps to take preventative measures.

Key components and stages of fundamental model development in FNPD research may include (see Fig. 14):

a: PROBLEM DEFINITION

Identifying specific aspects of FNPD to be modeled, such as FNPD propagation modelling, modeling, event based FNPD impact modeling, infiltration modeling, authors/disseminators modeling, segmentation modeling, bots detection, networking analysis, social impact modeling, etc.

b: BIG DATA COLLECTION

Collecting relevant FNPD data sets, often from social media platforms or news sources stemming form all related research domains.

c: FEATURE EXTRACTION

Identifying and quantifying relevant features of the data, such as linguistic patterns or network structures.

d: DESIGNING DIFFERENT RESEARCH MODELS

Developing mathematical or computational frameworks to represent different phenomena, often using combined or hybrid ML/DL, which could include, FNPD detection methods, authors/disseminators analysis, NLP analysis, context analysis, semantic analysis, sentiment analysis, botnet detection, networking analysis, fact checking (debunking), social impact assessments, echo chambers analysis, radicalization, marginalization analysis, fragmentation analysis, foreign information manipulation and interference, cognitive warfare narratives analysis, etc.

e: TRAINING AND VALIDATION

Testing models on well known constantly updated data sets to ensure accuracy and reliability.

f: REFINEMENT

Iteratively improving models based on performance and new insights.

g: CROSS-VALIDATION

Testing models in different contexts or data sets to ensure generalizability.

h: IMPLEMENTATION

Applying models to real-world scenarios for practical use in research or FNPD detection systems.

V. CONCLUSION AND DISCUSSION

A. MAIN INSIGHTS

In summary, there are a few key conclusions to be drawn:

- The PRISMA systemic review framework shaped the search process and helped to objectively select recent, well-cited papers for in-depth meta-analysis, covering the most important current research trends in the field.
- The combination of three specialized systemic reviews, rather than one broad review, provided a unique opportunity for in-depth analysis, comparison, identification of overlapping niches or complementarities of FNPD research approaches in each of the consequently related domains: starting with D#1 (analysis of authors/disseminators), then D#2 (content analysis), and finally D#3: (social impact).
- This research has shown that all three FNPD research domains have considerable overlap in terms of the data sets used. This is particularly evident in the first and second domains. However, the third research domain is much broader in terms of the social data used. However, the part of it that explores (using ML/DL) authors/disseminators overlaps considerably with the first research domain data. In summary, the systemic review found that multi-domain datasets are not only possible but also desirable to achieve more efficient, accurate and integrated research results.
- From a methodological point of view, the analysis of FNPD authors and disseminators is closely related to social impact modelling, as evidenced by the typical focus on factors such as user trustworthiness, engagement, profile analysis and interaction activities in the first and third research domains. This overlap highlights author/disseminator profiling and credibility metrics as critical to understanding and evaluating phenomena in both research domains.
- Bot detection models within the context of FNPD research stand out in their training and testing strategy. They use different datasets for training and testing, particularly to assess their adaptability to new, unseen data.
- In the domain (#2) of FNPD content analysis and classification, used methods depend on whether the training data consists only of FNPD news content or includes social context information. The analysis reveals trends toward static word embeddings for extracting features from news content and using hand-crafted features and graphical neural network models when social context information is included. Sentiment analysis is used in both categories, with more emphasis when social context information is added to the training data. Traditional vectorization techniques and contextual word embeddings are used when the training data consists only of news content.
- FNPD classification models have recently moved from traditional algorithms to more advanced neural networks, including RNNs and CNNs. However, conventional ML algorithms are still used with more advanced methods, such as ensemble approaches. Leveraging the collective strengths of multiple algorithms through

combined or hybrid approaches is becoming increasingly common and highly effective.
- The detection of logical fallacies is a popular standalone task, but its application within FNPD systems is a relatively new and emerging idea. Among the reviewed papers, only a few, all published within the last year, have addressed logical fallacies as part of FNPD frameworks. These recent studies highlight the potential of integrating fallacy detection into propaganda detection, providing models that go beyond simple text classification.
- In the field of social impact research, there are critical areas of analysis (niches) that have not yet received sufficient research attention, such as (i) reasons for successful deception, (ii) radicalization and polarization, (iii) social impact via social behavioral patterns analysis, (iv) social impact modelling, (v) cognitive warfare modelling, (vi) FNPD influence operations, (vii) echo chambers research, (viii) opinion dynamics modelling.
- The fight against FNPD on social media platforms is being tackled on multiple fronts, focusing on early detection techniques and broad system development (e.g., FakeNewsTracker) and external influences (e.g., Russian trolls), shifting towards more technical and AI-driven solutions (e.g., geometric deep learning, social bot detection, and network-based patterns), and employing deep sociological insights.
- To operationalize how and which author actions disseminate FNPD content and which content elements influence social behavior in OSNs, we propose to use the production rules approach (linking antecedents (IF) with consequents (THEN)) to provide a structured framework for representing knowledge and reasoning in FNPD research field. One of the advantages of this approach is that FNPD can then be modelled as a generative FNPD process that can respond flexibly to the rapidly evolving nature of FNPD by expanding the set of production rules, rather than creating fragmented one-time solutions.
- After analyzing the benefits of integrating FNPD research domains and the global perspectives for building fundamental models in different research domains, we came to the insight that the broad field of FNPD research is also maturing towards the realization of its own fundamental model. The FNPD Fundamental Model (FNPD FM) for training could use big data (datasets from all FNPD domains) and a variety of ML/DL models applied in the FNPD domains. By applying production rules, FNPD FM could be trained to integrate ensembled, combined or hybrid solutions from different FNPD domains. FNPD FM could then be applied to specific applications where solutions can be drawn from the fundamental model knowledge base to provide possible solutions at different scales using ML/DL model ensembles or hybrid solutions for author, content and social impact analysis, etc.

B. LIMITATIONS

Despite the many existing studies on FNPD detection, the field is still evolving, and new methods or evidence are needed to advance the state of the art. The main limitations of current methods were identified and summarized based on the limitations identified in the reviewed studies. Although not all studies clearly identify the limitations of the methods used, most papers discuss directions for future research that could improve the results. Consequently, we have formalized general limitations from these studies, focusing specifically on the application of FNPD author/disseminator, content, and social impact analysis. Hence, the main limitations of the research approaches are assessed and summarized below.

1) DATA LIMITATIONS

Most of the papers reviewed obtained their data from pre-labeled databases. The preference for pre-labeled databases is often due to their accessibility, saving researchers the time and effort of collecting and labeling data from scratch. However, this approach may introduce bias and limit the generalizability of the results.

A key limitation is the regional or activity-specific focus of studies, such as those focusing only on Russian trolls, which may lack comprehensive insights into different misinformation campaigns [113]. In addition, the scope of the datasets is limited, with some focusing narrowly on U.S. politics [127] or exclusively on English-language tweets [102].

Nonrepresentative samples of the populations are limiting the validity of the results [119]. The demographics of online evaluation may also lack diversity, further limiting the applicability of the research.

Reliance on fact-checking websites as a primary data source is also common. However, this can also lead to bias, and reliance on these sites requires time-consuming expert analysis.

The rapid evolution of content on platforms such as Twitter also affects the applicability of models. In addition, many models rely heavily on the availability and representativeness of user characteristics and labeled data. Inadequacy, obsolescence, or lack of diversity in this data significantly reduces the effectiveness of these models. As pointed out in the paper [84], many datasets, such as FakeNewsNet, rely on tweet IDs, most of which have been removed from Twitter, making the data no longer available on the original platforms. This problem limits the ability to collect comprehensive social context data.

Effective detection often requires a combination of content and metadata analysis. However, models that rely solely on one aspect tend to be less accurate, underscoring the need for comprehensive data integration. In addition, data collection methods, such as API rate limits, can affect the depth and breadth of data analysis, which in turn affects the learning and detection capabilities of models [78].

It is worth noting that all of the models analyzed were evaluated using binary data sets. This limits their application in more complex scenarios where news could fall into multiple categories beyond real or fake.

2) MODEL LIMITATIONS

The use of artificial neural networks in FNPD detection is considered promising. However, there is an implication that their potential in this context has not yet been fully realized or explored. The detection of FNPD in social media remains an unsolved problem due to several limitations. One of the main challenges is the rapid spread of information, which requires real-time detection. Many existing models may not work in real-time, further complicating the detection process. Models may not detect fake news until it has begun to spread, further emphasizing the need for real-time detection [52].

Furthermore, the dynamic nature of social media platforms and user behavior complicates the detection process, making it essential for models to continuously adapt to these changes [53], [54]. However, the dynamic nature of propagation introduces some limitations that static models cannot capture. In addition, graph-based methods face challenges because it is difficult to propagate information to neighboring nodes per interaction in real-time [82].

The intention to consider more complex or hybrid neural network architectures [59], [77] for data analysis indicates the need for more sophisticated models to handle the complexity of the FNPD detection process. On the other hand, such advanced methods require significant computational resources and may not be suitable for all environments [63]. It thus follows that there is a need for methods that emphasize the optimization of model hyperparameters [97].

Graph Neural Networks (GNNs) have been investigated for propaganda detection. However, applying current GNN-based methods to propaganda detection poses a significant challenge due to the various propaganda techniques such as repetition, logical fallacies, and doubt [90]. It is important to note that only a limited number of studies have addressed these challenges. In addition, while experimental results indicate that the proposed methods effectively detect propaganda [91], the training speed of the models is relatively slow.

In some studies, simulations only modified one parameter [120], so results are based on the Ceteris Paribus criterion. Some studies were not able to consider specific factors used in social networking site algorithms due to inaccessibility [116]. Most studies were limited to users on one SNS platform.

Given the limitations, it is important to note that the models may be vulnerable to adversarial attacks and manipulation, especially those that mimic human-like behavior or use sophisticated writing styles. There are concerns about the models' resilience to adversarial attacks [63]. In addition, the nuances and varying definitions of ''fake news'' are not universally agreed upon and may affect the robustness of the model. The linguistic and geographic independence of the models requires further investigation. In addition, using non-standard language and slang in social media posts can affect the accuracy of models based on text analysis [59].

3) GENERALIZABILITY AND SCOPE LIMITATIONS

A major limitation of current FNPD detection models is their lack of generalizability. These models perform well on specific datasets, but do not effectively generalize across datasets or contexts. Models struggle to generalize across different types of fake news, social media platforms, and the evolving tactics of FNPD purveyors. This includes the challenge of detecting novel bots or fake news strategies that were not part of the training dataset [61], [76].

Some measures assume that people organize themselves in a 1D opinion space with only two poles, which may not accurately describe all social environments [11]. Another multidisciplinary study [12] found an analytical solution under the adiabatic approximation, etc.

In addition, some models focus on specific features and omit potentially informative attributes. For example, focusing on propagation tree and stance network features or focusing on text-based communication. The issue of generalizability is partly due to their focus on domain-specific relationships, which may hinder cross-domain performance. In addition, the specific geographic focus may not be representative of global trends. There is a notable gap in research coverage, such as the lack of exploration of geolocation methods in online social networks, which could provide significant insights.

4) OTHER LIMITATIONS

When it comes to FNPD detection, interpretability and transparency are crucial. However, understanding the decision-making process of complex models can be challenging, raising concerns about interpretability and accountability [63]. Furthermore, distinguishing between sophisticated bots and human behavior can be challenging, especially when bots are designed to mimic human behavior closely [51]. In addition, rapid changes in user behavior can make traditional detection methods less effective [109], and the tendency of users to hide or delete their interaction records poses a challenge for data acquisition [107]. Furthermore, some studies may not delve deeply into certain aspects, leaving them for platforms themselves to counter [108].

The mentioned limitations indicate both the complexity of the problem and the research areas where the current research could be extended. Future research directions derived from the limitations are given in the following subsection.

C. INSIGHTS FOR FURTHER RESEARCH DIRECTIONS

The limitations of the FNPD research field are also challenges that can open up new opportunities and research niches, see Fig. 15. In this figure, we have linked the main classes of limitations found to directions for further research that may extend the boundaries of these constraints. Our analysis has highlighted within each class of constraints the most relevant and, in our view, promising avenues for further research (see bolded research areas in blue).

The systemic review showed that there is a clear need for advanced combined research approaches based on user profiles, textual (and multimodal including voice and video) content and social impact studies in a more integrated and coherent way. Multimodal research methods, which include audio and visual analysis alongside text, is a very promising direction for this research that we plan to explore in the future. However, in this systematic review, we have limited ourselves to an overview of FNPD authors, textual content, and social impact analysis domains.

In what follows, we have combined and summarized the most promising opportunities and niches for future research in the three areas of research that were examined.

In order to highlight the perspectives of FNPD R&D approaches, we have constructed the world-renowned Gartner Hype Cycle Curve (GHCC) based on the results of our systematic analysis, see Fig. 16. However, it should be borne in mind that we have focused on the FNPD research impact and R&D in this field than on the maturity of applied solutions. We have mapped the key FNPD R&D directions on the GHCC based on their current status and development potential. As a reminder, the GHCC is a graphical representation developed by Gartner, an IT research and advisory firm, to illustrate the maturity, adoption, and social application of specific technologies. It depicts five key phases that a technology typically goes through: (i) technology trigger, (ii) peak of inflated expectations, (iii) trough of disillusionment, (iv) slope of enlightenment, (v) plateau of productivity. By visualizing the journey of a FNPD technology from its inception to mainstream adoption, the GHCC serves as a valuable tool for navigating the complex landscape of technological innovation and its impact on society [136].

The presented GHCC reflects the current state of R&D in the FNPD field. The technologies at the ''Technology Trigger'' stage are emerging and show promise but are not yet fully developed. Those at the ''Peak of Inflated Expectations'' are generating excitement but may not have proven their full potential. The ''Trough of Disillusionment'' represents areas where initial excitement has waned, and challenges have become apparent. The ''Slope of Enlightenment'' includes technologies that are beginning to show practical applications, while those at the ''Plateau of Productivity'' are more established and show consistent results.

The placement of these FNPD research directions on the Hype Cycle is subjective and may change over time as the FNPD field continues to evolve. The time axes indicate the life cycle of FNPD R&D technologies. That is, all new technologies start from the left and move over time through the curve to the right.

Below are underlined the main perspectives for further research on FNPD.

1) EARLY FNPD DETECTION FOR MODELS BASED ON AUTHORS/DISSEMINATORS DATA

Systemic review shows that early detection using author/disseminator data is feasible, with some models achieving significant accuracy within a few hours of news circulation. However, this research area deserves further investigation. Investigating whether specific disseminators are present in the early or late stages of propagation could reveal patterns characteristic of bot profiles, especially since botnets often aim to disseminate non-credible news quickly.

2) DETECTION OF MASSIVE COORDINATED FNPD INFLUENCE OPERATIONS

There is a clear need for in-depth research into the internal and external collaboration patterns of authors, botnets, and troll communities, as they work in a coordinated way in massive influence operations (e.g., on elections).

3) FLEXIBILITY AND ADAPTABILITY

To ensure that models generalize effectively to new, unseen data, it is important to train and test on multiple data sets. This practice is common in bot detection, where rapid changes persist. However, it is rare for other FNPD detection models. Therefore, implementing this approach is essential to maintain the adaptability needed to keep pace with the dynamic nature of social media and the evolving strategies used to spread FNPD. Researchers should put more effort into improving the interpretability of models, generalizing their applications across different datasets or platforms, and exploring their adaptability and resilience.

4) COMBINED ANALYSIS OF FNPD MULTIMODAL (TEXT, AUDIO, AND VIDEO) DATA

Multimodal analysis is still in its infancy, but its importance is growing as more FNPD information becomes available on social media channels in text format and audio, and video formats. This presents new challenges and opportunities for researchers.

5) USE OF PROPAGANDA TECHNIQUES

The review revealed that very few studies address the specific propaganda techniques used in texts and their automatic detection capabilities. A few studies also use original expert or hybrid annotation techniques for propaganda and disinformation labeling based on pre-trained LLM (Large Language Models) annotations.

6) ADAPTATION TO OTHER LANGUAGES

FNPD studies focus on English language datasets. This is mainly driven by the fact that already existing datasets and, most of all, feature engineering techniques are built for the English language. There needs to be more research where models built for the English language would also be tested for other foreign languages.

7) ADAPTATION TO OTHER SOCIAL NETWORKS

The majority of all FNPD models are built on Twitter datasets, mainly because Twitter allowed its data to be freely crawled and used for research, but in general, Twitter does not even appear in the top ten largest online social

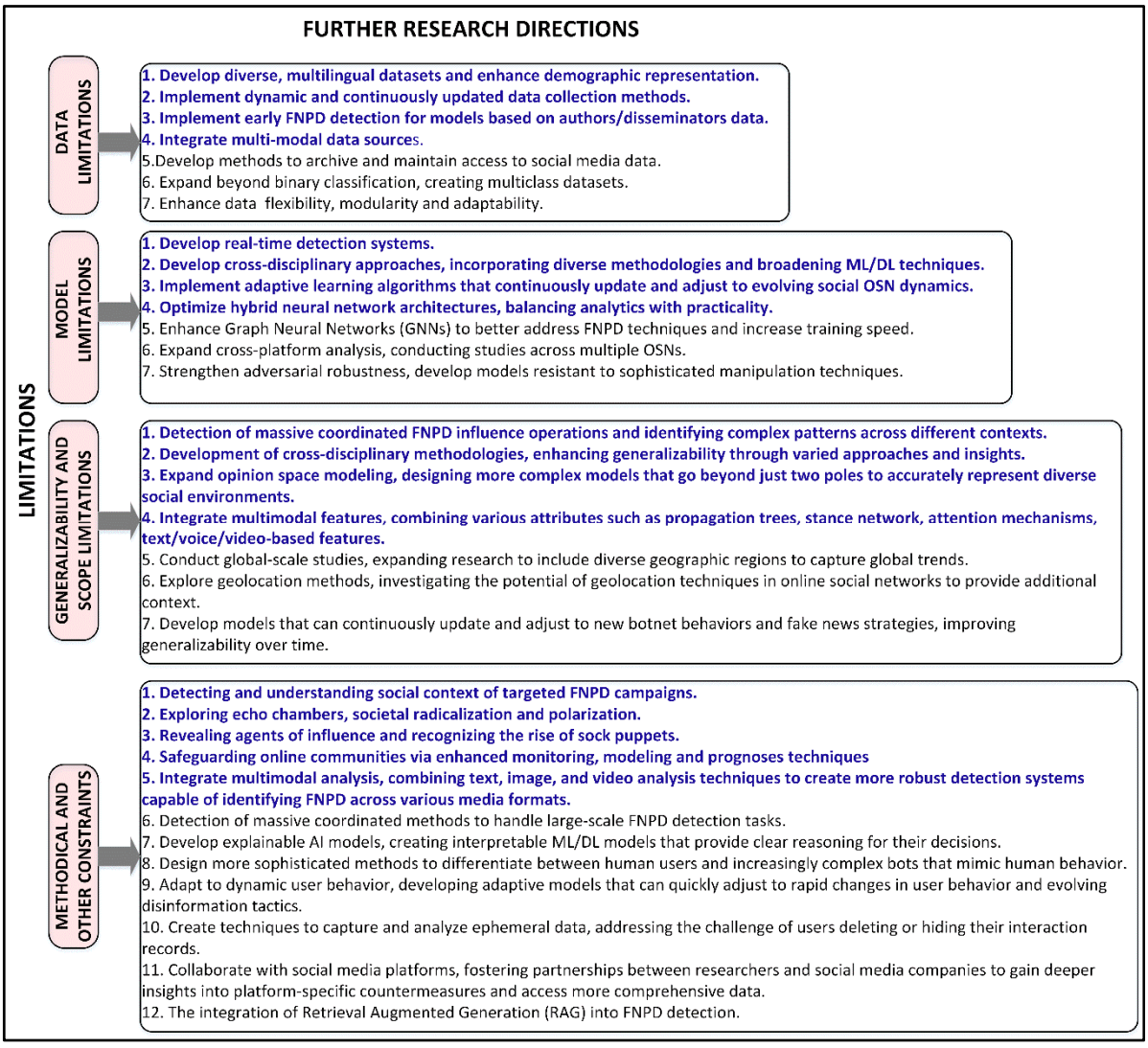


FIGURE 15. Linking the main limitations to the identified future R&D directions in the FNPD research field.

networks (https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/). For robustness evaluation, the models should be tested on other online social network datasets to see their adaptability.

8) THE DEVELOPMENT OF GENERALIZED MODELS

The literature review has highlighted a common goal for the future - the development of generalized models that can be adapted to diverse types of data and real-world scenarios. In addition to this goal, there are many research areas where there is a need to extend current research, such as exploring other algorithms and additional features, as well as the need for datasets with more complex scenarios.

9) QUALITY TRAINING DATA COLLECTION

Building accurate and effective news classification models using machine learning techniques depends on the availability and quality of training data. The training data is the most critical component in the training process. Future work could include extending an existing dataset to multi-class real-world fake news datasets or collecting new datasets with the aforementioned scenarios.

10) EXPLORING MORE COMPLEX NEURAL NETWORK ARCHITECTURES OR HYBRID APPROACHES

There is also room for further optimization to improve model accuracy or other performance metrics. This exploration can include the integration of transformer models with convolutional or recurrent neural networks to take advantage of these architectures. In addition, attention mechanisms can be used to focus on relevant parts of the input data more effectively. Different training techniques, such as transfer learning or adversarial training, can be experimented with.

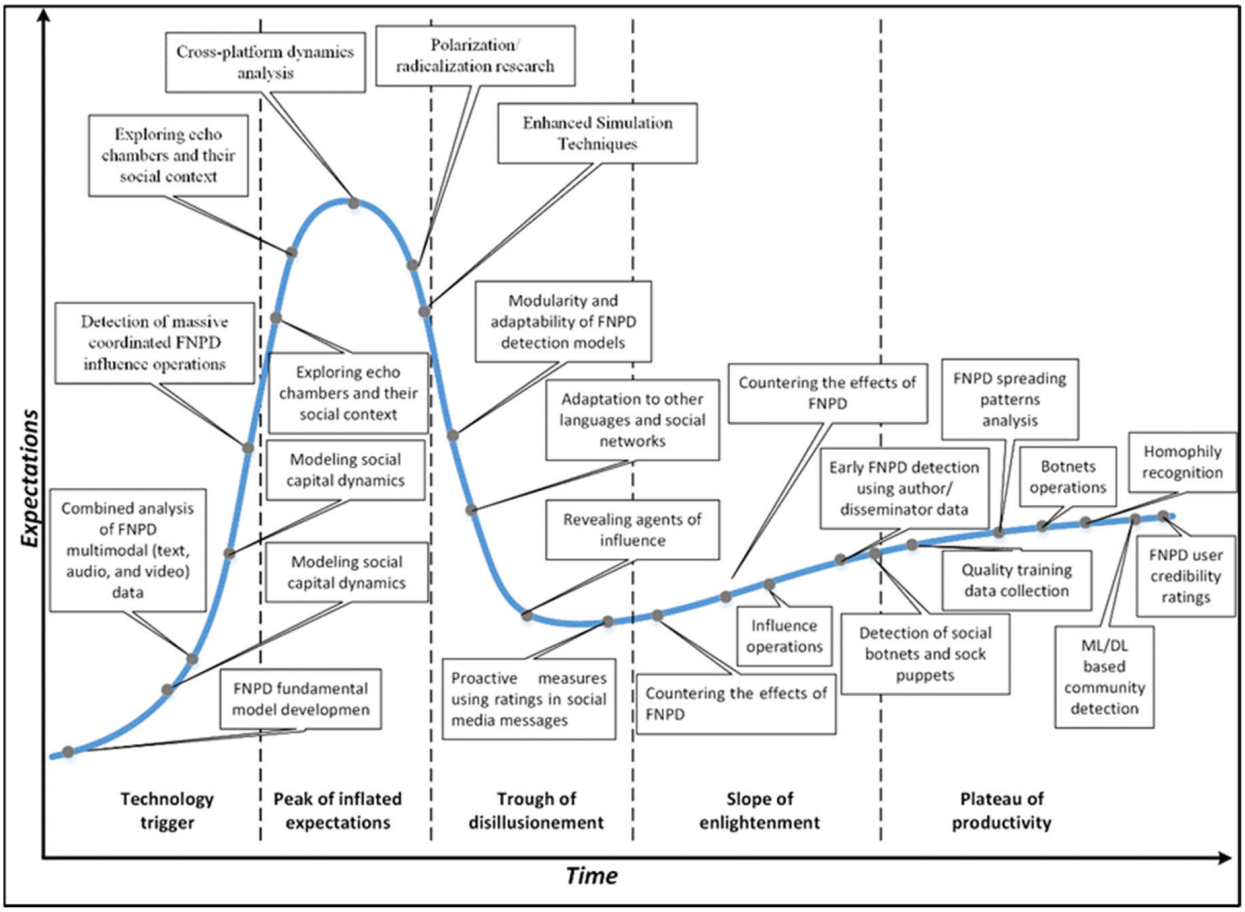


FIGURE 16. Gartner hype cycle curve constructed for the visualization of main FNPD R&D perspectives.

11) EXPLORING ENHANCED TEXT FEATURES

More text features can be explored in the future to improve the performance of the FNPD models. This could include integrating advanced linguistic analysis techniques to better understand the nuances of text. The focus could be on detecting stylistic elements such as narrative inconsistencies. Additionally, the inclusion of contextual embeddings such as those derived from transformer-based models, like BERT or GPT, could also be considered.

12) THE INTEGRATION OF RETRIEVAL AUGMENTED GENERATION (RAG) INTO FNPD DETECTION

RAG combines retrieval mechanisms with generative models to produce fact-based and context-aware output, providing a tool for improving fact-checking. While this approach is increasingly used in other NLP domains, its application in FNPD detection is still limited.

13) OVERCOMING THE CHALLENGES OF GENERATIVE AI

One challenge to FNPD detection is AI hallucination, where models generate factually incorrect content. While some approaches, such as RAG, mitigate this risk by grounding outputs in retrieved evidence, more work is needed to develop robust validation methods.

14) DETECTING AND UNDERSTANDING SOCIAL CONTEXT OF TARGETED FNPD CAMPAIGNS

It is important to estimate the social context and dynamics of FNPD information spread (target audiences, influencers, botnets, multimedia channels, etc.) and develop detection and mitigation techniques against such campaigns. There is a need for continuous early detection of organized FNPD campaigns to address the growing scale of disguised foreign influence operations.*Exploring echo chambers*. There is a lack of effective measures for exploring the social context of echo chambers, which play a significant role in the spread and acceptance of fake news. Multidisciplinary (including social sciences, psychology, cognitive science, behavioral science, etc.) research is needed to shed light on the root causes of the formation of closed echo chamber clusters. To gain deeper insights further research could focus on such aspects: experimenting with metrics for online polarization detection, developing countermeasures for echo chamber prevention or mitigation, exploring consensus-based measures embedded in community detection approaches, investigating

the contribution of different components to overall network polarization.

15) EXPLORING SOCIETAL RADICALIZATION AND POLARIZATION

There is a lack of research assessing the impact of the FNPD flows on the radicalization and polarization of society. There is an urgent need for metrics that find a causal or correlative relationship between radicalization and polarization and the long-term impact of FNPD flows on different demographic and psychographic groups.

16) DETECTION OF SOCIAL BOTNETS

There is a lack of effective tools to investigate in a timely manner the social impact of botnets, where they act in a synchronized and coordinated manner, stirring the emotions of the public and drowning out the voices of rational opponents in the avalanche of news on social media.

17) RECOGNIZING THE RISE OF SOCK PUPPETS

There is an increase in the use of sock puppets (fake online personas created to deceive others and manipulate information), suggesting research into advanced techniques to detect them and understand their motivations for deceptive manipulation, amplifying false narratives, undermining trust in online communities and information sources, exacerbating societal divisions by presenting extreme positions and stirring controversy, targeting vulnerable audiences (propaganda campaigns often target vulnerable or susceptible individuals with tailored messages), influencing electoral processes by spreading disinformation, supporting particular candidates, or sowing confusion and mistrust in the electoral system.

18) REVEALING AGENTS OF INFLUENCE

Governments often use their intelligence agencies or other state apparatus to influence public opinion both domestically and internationally. This can be done through propaganda, disinformation campaigns or control of media narratives. It is often done through covert actions by agents of influence. These are individuals or groups skilled in digital technologies who engage in cyber activities to influence opinions. This can include hacking, the release of sensitive information, manipulation of social media algorithms, and creation of fake news. There is therefore a need for innovative approaches and tools to detect covert influencers, who are orchestrating covert operations to spread FNPD, influence media narratives, or disrupt social cohesion.

19) SAFEGUARDING ONLINE COMMUNITIES

More broadly, online communities that uphold traditional and fundamental values are being virtually attacked, infiltrated, and disrupted from the outside and from the inside, using tailor-made impact strategies and operations. As part of information warfare, democratic online communities need new methods of self-defense, using ML/DL techniques that not only detect such operations in time, but also unmask the sources and prevent their impact.

20) ENHANCED MODELING AND SIMULATION TECHNIQUES

This narrative focuses on improving the accuracy and complexity of models used to study FNPD systems. Key aspects include (i) advance modeling of the social impact of FNPD influence operations, (ii) refining models for increased accuracy and inclusion of additional user dynamics, (iii) incorporating more parameters to capture a nuanced understanding of FNPD regional strategies, (iv) extending simulation models to capture various FNPD spreading and impact scenario variations.

21) CROSS-PLATFORM DYNAMICS

As disinformation often spreads across multiple platforms, researchers are developing models to track and analyze cross-platform information flows. This includes examining how content mutates or adapts as it moves between different social media environments.

22) PROACTIVE MEASURES USING RATINGS IN SOCIAL MEDIA MESSAGES

Further research might suggest a whole range of models and applications that could generate credibility ratings by assessing OSN news from different perspectives. These credibility ratings could be given, according to a defined algorithm, to that social media news and posts whose sources are doubtful and whose popularity or spreading rates exceed certain critical thresholds. By clicking on the ratings icon next to a post, people can see in a convenient and simple form the aggregated multimodal ratings of disinformation and propaganda, fact-checking estimates (e.g., using perplexity.ai via API), and expert feedback (if any). Expert judgments are typically delayed by 24 hours, but machine learning-based rankings would help in the initial stages of disinformation detection to reduce re-sharing and reposting in the next cascades of disinformation propagation. This would be done automatically in real-time, making it easier for users to assess the trustworthiness of posts, news, and messages. Such AI-based tools could give end-users an effective informative tool, impede propaganda and disinformation dissemination in the initial stages, and provide a proactive approach.

23) COUNTERING FNPD IMPACT

The researchers note that future work needs to focus on countering propaganda, suggesting that current approaches focus primarily on detection rather than mitigation or prevention of FNPD effects such as societal polarization. Another particularly important insight for further research comes from the fact that radicalization/polarization and social impact analysis via social behavior patterns are the least researched in the field. This is an astonishing revelation, as these two criteria should be the most important in FNPD social impact research. We imply that the ML/DL research community has walked around them, as it requires modeling of complex

social behavior patterns, multi-aspect clustering analysis, and complicated social impact metrics. Although such additional research composition is not yet well established in the ML/DL field, it is likely that in a few years prospective research efforts will challenge this new frontier.

24) FNPD FOUNDATION MODEL (FNPD FM) DEVELOPMENT

This systematic review highlights a novel key area of research – FNPD fundamental model development. Training fundamental models requires big data analysis, which involves different ML/DL models, combinations of them, hybrid models, and a large number of datasets. When trained on a large number of FNPD models and big data, such a fundamental model acquires the qualities of universality of application and generativity, which, for example in the case of LLM fundamental models, has led to one of the biggest explosions of AI applicability in almost all areas of human activity. In the case of FNPD FM, the prospective result could be a versatile tool with its own knowledge base, which can then be applied to narrower FNPD application tasks, with or without additional specific data cases provided for author/bot detection, content analysis, network analysis, social impact analysis, etc. In other words, the FNPD FM is trained on big data (all FNPD domains and all possible ML/DL models) for the subsequent applications to different tasks, using its knowledge base to provide possible solutions at different scales using ML/DL model ensembles or hybrid solutions.

APPENDIX A
INTERNATIONAL PROGRAMMES AND PROJECTS TO BRING TOGETHER FNPD RESEARCH DOMAINS

European Digital Media Observatory (EDMO) Launched in June 2020: EDMO supports the creation of a cross-border, multidisciplinary community of independent fact-checkers and academic researchers, integrating various research domains to detect, analyze, and expose potential disinformation threats.

Action Plan on Public Governance for Combatting Mis- and Disinformation: The OECD's Action Plan identifies three key areas for tackling disinformation, including implementing government policies to build resilient societies, increasing transparency and data sharing, and enhancing integrated research across multiple domains.

Horizon 2020 and Horizon Europe Research Programs. These EU-funded programs, with significant U.S. participation, have mobilized resources to address information veracity in social media and media. Projects like SOMA, FERMI, PROVENANCE, and SocialTruth demonstrate the integration of various research domains to combat disinformation.

Framework to Counter Foreign State Information Manipulation (US). Announced in January 2024, this framework aims to develop a common understanding of the FNPD threat and establish coordinated responses. It focuses on five key action areas, including national strategies, governance structures, human and technical capacity, civil society engagement, and multilateral cooperation.

UNESCO Consultations. UNESCO has undertaken comprehensive consultations across 134 countries on tackling mis- and disinformation. This global effort highlights the recognition of the need for integrated research.

European External Action Service (EEAS) Initiatives: The EEAS has developed a comprehensive framework to address FIMI (Foreign Information Manipulation and Interference) threats, which includes the analysis of 750 incidents of foreign manipulation between December 2022 and November 2023. Their reports emphasize the need for collaboration among stakeholders, including national governments, civil society, and independent media, to enhance resilience against FIMI. The framework focuses on five key action areas: national strategies, governance structures, human and technical capacity, civil society engagement, and multilateral cooperation.

Foreign Malign Influence Center (FMIC): Established by the U.S. government under the Office of the Director of National Intelligence (ODNI), the FMIC coordinates federal responses to FIMI. It utilizes a notification framework to ensure consistent communication across government sectors and with the public. The FMIC's strategy includes multiple goals aimed at mitigating FIMI impacts through integrated research efforts that encompass intelligence analysis, cybersecurity measures, and public awareness campaigns.

International IDEA's Project on Electoral FIMI: This initiative aims to develop actionable strategies to counter the adverse effects of FIMI on elections in various countries. A recent workshop brought together experts from across Europe to design a global methodology for analyzing electoral FIMI. This project highlights the importance of integrating insights from political science, communication studies, and civil society engagement to bolster democratic processes.

ATHENA Project: Funded by the European Union, this project aims to analyze instances of FIMI through a multidisciplinary lens. By examining tactics used by foreign actors, ATHENA seeks to develop tools for better detection and response strategies.

APPENDIX B
SELECTION PROCESS OF AUTHORS/DISSEMINATORS PAPERS (DOMAIN #1)

The search focused on studies using author or disseminator data and the detection of bots disseminating FNPD on social media.

Title words: "fake news," "propaganda," "disinformation," "deep learning," and "machine learning."

Keywords selected for this review included "bot," "spreader," "disseminator", "author," "creator," "user," "account," and "malicious."

Searches were conducted consistently across Scopus, Google Scholar, and CrossRef using title words and keywords

with 'OR.' Due to Semantic Scholar's limitations, two separate keyword searches were performed.

The article selection began with 1,099 articles, see Fig. 17. Initial exclusions of 339 records were due to duplication, irrelevant types, or incomplete fields. The remaining 760 articles were screened by title, eliminating non-relevant ones. Further filtering removed 42 articles not aligned with surveys or reviews. A keyword search excluded 434 unrelated records, leaving 252 for abstract review. After excluding papers lacking ML/DL analysis or relevant impact, 30 records remained for full-text meta-analysis.

Some observations from the first research domain (FNPD authors and disseminators analysis):

(i) The Google Scholar database provided the most extensive coverage, with the highest number of papers (469) and the highest average citation rate of 63.4 citations per paper.

(ii) The high citation rates of the selected articles indicate their relevance within the research field, with an overall average citation rate of 39.86 citations per paper across all databases. Notably, the 30 papers selected for the meta-analysis demonstrated a significantly higher average citation rate of 173.7 citations per paper, emphasizing their significance in the field.

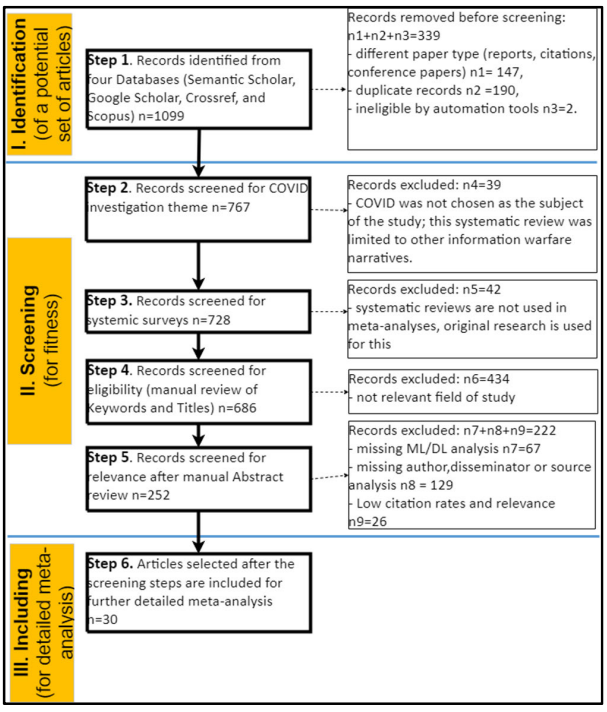


FIGURE 17. PRISMA flowchart for selecting the most relevant articles for detailed meta-analysis in Domain#1.

APPENDIX C
SELECTION PROCESS OF CONTENT ANALYSIS PAPERS (DOMAIN #2)

The selection process involved a systematic search for textual content analysis papers in Domain #2. The list of keywords and subject headings used is follows:

Title words: fake news OR propaganda OR disinformation AND detection OR identification OR classification

Keywords: deep learning OR machine learning OR neural networks

The number of records found before preprocessing was n = 1568. Before the screening, 546 records were removed (see Fig. 18). The main exclusion criteria were paper type (reports, citations, and conference papers were not considered) (n1 = 67), records with an empty field (n2 = 57), and duplicate records (n3 = 419). In addition, the automation tools identified some of these records as ineligible (n4 = 3).

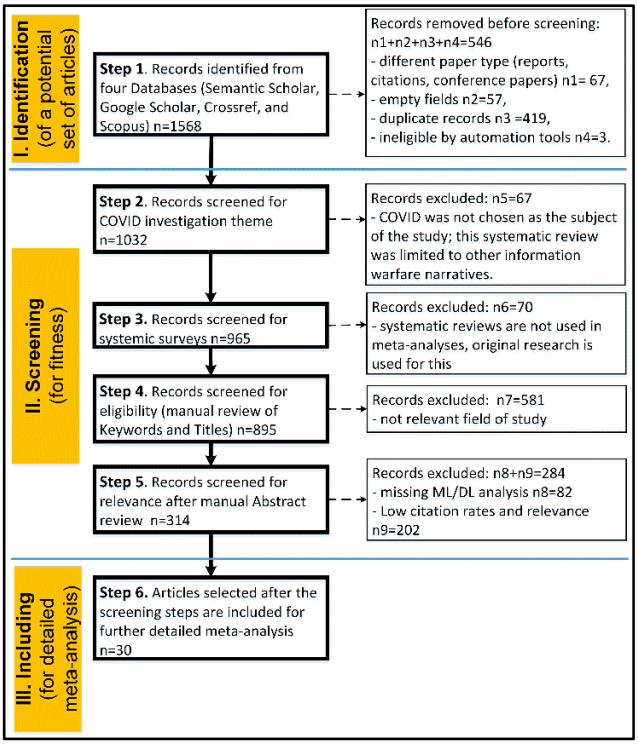


FIGURE 18. PRISMA flowchart for selecting the most relevant articles for detailed meta-analysis in Domain#2.

The eligibility screening process began with the exclusion of studies of coronavirus disease (n5 = 67) and systematic or narrative reviews (n6 = 70). The remaining records were manually screened for keywords and titles, resulting in further exclusions (n7 = 581). Abstract reviews followed, resulting in the exclusion of additional records: 82 papers that did not use ML/DL methods for FNPD (n8 = 82), and 202 papers due to low citation rates or relevance (n9 = 202). After screening, 30 papers were selected for meta-analysis.

Some observations on the 2nd review block (FNPD content analysis):

(i) The analysis of literature sources revealed several articles that were not related to the search criteria. After further analysis, we concluded that the irrelevant results were caused by some of the data sources performing full-text searches.

TABLE 6. Summary of the meta-analysis results for selected key criteria.

Key criteria	Articles																													Total per articles
	52	91	89	92	98	93	99	97	17	51	62	94	95	96	100	101	90	102	121	116	117	122	118	119	123	124	126	120	125	127	
Novelty																															
Graph-based Learning and Propagation Patterns	1						1					1																			3
Content and User Interaction Fusion			1					1					1																		3
Bot Detection and Influence				1										1																	2
Role Identification and Infiltration															1	1															2
Holistic Approaches and Comprehensive Data									1	1																					2
Advanced Analytical Frameworks		1															1	1													3
Disinformation Through Network Effects																				1	1		1	1				1			5
Detecting Disinformation Campaigns																			1			1			1						3
Measuring Echo Chambers' Polarization																										1			1		2
Cognitive Warfare																											1			1	2
Main Methods Used																															
Network & Graph-based Techniques	1	1					1	1			1																				5
Social Context & Interaction Analysis			1		1								1				1														4
Bot & User Role Analysis		1												1	1	1															4
Community detection and dynamics modelling																			1	1		1	1	1		1		1	1		8
Metrics employed																															0
Popular Metrics Used	1	1	1		1	1	1	1				1	1	1			1														11
Network analysis & modeling																				1	1		1			1	1	1	1		7
Auxiliary/Additional Metrics & Methods		1		1				1		1				1			1														6
Articles with Ambiguous or Not Explicitly Mentioned Metrics			1	1			1		1	1	1					1		1													9
Main results obtained																															
Fake News Detection Efficiency	1		1	1		1		1			1	1	1		1																9
Bot Detection and Influence		1							1					1		1		1													5
Insights on Content and Dissemination		1															1														2
Model Architecture and Methods					1		1			1			1	1																	5
Echo Chambers and Polarization Research																				1	1		1				1	1			5
Algorithmic Mechanisms and Countermeasures																				1	1	1	1		1					1	6
Study of the social impact																															
Fake News Impact on Political Events and Democracy	1	1	1	1				1			1	1					1														8
Social Bots and Their Influence					1									1				1													3
Real-World Consequences of Misinformation									1	1			1																		3
Infiltration and Manipulation by Digital Agents															1	1															2
Impact frameworks and countermeasures																					1		1			1	1	1		1	6
Total per criteria:	5	8	6	4	5	2	5	6	4	5	4	4	6	7	5	5	5	5	2	5	5	3	6	2	2	4	4	5	3	3	

(ii) The high number of duplicate records (n3 = 419) indicates that the searches were precise and that the same articles were retrieved from multiple databases.

(iii) For the full-text meta-analysis, 30 papers were selected. These papers were cited an average of 107.1 times.

APPENDIX D
SELECTION OF SOCIAL IMPACT ANALYSIS PAPERS (DOMAIN #3)

We conducted multiple searches in four scholarly databases - Semantic Scholar, Google Scholar, Crossref and Scopus. The initial total number of records found before preprocessing was n = 1388. For each scholarly database, we used a similar list of keywords and subject headings, such as

Title words: social media OR social networks.

Keywords: deep learning OR machine learning OR neural networks OR deep neural networks AND propaganda AND disinformation AND fake news.

After storing the initial dataset of records in a spreadsheet, we initiated the selection process, as shown in the flowchart in Fig. 19.

The flowchart shows the main selection phases: identification, screening and inclusion, where six exclusion steps were applied: Step#1 (different paper type, empty fields, duplicate records, ineligible by automation tools, Step#2 (COVID topic was excluded as systematic review was limited to other information warfare narratives), Step#3 (other systematic reviews), Step#4 (based on manual review of keywords and titles, not relevant field of study), Step#5 (based on manual Abstract review, missing ML/DL or social impact analysis), Step#6 (final list of 30 best cited recent articles), see Fig. 19. In this way we reached the final list of articles for full-text analysis in the meta-analysis phase.

Thus, following the approach presented, the flow chart consists of six stages. In the final stage, we sorted the papers according to two equally important criteria - number of citations and quality of the results obtained using the latest methodological advances. This ensured that the most cited older papers did not overshadow the most recent, advanced and relevant papers that had not yet been adequately cited.

Some remarks and observations regarding selection process:

- During the searching and selection process, we observed that modelling of the societal impact of propaganda and disinformation focuses mostly on traditional quantitative or qualitative social research methods. Meanwhile, ML and DL methods are just emerging in this research area.

TABLE 7. Summary of the meta-analysis results for selected specific criteria.

Specific criteria	Articles																												Total:	%
	52	91	89	92	98	93	99	97	17	51	62	94	95	96	100	101	90	102	121	116	117	122	118	119	123	124	126	120	125	127		
Community of users	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	1	1	0	1	0	0	0	1	1	1	0	1	21	70
Fake information bubbles: formation and dynamics	1	1	1	0	0	0	1	1	1	1	0	0	1	0	1	1	0	1	1	1	0	1	0	1	1	1	0	1	1		20	67
Users' profile analysis	1	1	1	1		1	1	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	16	53
FNPD spreading analysis	1	0	0	0	0	0	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	0	19	63
Social metrics of FNPD impact	1	1	0	0	1	0	0	1	0	1	1	1	1	1	0	0	0	1	0	0	1	0	1	0	0	1	1	1	0	0	15	50
Social impact modeling	1	1	1	0	0	0	1	0	0	0	1	1	1	1	0	0	1	0	0	1	1	0	0	1	0	0	1	1	1	0	15	50
Authors and content multimodal analysis	1	1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	0	0	0	1	0	0	1	0	1	1	0	0	20	67
Content and social (networking) impact multimodal analysis	1	0	1	1	1	0	1	0	1	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	1	1	0	0	15	50
Authors and social (networking) impact multimodal analysis	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	0	0	23	77
Reasons of successful deception	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	0	0	0	1	12	40
Analysis of propagation characteristics	1	0	0	0	0	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	20	67
Radicalization/polarization	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	1	1	1	10	33
Social impact analysis via social behavioral patterns	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0	1	1	1	1	0	1	11	37
Total:	12	8	7	5	5	3	9	6	8	9	9	9	9	11	3	8	12	6	6	7	9	6	7	3	6	7	11	8	3	5		

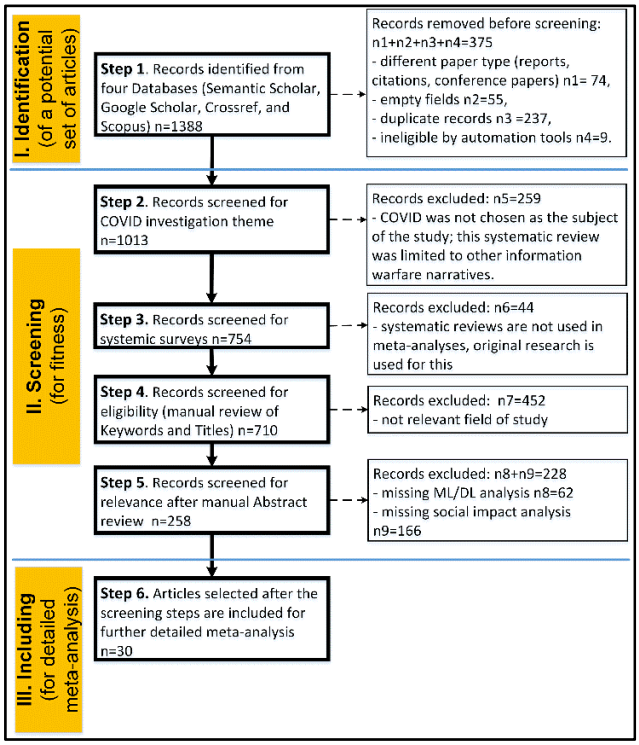


FIGURE 19. PRISMA flowchart for selecting the most relevant articles for detailed meta-analysis in Domain#3.

- The search results indicate that research domains #1 and #3 are often related, as they deal with authors' social networking through social media. In a sixth step, two relevant articles that were found in domain #1 of this study were also included.
- Thus, the final number of papers selected for full-text FNPD impact analysis at the meta-analysis stage is 30 (68.8 citations per paper).

APPENDIX E
KEY ANALYSIS CRITERIA (DOMAIN#3)

See Table 6.

APPENDIX F
SPECIFIC ANALYSIS CRITERIA (DOMAIN#3)

See Table 7.

ACKNOWLEDGMENT

Project title: 'Propaganda and Disinformation Research: Machine Learning Based Automatic Detection, Impact and Societal Resilience.'

REFERENCES

[1] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Social Netw. Anal. Mining*, vol. 13, no. 1, p. 30, Feb. 2023, doi: 10.1007/s13278-023-01028-5.
[2] J. Prier, "Commanding the trend: Social media as information warfare," *Strategic Stud. Quart.*, vol. 11, no. 4, pp. 50–85, 2017.
[3] T. Norri-Sederholm, E. Norvanto, K. Talvitie-Lamberg, and A. M. Huhtinen, "Misinformation and disinformation in social media as the pulse of Finnish national security," in *Social Media and the Armed Forces*. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-47511-6_12.
[4] M. Ajir and B. Vailliant, "Russian information warfare: Implications for deterrence theory," *Strategic Stud. Quart.*, vol. 12, no. 3, pp. 70–89, 2018.
[5] G. E. Center, "Pillars of Russia's disinformation and propaganda ecosystem," U.S. Dept. State, Washington, DC, USA, GEC Special Rep. GPO 0876, 2020.

[6] G. J. Stein, *Information Warfare*. Maxwell Air Force Base, Alabama: Air University Press, Mar. 26, 2022.
[7] *Information Warfare*, NATO, Washington, DC, USA, Mar. 26, 2022.
[8] J. Aro, "The cyberspace war: Propaganda and trolling as warfare tools," *Eur. View*, vol. 15, no. 1, pp. 121–132, Jun. 2016, doi: 10.1007/s12290-016-0395-5.
[9] J. D. Ohlin, K. H. Govern, and C. O. Finkelstein. (2015). *Cyber War: Law and Ethics for Virtual Conflicts*. [Online]. Available: https://doi.org/10.1093/ACPROF%3AOSO%2F9780198717492.001.0001
[10] M. J. Dupuis and A. Williams, "The spread of disinformation on the Web: An examination of memes on social networking," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 1412–1418.
[11] Y. Liu and Y.-F.-B. Wu, "FNED: A deep network for fake news early detection on social media," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–33, Jul. 2020, doi: 10.1145/3386253.
[12] A. Qayyum, J. Qadir, M. U. Janjua, and F. Sher, "Using blockchain to rein in the new post-truth world and check the spread of fake news," *IT Prof.*, vol. 21, no. 4, pp. 16–24, Jul. 2019, doi: 10.1109/MITP.2019.2910503.
[13] L. V. S. Lakshmanan, M. Simpson, and S. Thirumuruganathan, "Combating fake news: A data management and mining perspective," *Proc. VLDB Endowment*, vol. 12, no. 12, pp. 1990–1993, Aug. 2019, doi: 10.14778/3352063.3352117.
[14] K. Scrivens and C. Smith, "Four interpretations of social capital: An Agenda for measurement," Econ. Co-Oper. Develop. (OECD), Paris, France, STD/DOC(2013)6, Work. Paper 55, 2013. [Online]. Available: https://doi.org/10.1787/5JZBCX010WMT-EN
[15] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," 2017, *arXiv:1704.07506*.
[16] S. Gupta, R. Thirukovalluru, M. Sinha, and S. Mannarswamy, "CIMTDetect: A community infused matrix-tensor coupled factorization based method for fake news detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 278–281, doi: 10.1109/ASONAM.2018.8508408.
[17] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 312–320, doi: 10.1145/3289600.3290994.
[18] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, Jun. 2020, doi: 10.1089/big.2020.0062.
[19] Q. Zhang, Q. Qiu, W. Guo, K. Guo, and N. Xiong, "A social community detection algorithm based on parallel grey label propagation," *Comput. Netw.*, vol. 107, pp. 133–143, Oct. 2016, doi: 10.1016/j.comnet.2016.06.002.
[20] S. Shelke and V. Attar, "Source detection of rumor in social network—A review," *Online Social Netw. Media*, vol. 9, pp. 30–42, Jan. 2019, doi: 10.1016/j.osnem.2018.12.001.
[21] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 549–556, doi: 10.1609/aaai.v34i01.5393.
[22] M. Alkhamees, S. Alsaleem, M. Al-Qurishi, M. Al-Rubaian, and A. Hussain, "User trustworthiness in online social networks: A systematic review," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107159, doi: 10.1016/j.asoc.2021.107159.
[23] N. E. H. Ben Chaabene, A. Bouzeghoub, R. Guetari, and H. H. B. Ghezala, "Deep learning methods for anomalies detection in social networks using multidimensional networks and multimodal data: A survey," *Multimedia Syst.*, vol. 28, no. 6, pp. 2133–2143, Dec. 2022, doi: 10.1007/s00530-020-00731-z.
[24] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113383, doi: 10.1016/j.eswa.2020.113383.
[25] M. R. Kondamudi, S. R. Sahoo, L. Chouhan, and N. Yadav, "A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 6, Jun. 2023, Art. no. 101571, doi: 10.1016/j.jksuci.2023.101571.
[26] N. A. Patel and R. Patel, "A survey on fake review detection using machine learning techniques," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–6, doi: 10.1109/CCAA.2018.8777594.
[27] S. Ahmed, K. Hinkelmann, and F. Corradini, "Combining machine learning with knowledge engineering to detect fake news in social networks—A survey," 2022, *arXiv:2201.08032*.
[28] S. I. Manzoor, J. Singla, and Nikita, "Fake news detection using machine learning approaches: A systematic review," in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 230–234, doi: 10.1109/ICOEI.2019.8862770.
[29] M. F. Mridha, A. J. Keya, Md. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021, doi: 10.1109/ACCESS.2021.3129329.
[30] R. Katarya and M. Massoudi, "Recognizing fake news in social media with deep learning: A systematic review," in *Proc. 4th Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Sep. 2020, pp. 1–4, doi: 10.1109/ICCCSP49186.2020.9315255.
[31] H. F. Villela, F. Corrêa, J. S. D. A. N. Ribeiro, A. Rabelo, and D. B. F. Carvalho, "Fake news detection: A systematic literature review of machine learning algorithms and datasets," *J. Interact. Syst.*, vol. 14, no. 1, pp. 47–58, Mar. 2023, doi: 10.5753/jis.2023.3020.
[32] H. T. Phan, N. T. Nguyen, and D. Hwang, "Fake news detection: A survey of graph neural network methods," *Appl. Soft Comput.*, vol. 139, May 2023, Art. no. 110235, doi: 10.1016/j.asoc.2023.110235.
[33] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, "Fake news types and detection models on social media a state-of-the-art survey," in *Proc. Asian Conf. Intell. Inf. Database Syst.*, 2020, pp. 562–573, doi: 10.1007/978-981-15-3380-8_49.
[34] M. Haqi Al-Tai, B. M. Nema, and A. Al-Sherbaz, "Deep learning for fake news detection: Literature review," *Al-Mustansiriyah J. Sci.*, vol. 34, no. 2, pp. 70–81, Jun. 2023.
[35] M. Choras, K. Demestichas, A. Gielczyk, A. Herrero, P. Ksieniewicz, K. Remoundou, D. Urda, and M. Wozniak, "Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study," 2020, *arXiv:2101.01142*.
[36] S. Hakak, W. Z. Khan, S. Bhattacharya, G. T. Reddy, and K. Choo, "Propagation of fake news on social media: Challenges and opportunities," in *Proc. Int. Conf. Comput. Social Netw.*, 2020, pp. 345–353, doi: 10.1007/978-3-030-66046-8_28.
[37] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, and D. M. F. Mattos, "Identifying fake news on social networks based on natural language processing: Trends and challenges," *Information*, vol. 12, no. 1, p. 38, Jan. 2021, doi: 10.3390/info12010038.
[38] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, "Content-based fake news detection with machine and deep learning: A systematic review," *Neurocomputing*, vol. 530, pp. 91–103, Apr. 2023, doi: 10.1016/j.neucom.2023.02.005.
[39] S. Rastogi and D. Bansal, "A review on fake news detection 3T's: Typology, time of detection, taxonomies," *Int. J. Inf. Secur.*, vol. 22, no. 1, pp. 177–212, Feb. 2023, doi: 10.1007/s10207-022-00625-3.
[40] M. Ahsan, M. Kumari, and T. P. Sharma, "Rumors detection, verification and controlling mechanisms in online social networks: A survey," *Online Social Netw. Media*, vol. 14, Nov. 2019, Art. no. 100050, doi: 10.1016/j.osnem.2019.100050.
[41] I. Varlamis, D. Michail, F. Glykou, and P. Tsantilas, "A survey on the use of graph convolutional networks for combating fake news," *Future Internet*, vol. 14, no. 3, p. 70, Feb. 2022, doi: 10.3390/fi14030070.
[42] A. Figueira, N. Guimaraes, and L. Torgo, "Current state of the art to detect fake news in social media: Global trendings and next challenges," in *Proc. 14th Int. Conf. Web Inf. Syst. Technol.*, 2018, pp. 332–339, doi: 10.5220/0007188503320339.
[43] S. Kumar and N. Shah, "False information on web and social media: A survey," 2018, *arXiv:1804.08559*.
[44] A. M. Abbas, "Social network analysis using deep learning: Applications and schemes," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–21, Dec. 2021, doi: 10.1007/s13278-021-00799-z.
[45] S. Nurulain Mohd Rum, R. Mohamed, and A. Asfarian, "Identifying political polarization in social media: A literature review," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 34, no. 1, pp. 80–89, Nov. 2023, doi: 10.37934/araset.34.1.8089.
[46] A. Mahmoudi, D. Jemielniak, and L. Ciechanowski, "Echo chambers in online social networks: A systematic literature review," *IEEE Access*, vol. 12, pp. 9594–9620, 2024, doi: 10.1109/ACCESS.2024.3353054.
[47] A. S. Booth and D. Papaioannou, *Systematic Approaches to a Successful Literature Review*. Newbury Park, CA, USA: Sage, 2016.

[48] K. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and generalizable social bot detection through data selection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 1096–1103, doi: 10.1609/aaai.v34i01.5460.

[49] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "Detection of novel social bots by ensembles of specialized classifiers," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2725–2732, doi: 10.1145/3340531.3412698.

[50] M. Mendoza, E. Providel, M. Santos, and S. Valenzuela, "Detection and impact estimation of social bots in the Chilean Twitter network," *Sci Rep.*, vol. 14, p. 6525, 2024, doi: 10.1038/s41598-024-57227-3.

[51] E. Van Der Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018, doi: 10.1109/ACCESS.2018.2796018.

[52] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," 2019, *arXiv:1906.04210*.

[53] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," in *Proc. Int. Conf. Web Social Media*, 2019, pp. 626–637, doi: 10.1609/icwsm.v14i1.7329.

[54] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Trans. Web*, vol. 13, no. 2, pp. 1–22, May 2019, doi: 10.1145/3316809.

[55] P. G. Efthimion, S. Payne, and N. Proferes, "Supervised machine learning bot detection techniques to identify social Twitter bots," *SMU Data Sci. Rev.*, vol. 1, no. 2, p. 5, 2018.

[56] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on Twitter," *Comput. Secur.*, vol. 91, Apr. 2020, Art. no. 101715, doi: 10.1016/j.cose.2020.101715.

[57] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci.*, vol. 467, pp. 312–322, Oct. 2018, doi: 10.1016/j.ins.2018.08.019.

[58] F. Wei and U. T. Nguyen, "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings," in *Proc. 1st IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Dec. 2019, pp. 101–109, doi: 10.1109/TPS-ISA48467.2019.00021.

[59] M. Heidari and J. H. Jones, "Using BERT to extract topic-independent sentiment features for social media bot detection," in *Proc. 11th IEEE Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2020, pp. 542–547, doi: 10.1109/uemcon51285.2020.9298158.

[60] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "DEFEND: Explainable fake news detection," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 395–405, doi: 10.1145/3292500.3330935.

[61] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106983, doi: 10.1016/j.asoc.2020.106983.

[62] Y. Han, S. Karunasekera, and C. Leckie, "Graph neural networks with continual learning for fake news detection from social media," 2020, *arXiv:2007.03316*.

[63] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," 2019, *arXiv:1902.06673*.

[64] A. Giachanou, B. Ghanem, and P. Rosso, "Detection of conspiracy propagators using psycho-linguistic characteristics," *J. Inf. Sci.*, vol. 49, no. 1, pp. 3–17, Feb. 2023, doi: 10.1177/0165551520985486.

[65] S. Haider, L. Luceri, A. Deb, A. Badawy, N. Peng, and E. Ferrara, "Detecting social media manipulation in low-resource languages," in *Companion Proc. ACM Web Conf.*, New York, NY, USA, Apr. 2023, pp. 1358–1364, doi: 10.1145/3543873.3587615.

[66] A. M. U. D. Khanday, M. A. Wani, S. T. Rabani, and Q. R. Khan, "Hybrid approach for detecting propagandistic community and core node on social networks," *Sustainability*, vol. 15, no. 2, p. 1249, Jan. 2023, doi: 10.3390/su15021249.

[67] C. Marche, I. Cabiddu, C. G. Castangia, L. Serreli, and M. Nitti, "Implementation of a multi-approach fake news detector and of a trust management model for news sources," *IEEE Trans. Services Comput.*, vol. 16, no. 6, pp. 4288–4301, Nov. 2023, doi: 10.1109/TSC.2023.3311629.

[68] T. Hamdi, H. Slimi, I. Bounhas, and Y. Slimani, "A hybrid approach for fake news detection in Twitter based on user features and graph embedding," in *Proc. Int. Conf. Distrib. Comput. Internet Technol.*, 2020, pp. 266–280, doi: 10.1007/978-3-030-36987-3_17.

[69] J. Zhang, B. Dong, and P. S. Yu, "FakeDetector: Effective fake news detection with deep diffusive neural network," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 1826–1829, doi: 10.1109/ICDE48307.2020.00180.

[70] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User preference-aware fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2051–2055, doi: 10.1145/3404835.3462990.

[71] V.-H. Nguyen, K. Sugiyama, P. Nakov, and M.-Y. Kan, "FANG: Leveraging social context for fake news detection using graph representation," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 1165–1174, doi: 10.1145/3340531.3412046.

[72] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: A tool for fake news collection, detection, and visualization," *Comput. Math. Org. Theory*, vol. 25, no. 1, pp. 60–71, Mar. 2019, doi: 10.1007/s10588-018-09280-3.

[73] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. 12th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2019, pp. 312–320.

[74] A. Jarrahi and L. Safari, "Evaluating the effectiveness of publishers' features in fake news detection on social media," *Multimedia Tools Appl.*, vol. 82, no. 2, pp. 2913–2939, Jan. 2023, doi: 10.1007/s11042-022-12668-8.

[75] L. Abualigah, Y. Y. Al-Ajlouni, M. S. Daoud, M. Altalhi, and H. Migdady, "Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe," *Social Netw. Anal. Mining*, vol. 14, no. 1, pp. 1–16, Feb. 2024, doi: 10.1007/s13278-024-01198-w.

[76] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3834–3840, doi: 10.24963/ijcai.2018/533.

[77] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 354–361, doi: 10.1609/aaai.v32i1.11268.

[78] S. A. Alhosseini, R. B. Tareaf, P. Najafi, and C. Meinel, "Detect me if you can: Spam bot detection using inductive representation learning," in *Proc. Companion World Wide Web Conf.*, May 2019, pp. 148–153, doi: 10.1145/3308560.3316504.

[79] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/MIS.2019.2899143.

[80] K. Shu, S. Dumais, A. H. Awadallah, and H. Liu, "Detecting fake news with weak social supervision," *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 96–103, Jul. 2021.

[81] C.-O. Truică and E.-S. Apostol, "It's all in the embedding! Fake news detection using document embeddings," *Mathematics*, vol. 11, no. 3, p. 508, Jan. 2023, doi: 10.3390/math11030508.

[82] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," *Inf. Process. Manage.*, vol. 58, no. 6, Nov. 2021, Art. no. 102712.

[83] M. Davoudi, M. R. Moosavi, and M. H. Sadreddini, "DSS: A hybrid deep model for fake news detection using propagation tree and stance network," *Expert Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116635.

[84] C.-O. Truică, E.-S. Apostol, and P. Karras, "DANES: Deep neural network ensemble architecture for social and textual context-aware fake news detection," *Knowl. Syst.*, vol. 294, Jun. 2024, Art. no. 111715.

[85] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, and P. Karras, "MCWDST: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media," *IEEE Access*, vol. 11, pp. 125861–125873, 2023.

[86] Y.-F. Huang and P.-H. Chen, "Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms," *Expert Syst. Appl.*, vol. 159, Nov. 2020, Art. no. 113584, doi: 10.1016/j.eswa.2020.113584.

[87] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with capsule neural networks," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 106991.

[88] A. Altheneyan and A. Alhadlaq, "Big data ML-based fake news detection using distributed learning," *IEEE Access*, vol. 11, pp. 29447–29463, 2023.

[89] I. Ahmad, M. N. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complex*, vol. 2020, Oct. 2020, Art. no. 8885861, doi: 10.1155/2020/8885861.

[90] X. Liu, K. Ma, Q. Wei, K. Ji, B. Yang, and A. Abraham, "G-HFIN: Graph-based hierarchical feature integration network for propaganda detection of we-media news articles," *Eng. Appl. Artif. Intell.*, vol. 132, Jun. 2024, Art. no. 107922.
[91] X. Liu, K. Ma, K. Ji, Z. Chen, and B. Yang, "Graph-based multi-information integration network with external news environment perception for propaganda detection," *Int. J. Web Inf. Syst.*, vol. 20, no. 2, pp. 195–212, Feb. 2024.
[92] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," *Cognit. Syst. Res.*, vol. 61, pp. 32–44, Jun. 2020, doi: 10.1016/j.cogsys.2019.12.005.
[93] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (CNN-LSTM)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020, doi: 10.1109/ACCESS.2020.3019735.
[94] A. Mallik and S. Kumar, "Word2Vec and LSTM based deep learning technique for context-free fake news detection," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 919–940, Jan. 2024.
[95] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *Int. J. Inf. Manage. Data Insights*, vol. 1, no. 1, Apr. 2021, Art. no. 100007, doi: 10.1016/j.jjimei.2020.100007.
[96] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "DeClarE: Debunking fake news and false claims using evidence-aware deep learning," 2018, *arXiv:1809.06416*.
[97] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali, and M. Abomhara, "Advancing fake news detection: Hybrid deep learning with FastText and explainable AI," *IEEE Access*, vol. 12, pp. 44462–44480, 2024.
[98] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 4, pp. 881–893, Aug. 2021.
[99] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114171, doi: 10.1016/j.eswa.2020.114171.
[100] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Phys. A, Stat. Mech. Appl.*, vol. 540, Feb. 2020, Art. no. 123174.
[101] M. I. Nadeem, S. A. H. Mohsan, K. Ahmed, D. Li, Z. Zheng, M. Shafiq, F. K. Karim, and S. M. Mostafa, "HyproBert: A fake news detection model based on deep hypercontext," *Symmetry*, vol. 15, no. 2, p. 296, Jan. 2023.
[102] G. Güler and S. Gündüz, "Deep learning based fake news detection on social media," *Int. J. Inf. Secur. Sci.*, vol. 12, no. 2, pp. 1–21, 2023.
[103] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, "Behind the cues: A benchmarking study for fake news detection," *Expert Syst. Appl.*, vol. 128, pp. 201–213, Aug. 2019.
[104] R. K. Kaliyar, A. Goswami, and P. Narang, "DeepFakE: Improving fake news detection using tensor decomposition-based deep neural network," *J. Supercomput.*, vol. 77, no. 2, pp. 1015–1037, Feb. 2021, doi: 10.1007/s11227-020-03294-y.
[105] J. Choi, T. Ko, Y. Choi, H. Byun, and C.-K. Kim, "Dynamic graph convolutional networks with attention mechanism for rumor detection on social media," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256039, doi: 10.1371/journal.pone.0256039.
[106] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/s11042-020-10183-2.
[107] Y. Lu and C. Li, "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020. [Online]. Available: https://doi.org/10.18653/V1
[108] N. G. des Mesnards, D. S. Hunter, Z. el Hjouji, and T. Zaman, "Detecting bots and assessing their impact in social networks," *Operations Res.*, vol. 70, no. 1, pp. 1–22, Jan. 2022, doi: 10.1287/opre.2021.2118.
[109] G. Lingam, R. R. Rout, and D. V. L. N. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks," *Appl. Intell.*, vol. 49, no. 11, pp. 3947–3964, Nov. 2019, doi: 10.1007/s10489-019-01488-3.
[110] B. Huang and K. M. Carley, "Discover your social identity from what you tweet: A content based approach," in *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Cham, Switzerland: Springer, 2020, pp. 23–37.
[111] C. Schwartz and R. Overdorf, "Disinformation from the inside: Combining machine learning and journalism to investigate sockpuppet campaigns," in *Proc. Companion Web Conf.*, Apr. 2020, pp. 623–628, doi: 10.1145/3366424.3385777.
[112] J. Jing, F. Li, B. Song, Z. Zhang, and K. R. Choo, "Disinformation propagation trend analysis and identification based on social situation analytics and multilevel attention network," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 2, pp. 507–522, Apr. 2023, doi: 10.1109/TCSS.2022.3169132.
[113] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 258–265, doi: 10.1109/ASONAM.2018.8508646.
[114] N. Hajli, U. Saeed, M. Tajvidi, and F. Shirazi, "Social bots and the spread of disinformation in social media: The challenges of artificial intelligence," *Brit. J. Manage.*, vol. 33, no. 3, pp. 1238–1253, Jul. 2022, doi: 10.1111/1467-8551.12554.
[115] J. Stein, M. Keuschnigg, and A. van de Rijt, "Network segregation and the propagation of misinformation," *Sci Rep.*, vol. 13, no. 1, p. 917, Jan. 2023, doi: 10.1038/s41598-022-26913-5.
[116] A. Concepcion and C. Sy, "Modeling the spread of fake news on social networking sites using the system dynamics approach," *ASEAN Eng. J.*, vol. 13, no. 4, pp. 69–78, Oct. 2023, doi: 10.11113/aej.v13.19251.
[117] N. Kratzke, "How to find orchestrated trolls? A case study on identifying polarized Twitter echo chambers," *Computers*, vol. 12, no. 3, p. 57, 2023, doi: 10.20944/preprints202302.0032.v1.
[118] R. Pandey, M. Pandey, and A. N. Nazarov, "Modelling information warfare dynamics to counter propaganda using a nonlinear differential equation with a PINN-based learning approach," *Int. J. Inf. Technol.*, vol. 16, no. 3, pp. 1527–1538, Mar. 2024, doi: 10.1007/s41870-023-01684-y.
[119] M. Hohmann, K. Devriendt, and M. Coscia, "Quantifying ideological polarization on a network using generalized Euclidean distance," *Sci Adv.*, vol. 9, Mar. 2023, Art. no. eabq2044, doi: 10.1126/sciadv.abq2044.
[120] A. G. Rincón, S. B. Moreno, B. Rodríguez-Canovas, R. L. C. Barbosa, and D. R. A. Franco, "Social networks, disinformation and diplomacy: A dynamic model for a current problem," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–14, Aug. 2023, doi: 10.1057/s41599-023-01998-z.
[121] N. A. Gabriel, D. A. Broniatowski, and N. F. Johnson, "Inductive detection of influence operations via graph learning," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, Art. no. 22571, doi: 10.1038/s41598-023-49676-z.
[122] H. W. A. Hanley, D. Kumar, and Z. Durumeric, "Specious sites: Tracking the spread and sway of spurious news stories at scale," 2023, *arXiv:2308.02068*.
[123] F. Alatawi, P. Sheth, and H. Liu, "Quantifying the echo chamber effect: An embedding distance-based approach," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, New York, NY, USA, Nov. 2023, pp. 38–45, doi: 10.1145/3625007.3627731.
[124] J. Liu, S. Huang, N. M. Aden, N. F. Johnson, and C. Song, "Emergence of polarization in coevolving networks," *Phys. Rev. Lett.*, vol. 130, no. 3, Jan. 2023, Art. no. 037401, doi: 10.1103/PhysRevLett.130.037401.
[125] M. Amendola, D. Cavaliere, C. De Maio, G. Fenza, and V. Loia, "Towards echo chamber assessment by employing aspect-based sentiment analysis and GDM consensus metrics," *Online Social Netw. Media*, vols. 39–40, Jan. 2024, Art. no. 100276, doi: 10.1016/j.osnem.2024.100276.
[126] S. Miller, "Cognitive warfare: An ethical analysis," *Ethics Inf. Technol.*, vol. 25, no. 3, pp. 1–10, 2023, doi: 10.1007/s10676-023-09717-7.
[127] G. Sansonetti, F. Gasparetti, G. D'aniello, and A. Micarelli, "Unreliable users detection in social media: Deep learning techniques for automatic detection," *IEEE Access*, vol. 8, pp. 213154–213167, 2020.
[128] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: A transformer-based approach," *Int. J. Data Sci. Analytics*, vol. 13, no. 4, pp. 335–362, May 2022, doi: 10.1007/s41060-021-00302-z.
[129] M. Al Atiqi, S. Chang, and H. Deguchi, "Agent-based approach to resolve the conflicting observations of online echo chamber," in *Proc. 11th Int. Conf. Soft Comput. Intell. Syst., 21st Int. Symp. Adv. Intell. Syst. (SCIS-ISIS)*, Dec. 2020, pp. 1–6.
[130] I. V. Kozitsin and A. G. Chkhartishvili, "Users' activity in online social networks and the formation of echo chambers," in *Proc. 13th Int. Conf. Manage. Large-Scale Syst. Develop. (MLSD)*, Sep. 2020, pp. 1–5.
[131] M. Al Atiqi, S. Chang, and D. Hiroshi, "Agent-based approach to echo chamber reduction strategy in social media," in *Proc. Joint 10th Int. Conf. Soft Comput. Intell. Syst. (SCIS), 19th Int. Symp. Adv. Intell. Syst. (ISIS)*, Dec. 2018, pp. 1301–1306.

[132] L. Shi, Y. Cheng, J. Shao, X. Wang, and H. Sheng, "Leader-follower opinion dynamics of signed social networks with asynchronous trust/distrust level evolution," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 495–509, Mar. 2022.

[133] M. Z. Rácz and D. E. Rigobon, "Towards consensus: Reducing polarization by perturbing social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 6, pp. 3450–3464, Dec. 2023.

[134] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Sci.*, vol. 8, no. 1, p. 35, Dec. 2019.

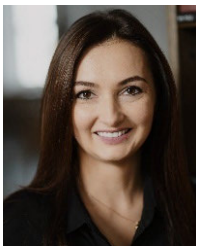
[135] J. Zhu, P. Ni, G. Tong, G. Wang, and J. Huang, "Influence maximization problem with echo chamber effect in social network," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 5, pp. 1163–1171, Oct. 2021.

[136] O. Dedehayir and M. Steinert, "The hype cycle model: A review and future directions," *Technol. Forecasting Social Change*, vol. 108, pp. 28–41, Jul. 2016.

DARIUS PLIKYNAS was born in Lithuania, in November 1972. He received the degree in physics engineering and the Ph.D. degree in economics from Vilnius University, in 1997 and 2003, respectively.

Currently, he is a Senior Research Fellow and a Professor with the Department of Mathematics and Informatics (Institute of Data Science and Digital Technologies), and the Department of Communications, Vilnius University. He is the author of several monographs, numerous research projects, and more than 60 research articles. His areas of interest include modeling individual cognitive and group social processes using computational intelligence techniques, agent-based simulation systems, neuroscience, physics-based methods, complexity theory, distributed cognitive systems, and social networks. His research interests include multidisciplinary sciences, integrating social, natural, and technological domains.

IEVA RIZGELIENĖ received the master's degree in financial and actuarial mathematics from Vilnius University, in 2016, and the master's degree in big data analytics from Kaunas University of Technology, in 2022. She is currently an experienced Data Scientist and a Ph.D. Student with the Institute of Data Science and Digital Technologies, Vilnius University. Her primary research interests include propaganda detection using deep learning models, emphasizing low-resource languages, and the analysis of propaganda narratives and techniques.



GRAŽINA KORVEL (Member, IEEE) received the Ph.D. degree from the Institute of Data Science and Digital Technologies, Vilnius University, in 2013. She is currently a Senior Researcher with the Institute of Data Science and Digital Technologies, Vilnius University. Since 2022, she has been a member with the Young Academy of the Lithuanian Academy of Sciences. Her research interests include speech and music signal processing, natural language processing, development of mathematical models, and applications of computational intelligence. She is a three-time winner of the Lithuanian Academy of Sciences Young Scientist Award. She received acknowledgment from the Prime Minister of Lithuania for her obtained scientific results, in 2013 and 2019.