SYSTEMATIC OVERVIEW OF MACHINE LEARNING APPLIED FOR PROPAGANDA SOCIAL IMPACT RESEARCH

Darius Plikynas

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

ABSTRACT

The proliferation of fake news, propaganda, and disinformation (FNPD) in the era of generative AI and information warfare poses significant challenges to societal cohesion and democratic processes. This systematic review examines recent advances in machine learning (ML) techniques for detecting and assessing the social impact of FNPD. Employing the PRISMA framework, we analyze promising ML/DL methodologies and hybrid approaches in combating the spread of conspiracy theories, echo chambers, and filter bubbles that contribute to social polarization and radicalization. Our findings highlight the potential of AI-driven solutions in identifying malicious social media accounts, organized troll networks, and bot activities that target specific demographics and manipulate public discourse. We also explore future research directions for developing more robust FNPD detection systems and mitigating the fragmentation of social networks of trust and cooperation. This review provides valuable insights for researchers and policymakers addressing the complex challenges of information integrity in the digital age.

KEYWORDS

Machine Learning, Deep learning, Propaganda and Disinformation, Social Impact Analysis, PRISMA Systematic Review

1. INTRODUCTION

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.

A robust social media ecosystem refers to a considerable proportion of the public that is regularly active on social media. This activity generates large, dense networks that facilitate the rapid spread of information. These platforms vary in format, regulation, online culture, and popularity. Therefore, social media users increasingly become the target of FNPD activities aimed at influencing their perception of reality [1,2].

Unlike more traditional forms of cyber-attack, cyber operations today target people within a society, influencing their beliefs as well as their behaviour 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 [3]. Socio-political cleavages are key to increasing the likelihood of domestic

David C. Wyld et al. (Eds): NLAI, IBCOM, MLDS, GridCom, SCAI, ICCSEA, SPPR – 2024
pp. 25-40, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142203

political instability, including atrocities. These include significant social and political polarization, anti-democratic or weakened democratic regimes, and severe governance or security crises.

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 [4]. 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 advanced ML/DL and other approaches while addressing FNPD social impact.

This is a less explored area of research in terms of ML/DL deployment and is more challenging in estimating FNPD social impact metrics. 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. [5-7].

Several previous systemic reviews have analyzed the social implications of the FNPD. For example, Ahmed, Hinkelmann, and Corradini (2022) [8] and Ahsan et al. (2019) [9] 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. Chorąś et al. (2021) [10] and Varlamis et al. (2022) [11], 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. (2018) [12] and Kumar and Shah (2018) [13] 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.

Abbas (2021) [14] provides an overview of the state of the art in different applications of OSN analysis using deep learning techniques. Similarly, Chaabene et al. (2021) [15] provided an overview of several methods that aim to solve the problem of detecting abnormal behavior in social media. Aïmeur, Amri, and Brassard (2023) [16] 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. Siti Nurulain Mohd Rum, Raihani Mohamed, and Auzi Asfarian (2024) [17] examined computing methods and approaches employed by the existing works for identifying political polarization in social media. Mahmoudi A., Jemielniak D., and Ciechanowski L. (2024) [18] 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.

Despite the large number of other systematic reviews mentioned above, there are important areas of analysis (niches) in the field of social impact research that have not yet received sufficient research attention, such as (i) social impact research via social behavioural patterns analysis, (ii) radicalization and polarization research, (iii) reasons for successful deception, (iv) social impact modelling, (v) echo chamber polarization effect, (vi) cognitive warfare.

Let us briefly summarize the content of the systemic review presented. The second section describes the selection process of relevant articles using the PRISMA systemic methodology. The third section presents the main results of a meta-analysis using a set of qualitative and quantitative criteria. The fourth section presents the discussion in terms of the main findings, limitations, and further research. The fifth and final section is a summary of the main findings.

2. SEARCHING: PRISMA METHODOLOGY

Methodology plays a crucial role in conducting a systematic literature review (SLR). For this study, we have chosen to follow the approach outlined in the PRISMA Statement [17, 18], which is a widely accepted checklist used by researchers worldwide to guide and inform the development of systemic literature reviews.

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 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 Figure 1. The four-phase flowchart shows the main selection phases: identification, screening and inclusion. Such an approach makes the selection process transparent by reporting the exclusion decisions made at different stages of the systematic review. Articles were removed at the different selection stages until we reached the final list of records for full-text analysis in the meta-analysis phase.

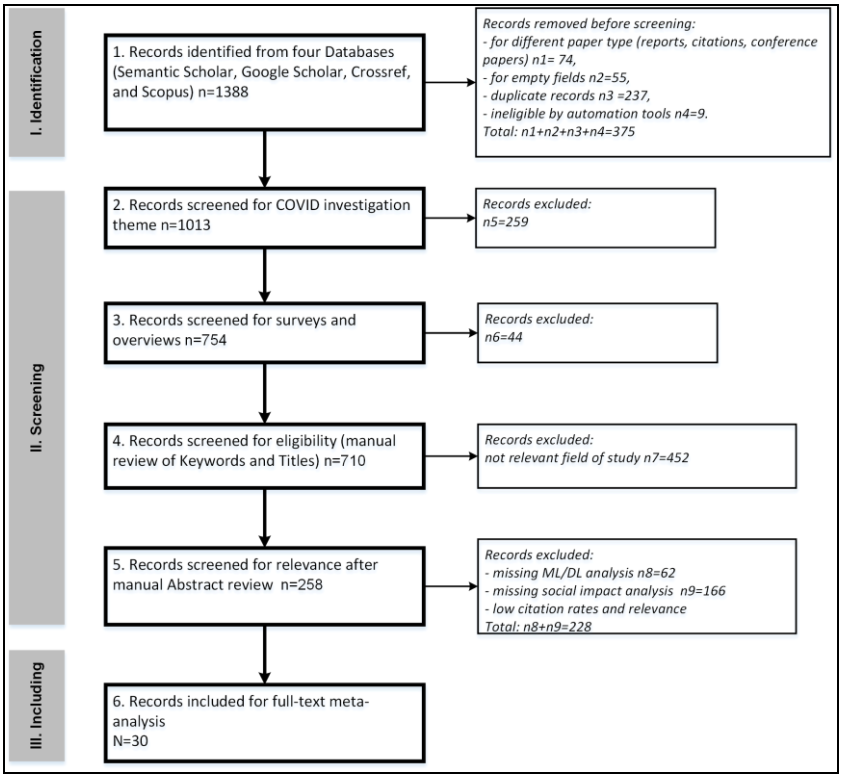

Figure 1. PRISMA flow diagram for the most relevant papers selection

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. Thus, 30 papers were selected for the full-text FNPD impact analysis at the meta-analysis stage.

3. META- ANALYSIS

Below we have conducted a meta-analysis to identify patterns of ML/DL application in the FNPD social impact research literature. The systematic nature of the present review minimizes bias, ensures transparency, and enhances the replicability of findings. We used a set of criteria for the systematic analysis of selected papers in each research domain.

The following meta-analysis of 30 selected articles provides an overview of FNPD social impact research. First, some statistics. Selected articles were cited on average of 69 times, average publication date 2021, average use of the term 'social' 82 times. 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%.

After careful consideration, we have identified several key meta-analysis criteria.
Novelty. The main novelties of the selected articles can be summarized from different perspectives as follows.

(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 [19-21].

(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 [22-24].

(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 [25-27].

(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 sockpuppets, particularly "infiltrators," who blend into genuine online communities [27,28].

(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 [29-31].

(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 [32-34].

(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 [36-39].

(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 [40-42].

(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 [43,44].

(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 [45,46].

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. Figure 2 shows the temporal dynamics of novelty in this domain.

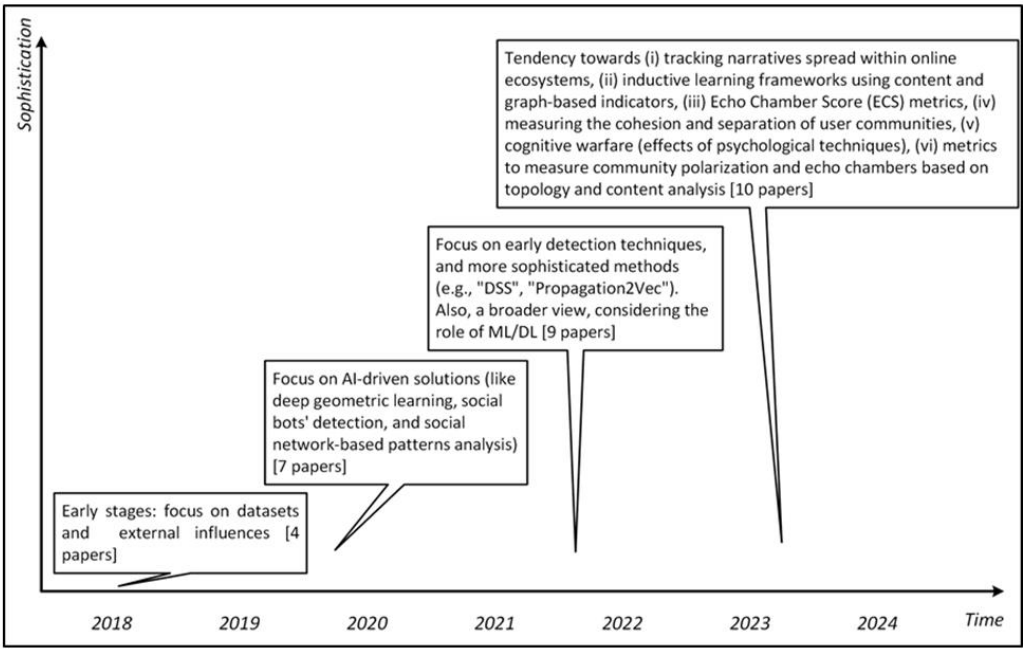

Figure 2. Key innovation trends for the period 2018-2024

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 [19], label propagation (method used to infer the ideological leanings of users within a network, demonstrating how beliefs or labels may spread in OSNs) [33], graph convolutional networks (GCN) with attention mechanisms (captures evolving rumor propagation patterns in social structures, emphasizing the temporal dynamics [21], DSS model (incorporates dynamic, static, and structural analysis to understand how information or content traverses through OSNs [24], network-based pattern-driven model (focuses on extracting features from patterns of fake news dissemination on social platforms) [47].

(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) [22], 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) [26], GCAN model (integrates word embeddings, neural networks, and a dual co-attention mechanism to analyze correlations between source content, retweet propagation, and user interactions) [23], 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) [32].

(iii) Bot & User Role Analysis: advanced machine learning techniques for bot detection (underscores the non-human entities that might manipulate social dynamics online) [33], using a statistical physics model (to identify ots and measure their influence on shifting opinions within OSN [25], hierarchical self-attention neural network (delineates how different user roles might influence or be influenced in social contexts) [27], supervised machine learning guided by journalistic investigations (by integrating journalistic insights, this method underscores the human-social perspective in validating and understanding online content) [28].

(iv) Community detection and dynamics modelling: agent-based simulation (simulates the behavior of individual users within a social network to understand how information spreads) [35], physics-informed neural networks (modeling of complex social systems) [38], system dynamics modeling (explores how interconnected parts of a system influence each other over time) [40,41], community detection algorithms (identify groups of users within a network who are more likely to interact with each other) [37], user embedding models (analyzing how users with similar ideologies connect) [43], 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) [39], scaling law analysis (explores how different aspects of a system change in relation to each other) [44].

In summary, from the data, 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.

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%).

Size of Datasets (where specified): largest 1858575 entries [32], smallest 30,000 tweets [34].
In summary, 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.

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

(i) Popular Metrics Used: accuracy [48, 23-25, 27, 34], precision [22, 26, 33], recall [22, 33, 49], F1-Score [20, 49], ROC AUC [19].

(ii) Network analysis & modeling: user segregation [35], metrics of systems dynamics [36], community detection algorithms [37], community detection algorithms [43], network distance measures [39], opinion dynamics [44], consensus metrics [45].

(iii) Auxiliary/additional metrics & methods: descriptive statistics [33], early detection rates [24, 50], linguistic features and social engagements [31], shift in equilibrium opinions [25], statistical indicators (Lorentz curve and Gini coefficient) [32].

(iv) Articles with ambiguous or not explicitly mentioned metrics: [29,47,31,20-22,28,34].

In summary, accuracy emerges as the most popular metric used across the studies, being 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 [32,33]. 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.

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 [19] achieved a 92.7% ROC AUC, the DeepFake model in [22] obtained validation accuracies of 85.86% and 88.64% on two different datasets, and the model in [48]

demonstrated improved accuracy and early detection capabilities compared to existing methods. In article [49] model distinguished between real and fake news with 90% accuracy, and the DSS model [24] surpassed state-of-the-art methods by up to 8.2%. Meanwhile, the network-based pattern-driven approach [47] was robust against manipulations and effective even with limited network data, and Propagation2Vec from [20] outperformed other models by up to 5.55% in F1-score. GCAN, from [51], significantly outperformed existing methods, and the model in [27] boosted its accuracy when combined with a transfer learning scheme.

(ii) Bot Detection and Influence: Significant information about bots emerged from the articles. Article [33] found that 4.9% of liberal users and 6.2% of conservative users were bots. Article [29] observed that bot users are more involved in spreading fake news, while Article 96's Ising model algorithm efficiently identified bots. Article [28] unveiled that as bot detection methods improve, disinformation agents are now more focused on using sockpuppets, especially infiltrators. Article [115] highlighted the critical role of bots in influencing online public opinion and spreading false narratives.

(iii) Insights on Content and Dissemination: Article [33] 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 [52] 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 [26] deep Q-learning algorithm integrated with various social attributes demonstrated improved precision over other algorithms. The Dynamic GCN from [53] outperformed other leading methods in rumor detection. In Article [31], FakeNewsTracker was effective in using linguistic and social engagement features for fake news detection. Article [51] GCAN highlighted suspicious retweeters and specific tweet segments, adding a layer of explainability to the model. Lastly [25] 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 [35]. Confirmation bias, sharing of posts, and algorithmic ranking are critical variables driving this process [36]. Coevolving dynamics of opinions and network structures can lead to stable bipolarized community structures, with phase transitions across different polarization phases [44]. An inductive learning framework identified how echo chambers foster polarization and dysfunctional political discourse [37]. Complementing topology-based metrics with semantic analysis of viewpoints and beliefs is essential to fully capture community closeness and prevailing beliefs [45]. 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 [39].

(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 [41]. Systems for identifying prevalent narratives and aiding fact-checkers in addressing misinformation more quickly are also highlighted [42]. Social media algorithms monitor user behavior, interests, and actions to

recommend relevant content, refining suggestions by adapting and learning from user interactions [35,41,46]. 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 [36,37]. 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.

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 [19]. The dissemination of misinformation can heavily influence democratic discussions, leading to societal confusion and potential instability [33]. 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 [47,22,24]. Instances like the anti-vaccine misinformation during the COVID-19 pandemic underscore the importance of addressing the challenge of fake news [20,32,148].

(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 [26]. The presence of politically motivated bots on OSNs poses a considerable threat to democratic processes [25]. The acceleration of information spread, both factual and fictitious, by social bots emphasizes the need for thorough research to mitigate potential threats [34].

(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 [29]. Events such as the "Pizzagate" tweets during the US elections provide tangible evidence of the consequences of misinformation [31]. The proliferation of false news can potentially benefit certain factions unjustly, whether in political, economic, or psychological domains [23].

(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 [27,28]. Recognizing the roles of various bots and entities offers deeper insights into the dynamics of misinformation spread on digital platforms [27].

(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 [37, 43,45]. They emphasize how echo chambers can make political discourse dysfunctional and exacerbate polarization in open societies, contributing to the identification of problematic interaction patterns [36,39]. 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 [46].

The development of trends in the analysis of the social impact of FNPD has only recently gained momentum, see Figure 3. 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.

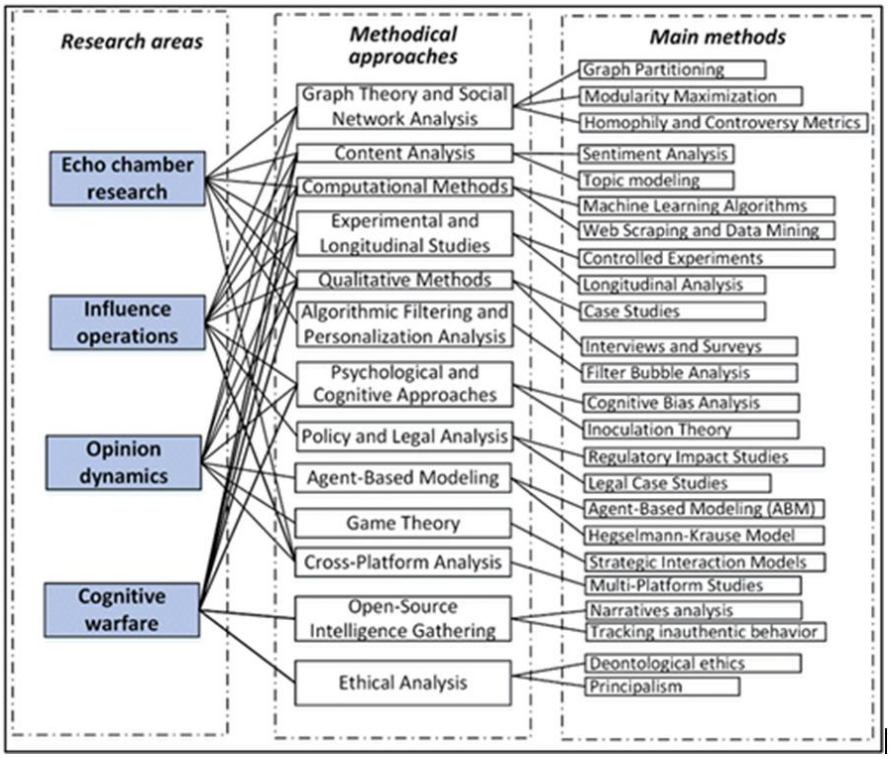

Figure 3. Relationships between main social impact research methodological approaches

4. DISCUSSION

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 the research approaches are assessed and summarized below.

4.1. 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 [33]. In addition, the scope of the datasets is limited, with some focusing narrowly on U.S. politics [49] or exclusively on English-language tweets [54]. Nonrepresentative samples of the populations are limiting the validity of the results [39]. 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 [55]. This problem limits the ability to collect comprehensive social context data.

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, and models may not detect fake news until it has begun to spread, further emphasizing the need for real-time detection [47].

Furthermore, the dynamic nature of social media platforms and user behavior further complicates the detection process, making it essential for models to continuously adapt to these changes [56,57,58].

4.2. Further Research Directions

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. 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.

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.

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).

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.

Adaptation to other social networks. The majority of all FNPD models are built on Twitter datasets, mainly because Twitter allows its data to be freely crawled and used for research. For robustness evaluation, the models should be tested on other online social network datasets to see their adaptability.

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, exploring other algorithms and additional features, as well as the need for datasets with more complex scenarios.

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.

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.

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.

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.

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. In an information war, the strategies employed by these agents often rely on psychological manipulation, exploiting cognitive biases, and capitalizing on existing social or political divides. Therefore, democracies need innovative approaches and tools to discover disguised agents of influence, who are orchestrating covert operations to spread FNPD, influence media narratives, or disrupt social cohesion.

5. CONCLUSIONS

In summary, there are a few key conclusions to be drawn from the systemic research above:

(i) Using the PRISMA systemic review framework, we were able to select well-cited recent papers for detailed meta-analysis covering the most important current research trends.

(ii) 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 user credibility metrics as critical to understanding and evaluating the phenomena in both research domains.

(iii) 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.

(iv) 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, (vi) influence operations, (vii) echo chambers, (viii) opinion dynamics.

(v) 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.

ACKNOWLEDGEMENTS

We are grateful for the financial support provided 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 number S-VIS-23-8. Project title: 'Propaganda and Disinformation Research: Machine Learning Based Automatic Detection, Impact and Societal Resilience'.

REFERENCES

[1] Aro, J. (2016). "The Cyberspace War: Propaganda and Trolling as Warfare Tools", *European View*, Vol. 15, No. 1, pp 121-132.

[2] Liu, Y., & Wu, Y. F. B. (2020). "Fned: a deep network for fake news early detection on social media*", ACM Transactions on Information Systems (TOIS)*, Vol. 38, No. 3, pp 1-33.

[3] Prier, J. (2017). "Commanding the trend: social media as information warfare", Strategic Studies Quarterly, Vol.11, No. 4, pp. 50-85.

[4] Lakshmanan, L. V., Simpson, M., & Thirumuruganathan, S. (2019). "Combating fake news: a data management and mining perspective", *Proceedings of the VLDB Endowment*, Vol. 12, No. 12.

[5] Scrivens, K., & Smith, C. (2013). "Four Interpretations of Social Capital: An Agenda for Measurement", *https://doi.org/10.1787/5JZBCX010WMT-EN*

[6] Tacchini, E., Ballarin, G., Vedova, M.L., Moret, S., & Alfaro, L.D. (2017). "Some Like it Hoax: Automated Fake News Detection in Social Networks", *ArXiv, abs/1704.07506*.

[7] Gupta, S., Thirukovalluru, R., Sinha, M., & Mannarswamy, S. (2018). "CIMTDetect: A Community Infused Matrix-Tensor Coupled Factorization Based Method for Fake News Detection", *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp 278-281.

[8] Ahmed, S., Hinkelmann, K., & Corradini, F. (2022). "Combining machine learning with knowledge engineering to detect fake news in social networks-a survey", *arXiv preprint arXiv:2201.08032*.

[9] Ahsan, M., Kumari, M., & Sharma, T.P. (2019). "Rumors detection, verification and controlling mechanisms in online social networks: A survey", *Online Soc. Networks Media*, Vol. 14.

[10] Choraś, M., Demestichas, K.P., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., Urda, D., & Woźniak, M. (2020). "Advanced Machine Learning Techniques for Fake News (Online Disinformation) Detection: A Systematic Mapping Study", *ArXiv, abs/2101.01142*.

[11] Varlamis, I., Michail, D., Glykou, F., & Tsantilas, P. (2022). "A Survey on the Use of Graph Convolutional Networks for Combating Fake News. Future Internet", Vol. 14, No. 70.

[12] Figueira, Á., Guimarães, N., & Torgo, L. (2018). "Current State of the Art to Detect Fake News in Social Media: Global Trendings and Next Challenges. *International Conference on Web Information Systems and Technologies*.

[13] Kumar, S., & Shah, N. (2018). "False Information on Web and Social Media: A Survey", *ArXiv, abs/1804.08559*.

[14] Abbas, A.M. (2021). 'Social network analysis using deep learning: applications and schemes", *Social Network Analysis and Mining*, Vol. 11, pp 1-21.

[15] Ben Chaabene, N.E., Bouzeghoub, A., Guetari, R., & Ghézala, H.H. (2021). "Deep learning methods for anomalies detection in social networks using multidimensional networks and multimodal data: a survey", *Multimedia Systems*, Vol. 28, pp 2133 - 2143.

[16] Aïmeur, E., Amri, S., & Brassard, G. (2023, February 9). "Fake news, disinformation and misinformation in social media: a review", *Springer Science+Business Media*, Vol. 13 No 1.

[17] Rum S. N. M., Mohamed R., and Asfarian A. (2024). "Identifying Political Polarization in Social Media: A Literature Review", *Journal of Advanced Research in Applied Sciences and Engineering Technology*, Vol. 34, No. 1.

[18] Mahmoudi A., Jemielniak D., and Ciechanowski L. (2024) "Echo Chambers in Online Social Networks: A Systematic Literature Review," *IEEE Access*, Vol. 12, pp. 9594-9620.

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

[20] Kaliyar, R.K., Goswami, A., & Narang, P. (2020). "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network", *The Journal of Supercomputing*, pp 1-23.

[21] Choi, J., Ko, T., Choi, Y., Byun, H., & Kim, C. (2021). "Dynamic graph convolutional networks with attention mechanism for rumor detection on social media", *PLoS ONE*, Vol. 16.

[22] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach", *Multimedia tools and applications*, Vol. 80, No 8, pp 11765-11788.

[23] Lu, Y., & Li, C. (2020). "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media", *Annual Meeting of the Association for Computational Linguistics*.

[24] Davoudi, M., Moosavi, M.R., & Sadreddini, M.H. (2022). "DSS: A hybrid deep model for fake news detection using propagation tree and stance network", *Expert Syst. Appl.*, Vol 198, 116635.

[25] Mesnards, N.G., KIM, S., Hjouji, Z.E., & Zaman, T. (2018). "Detecting Bots and Assessing Their Impact in Social Networks", *Oper. Res.*, Vol. 70, pp 1-22.

[26] Lingam, G., Rout, R.R., & Somayajulu, D.V. (2019). "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks", *Applied Intelligence*, pp 1-18.

[27] Huang, B., & Carley, K. M. (2020). "Discover your social identity from what you tweet: a content-based approach", *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pp. 23-37

[28] Schwartz, C., & Overdorf, R. (2020). "Disinformation from the Inside: Combining Machine Learning and Journalism to Investigate Sockpuppet Campaigns", *Companion Proceedings of the Web Conference 2020*.

[29] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media", *Big data*, Vol. 8, No 3, pp 171-188.

[31] Shu, K., Mahudeswaran, D., & Liu, H. (2019). "FakeNewsTracker: a tool for fake news collection, detection, and visualization", *Computational and Mathematical Organization Theory*, Vol. 25, pp 60-71.

[32] Jing, J., Li, F., Song, B., Zhang, Z., & Choo, K.R. (2023). Disinformation Propagation Trend Analysis and Identification Based on Social Situation Analytics and Multilevel Attention Network. IEEE Transactions on Computational Social Systems, 10, 507-522

[33] Badawy, A., Ferrara, E., & Lerman, K. (2018). "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign", *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp 258-265

[34] Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2021). "Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence", *British Journal of Management*, Vol. 33, No. 3, pp 1238-1253

[35] Stein J, Keuschnigg M, van de Rijt A. "Network segregation and the propagation of misinformation", *Sci Rep*., Vol. 13, No. 1, pp 917.

[36] Concepcion, Aleena & Sy, Charlle. (2023). "Modeling the spread of fake news on social networking sites using the system dynamics approach", *ASEAN Engineering Journal*., Vol. 13, pp 69-78.

[37] Kratzke, Nane. (2023). "How to find Orchestrated Trolls? A Case Study on Identifying Polarized Twitter Echo Chambers". *10.20944/preprints202302.0032.v1*.

[38] Pandey, Rashmikiran & Pandey, Mrinal & Nazarov, Alexey. (2023). "Modelling information warfare dynamics to counter propaganda using a nonlinear differential equation with a PINN-based learning approach", *International Journal of Information Technology*, Vol. 16. *10.1007/s41870-023-01684-y*.

[39] Hohmann M, Devriendt K, Coscia M. (2023). "Quantifying ideological polarization on a network using generalized Euclidean distance. *Sci Adv*., Vol. 9, No. 9.

[40] Guzman Rincon, Alfredo & Barragán, Sandra & Canovas, Rodriguez & Carrillo Barbosa, Ruby & Africano Franco, David. (2023). "Social networks, disinformation and diplomacy: a dynamic model for a current problem", *Humanities and Social Sciences Communications, Vol*. 10, No. 1, pp 1-14.

[41] Gabriel, N.A., Broniatowski, D.A. & Johnson, N.F. (2023). "Inductive detection of influence operations via graph learning", *Sci Rep, Vol* 13, No. 1, pp 22571.

[42] Hanley, H. W., Kumar, D., & Durumeric, Z. (2023). "Specious Sites: Tracking the Spread and Sway of Spurious News Stories at Scale", *arXiv preprint arXiv:2308.02068*.

[43] Faisal Alatawi, Paras Sheth, and Huan Liu. 2024. "Quantifying the Echo Chamber Effect: An Embedding Distance-based Approach", *In Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '23).* Association for Computing Machinery, New York, NY, USA, pp 38–45.

[44] Liu, Jiazhen & Huang, Shengda & Aden, Nathaniel & Johnson, Neil & Song, Chaoming. (2023). "Emergence of Polarization in Coevolving Networks", *Physical Review Letters*, Vol. 130.

[45] Amendola, Miriam & Cavaliere, Danilo & De Maio, Carmen & Fenza, Giuseppe & Loia, Vincenzo. (2024). "Towards echo chamber assessment by employing aspect-based sentiment analysis and GDM consensus metrics", *Online Social Networks and Media*. Pp 39-40.

[46] Miller,S. (2023) "Cognitive warfare: An ethical analysis", *Ethics Inf. Technol*., Vol. 25, No. 3, pp 1–10.

[47] Zhou, X., & Zafarani, R. (2019). "Network-based Fake News Detection: A Pattern-driven Approach", *ACM SIGKDD explorations newsletter*, Vol. 21, No. 2, pp 48-60.

[48] Raza, S., & Ding, C. (2022). "Fake news detection based on news content and social contexts: a transformer-based approach", *International Journal of Data Science and Analytics*, Vol. 13, No. 4, pp 335-362.

[49] Sansonetti, G., Gasparetti, F., D'aniello, G., & Micarelli, A. (2020). "Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection", *IEEE Access*, Vol. 8, pp 213154-213167.

[50] Al Atiqi, M., Chang, S., & Deguchi, H. (2020). "Agent-based approach to resolve the conflicting observations of online echo chamber", *IEEE Joint 11th international conference on soft computing and intelligent systems and 21st international symposium on advanced intelligent systems (SCIS-ISIS)*, pp 1-6).

[51] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). "Fake news detection: A hybrid CNN-RNN based deep learning approach", *International Journal of Information Management Data Insights*, Vol. 1, No. 1, pp 100007.

[52] Liu, X., Ma, K., Wei, Q., Ji, K., Yang, B., & Abraham, A. (2024). "G-HFIN: graph-based hierarchical feature integration network for propaganda detection of we-media news articles", *Engineering Applications of Artificial Intelligence*, Vol. 132, pp 107922

[53] Choudhary, A., & Arora, A. (2021). "Linguistic feature-based learning model for fake news detection and classification", *Expert Systems with Applications*, Vol. 169, pp 114171.

[54] Güler, G., & Gündüz, S. (2023). "Deep learning-based fake news detection on social media", *International journal of information security science*, Vol. 12, No. 2, pp 1-21.

[55] Truică, C. O., Apostol, E. S., & Karras, P. (2024). "DANES: Deep neural network ensemble architecture for social and textual context-aware fake news detection", *Knowledge-Based Systems*, Vol. 294, pp 111715.

[56] Shu, K., Mahudeswaran, D., Wang, S., & Liu, H. (2019). "Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation", *International Conference on Web and Social Media*.

[57] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. (2019). "Polarization and Fake News: Early Warning of Potential Misinformation Targets", *ACM Trans. Web*, Vol. 13, No. 2, pp 22.

[58] Song, C., Shu, K., & Wu, B. (2021). "Temporally evolving graph neural network for fake news detection. Information Processing & Management:, Vol. 58, No. 6, pp 102712.

AUTHOR

DARIUS PLIKYNAS born on November 19, 1972, in Lithuania. He holds a degree in Physics Engineering (1997) and a PhD in Economics (2003) from Vilnius University. His research spans multidisciplinary sciences, integrating social, natural, and technological domains. 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.

Currently, he is a senior research fellow and professor at Vilnius University, affiliated with both the Department of Mathematics and Informatics (Institute of Data Science and Digital Technologies) and the Department of Communications. He is the author of several monographs, numerous research projects, and over 60 research papers.