Literature Review for Computational Detection and Analysis of Manipulation Techniques on Ukrainian Telegram

Nataliia Volkova

[1] Ukrainian Catholic University
[2] Ilariona Svjentsits'koho Street, 17, Lviv, Ukraine
[3] natallia.volkova@ucu.edu.ua

Abstract. The prevalence of Russian propaganda news is most pronounced within the media landscape of Ukraine, particularly intensifying since the Russian annexation of Crimea in 2014 and experiencing a further surge in 2022 with the full-scale war. Social media platforms witness a proliferation of numerous Russian political-oriented channels disseminating partial truths and outright fabrications. The problem of computational detection and analysis of manipulated news in the Ukrainian social media landscape is urgent, and the solution may lead to a more resilient Russian propaganda society. In this work, we provide an overview of existing computational research on fake news detection and the characteristics of fake news from psychological, social, and political theories.

Keywords: disinformation · fake news · manipulation techniques · Russian propaganda.

1 Introduction and Motivation

The widespread dissemination of fake news poses a significant threat to individuals and society, disrupting the authenticity of the news ecosystem. This global impact has been evident during critical events, such as the 2016 U.S. Presidential election and the U.K. Brexit referendum, the COVID-19 pandemic, the Crimea annexation, and the fully escalated Russian war against Ukraine. In recent incidents, like Israel's defense campaign in Sector Gaza, fake news has influenced decision-making, shaped attitudes, and had tangible consequences.

Fake news, intentionally false and verifiable information, is crafted to potentially mislead readers by distorting facts, taking information out of context, or presenting partial truths. While traditionally associated with entirely false information, recent research highlights a nuanced perspective. Scholars recognize that fake news is rarely devoid of truth, and even factual information can be manipulated to convey a misleading narrative, emphasizing the complexity of evaluating news authenticity.

Social media, due to its nature and business model, has become a weapon for propaganda and a vital tool for disseminating fake news. Complex algorithms

aim to precisely target audiences and ensure repeated visibility of paid posts, potentially contributing to the formation of "filter bubbles," "echo chambers," and "polarization" effects. Such effects contribute to an environment where fake news can attract a substantial audience.

The Russian war against Ukraine underscores the urgent need for an effective tool to detect fake news in the social media landscape. Social media platforms, especially Telegram, as the most popular among others, witness a proliferation of numerous channels disseminating partial truths and outright fabrications to the Ukrainian audience. The contemporary strategy of Russia's "information warfare" deliberately seeks to sow confusion, polarize opinions, instill distrust, and construct a fundamentally distorted worldview. These tactics serve Russia's overarching goals, allowing the state to address its political agenda directly to users as an obvious objective solution.

Detecting fake news on social media in Ukraine is an urgent and essential problem. It could empower fact-checkers, journalists, public figures, and bloggers to catch manipulated news early, preventing its dissemination. Telegram users could gain insights into the reasons behind the stylistic aspects of manipulation texts, fostering resilience to fake news delivered by Russia.

This report aims to summarize existing computational research on fake news detection, encompassing characterizations from psychological, social, and political theories. The structure of the review is as follows: Section 2 describes the literature search and selection method. Section 3 provides the characterization of fake news with an examination of the definition and introduces foundational social, psychological, and political theories related to the spread of fake news while also discussing patterns introduced by social media. Section 4 reviews current work on the style-based news-content model-oriented direction for fake news on social media detection from a text analysis perspective only. Section 5 examines the research gap analyses and outlines future work toward addressing the stated problem. Finally, Section 6 provides our conclusions.

2 The Method for Literature Search and Selection

2.1 Method Motivation and Description

The identification of fake news detection in the media is a pressing subject explored across various disciplines. To encompass the collective efforts of the multidisciplinary research community and enhance our comprehension of the concept, framework, and propaganda techniques, we will draw upon literature from both Computer Science and Social Science fields, including Sociology, Psychology, Linguistics, and Politics.

Employing scholarly platforms, we seek noteworthy and pivotal research articles in these chosen domains using relevant key phrases. We carefully weigh the advantages and disadvantages of these platforms during our search. Additionally, we consider the publication year, focusing on articles from 2014 onwards. Subsequently, we utilize a recommendation platform to identify similar articles,

further diversifying our collection. In the concluding phase, we filtered selecting articles based on exclusion criteria such as a systematic and thorough approach, relevance to the chosen problem, and open access versions.

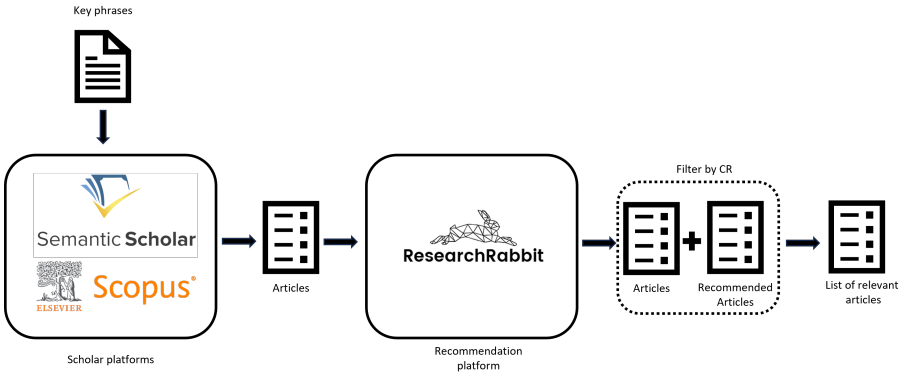

Fig. 1. Method of literature search

2.2 Documents Collection

Step 1: Querying by Keywords and Key Phrases. The first step involves searching for relevant articles through the use of keywords and key phrases. These are essential to identify what we call "seed papers", forming the foundation of our collection. It's crucial to carefully choose keywords and phrases that encompass all vital aspects of our research topic. Our goal is to ensure that the articles in our collection address (i) disinformation and propaganda across various disciplines, (ii) computational research for detecting propaganda techniques in the media, and (iii) the application of manipulative techniques by Russian authorities. The selected keywords and key phrases include: disinformation OR propaganda in Social Studies OR Computer Sciences; computational disinformation OR computational propaganda detection; manipulation OR propaganda techniques in news; Russian manipulation techniques; Russian information warfare OR information operations.

We utilize two widely recognized scholarly platforms, Semantic Scholar and Scopus, to identify relevant articles for our research. Both platforms offer sophisticated filtering capabilities, allowing us to refine our searches based on criteria such as publication years, fields, document types, and the number of citations. Our selection criteria encompass review and research articles from 2014 onwards, prioritizing those with a higher number of citations. Notably, we intentionally included two recent articles from the past year, irrespective of citation count.

Semantic Scholar functions as an academic search engine and research discovery platform, leveraging artificial intelligence and natural language processing

to provide highly pertinent and context-aware results from the scientific literature. The platform offers features such as citation analysis, links to open-access versions of research papers, and encompasses a diverse spectrum of disciplines and sources. However, it is essential to note that Semantic Scholar might not consistently provide the most recent articles. Additionally, the platform's algorithm relies on the completeness of article metadata; incomplete metadata may not be recommended.

Scopus is a comprehensive abstract and citation database spanning various academic disciplines. In addition to journal articles, Scopus includes conference proceedings, an aspect we also incorporate into our literature search. The platform undergoes regular updates, ensuring that researchers maintain access to the most current scholarly information. Nevertheless, it is pertinent to mention that Scopus has limitations in providing open access to certain publications.

Following our meticulous query process utilizing keywords and key phrases, we successfully amassed a collection of 30 articles in alignment with our predefined criteria.

Step 2: Querying by Recommendation Platform. We employed the ResearchRabbit recommendation platform to enhance our collection of articles. ResearchRabbit serves as a literature mapping designed to assist researchers in exploring, organizing, and visualizing relevant information from academic literature. Its specialized features facilitate the identification of connections and gaps within a designated research area or set of articles. While the complete algorithm underlying the search process remains undisclosed to the public, the platform's functionalities include the use of citation relationships in the "Earlier work" and "Later work" functions, while the "Similar work" function incorporates additional criteria and encompasses papers situated a bit further apart in the conceptual space compared to the other two options.

To expand our article collection, we conducted searches based on four criteria: (i) similar works via citation, (ii) references of the relevant paper (backward citation), (iii) citations of the relevant paper (forward citations), and (iv) suggested authors related to our existing collection.

Following this systematic procedure, we expanded our collection by an additional 30 articles.

2.3 Articles Selection

We sorted the list of articles from our collection based on the number of citations, publication year, and relevance to the research topic and carefully reviewed them. We applied the following exclusion criteria to include papers in the next phases of our research:

CR1: We excluded articles that mention false information problems without any attempt to provide systematic classification or even explanations of the problem.

CR2: We excluded articles that discuss the general classification problem without any connection to fake news or style-based text classification.

CR3: We exclude articles with no open-access versions as we could not examine them.

The following section outlines the review of the refined collection of articles

3 Characterization of Fake News

The spreading of disinformation and fake news has long been a concern in the media landscape, but it has gained significant urgency with proficiently incorporated propaganda in social media. The global impact of fake news became particularly apparent during pivotal events such as the U.S. Presidential election and the U.K. Brexit referendum in 2016, the COVID-19 pandemic in 2020, the Crimea annexation in 2014, and the fully escalated Russian war against Ukraine in 2022. A recent example involves the HAMAS attack on Israel and the subsequent Israel defense campaign in Sector Gaza, where fake news in media influenced decision-making, shaped attitudes, and had real-life consequences.

This section aims to accomplish three main objectives: (1) examine the definitions of disinformation, fake news, and propaganda news and their interconnections, (2) introduce foundational social, psychological, and political theories related to the spread of fake news, while also discuss patterns introduced by social media. We took the structure of examining the fake news detection problem proposed by Kai Shu et al.[1], which modification.

3.1 Definitions

To begin, it is imperative to establish a clear understanding of the terminology employed in our study. In this section, we examine the definitions of disinformation, fake news, and politically oriented fake news, meticulously outlining the criteria for inclusion and exclusion of certain concepts.

Nevertheless, within the scholarly community, there exists a lack of consensus regarding the precise definitions and boundaries of the terms "disinformation" and "fake news." To ensure clarity and coherence in the subsequent literature analysis, we establish the following definitions for these terms:

Disinformation is false information, spread deliberately with the intention to mislead and/or deceive[2]. It distinguishes itself from misinformation, which involves the unintentional spread of false news, by its explicit intent to cause harm. Additionally, disinformation differs from malinformation, where truthful information is spread with harmful intentions, by its inherent falsehood.

Fake news is a news article that is intentionally and verifiably false, designed to potentially mislead readers. This may involve distorting facts, taking information out of context, or presenting partial truths to manipulate or mislead the audience[1]. The objective is to construct a false narrative, often sensationalized or biased, with the aim of influencing opinions or actions. Consequently, fake news falls within the broader category of disinformation within the media landscape.

While fake news is typically associated with entirely false information that can be verified, recent research highlights a nuanced perspective. Scholars now recognize that fake news is rarely devoid of truth, and even factual true information can be manipulated to convey a misleading narrative. This acknowledgment underscores the intricate nature of evaluating news authenticity and the potential for manipulation within ostensibly truthful content.

When fake news is intentionally generated and disseminated with a political agenda to influence public opinion or manipulate perceptions, it can be considered a form of *propaganda*. This type of politically motivated fake news can be identified by its persuasive function, the use of faulty reasoning and emotional appeals, and the representation of a specific agenda. Often presented as credible news, such fake news tends to blend truths, falsehoods, and ambiguities without sufficient transparency about its source or underlying motives. Similar to any form of propaganda, news propaganda can be disseminated for various reasons, including governance, political or ideological motives, partisan agendas, religious or ethnic considerations, and commercial or business interests.

Our research focuses on the harmful intent inherent in fabricated content and the news format as pivotal elements in our study. Consequently, certain concepts are excluded from our definition of fake news: (1) satirical news with proper context, which lacks the intent to mislead or deceive and is unlikely to be misconstrued as factual; (2) rumors not originating from news events; (3) conspiracy theories, challenging to verify as true or false and extending beyond the scope of news; and (4) hoaxes driven solely by amusement or to deceive targeted individuals[2].

3.2 Fake News in Social Science

Prior to delving into the computational challenges associated with detecting fake news, it is imperative to grasp the underlying nature of the problem: why is fake news exceptionally effective nowadays in its dissemination and harmful impact on social media? To unravel this, a comprehensive exploration of fake news is warranted, scrutinizing it through psychological, social, and political lenses and examining the patterns and tools inherent in social media that empower the creation and dissemination of fake news.

Psychological Foundations of Fake News.

The discernment of whether news is true extends beyond rational considerations, as it is also shaped by various psychological factors that could be hidden or underestimated by humans: (1) Social credibility: individuals are inclined to view a source as credible if others deem it credible. (2) Confirmation Bias: individuals prefer to receive information that confirms their preexisting beliefs. (3) Frequency heuristic: individuals may naturally trust information they come across frequently, even if it happens to be fake news[1].

Due to these cognitive biases inherent in human nature, fake news can often be perceived as real by people and fake news spreading in social media capitalizes on these biases. First of all, the endorsement or perceived credibility of a source by one individual can significantly impact the perceptions of others, particularly

in social networks where people often trust sources simply because their friends follow them. Secondly, in situations of uncertainty or when credible sources are not readily available, individuals may experience an information vacuum. Social networks, characterized by the rapid exchange of information, fill this gap by circulating news from less reliable accounts. Thirdly, social media algorithms may amplify confirmation bias by showing users content that reinforces their perspectives. Fourthly, on social media platforms, where content is constantly circulated and shared, the frequency heuristic leads individuals to unconsciously trust the information they encounter more often, even if it is misleading or false. Furthermore, correcting fake news is exceptionally challenging once it has taken root. Psychological studies suggest that the efforts of fact-checkers to debunk false information do not fully negate the impact of the initial exposure to the news. Additionally, certain ideological groups might perceive such corrective actions as an attempt to conceal the truth rather than an earnest effort to provide accurate information.

Social Foundations of Fake News.

Social dynamics, encompassing interactions within groups and adherence to social norms, play a crucial role in the spreading of fake news. Two key social concepts contributing to this understanding are (1) Social Identity Theory, which suggests individuals categorize themselves and others based on shared characteristics, and (2) Normative Influence Theory, exploring how individuals conform to gain approval or avoid disapproval[1]. The innate desire for social acceptance leads users "to choose "socially safe" options when consuming and disseminating news information, following the norms established in the community even if the news being shared is fake news"[1]. The power of fake news lies in its ability to manipulate these social tendencies and capitalize on the human need for acceptance and approval within social groups.

Political Foundation of Fake News.

Throughout history, politics has consistently employed propaganda as a means to sway the beliefs of various groups. Notably, the Nazis and later the Communists expanded the arsenal of propaganda tools in the 20th century. Political propaganda, especially in the digital age, employs a diverse array of information types to manipulate people's beliefs, attitudes, and actions online. This includes fake news, rumors, intentionally inaccurate information, unintentionally incorrect information, politically biased content, and "hyperpartisan" news. The impact of propaganda has been further intensified by authoritarian governments engaging in organized "information warfare" domestically and abroad.

Considerable journalistic and scholarly focus has been directed towards Russian attempts to disseminate disinformation and change discord during the 2016 US presidential election. Researchers reveal[4] that the Russian strategy encompassed a blend of covert intelligence operations, including cyber activities, alongside overt initiatives orchestrated by Russian Government agencies, state-funded media outlets such as RT, proxy news sources, third-party intermediaries, and both automated and human-operated bot farms generating deceptive posts and comments in public forums to serve political agendas.

However, the prevalence of fake news is most pronounced within the media landscape of Ukraine, particularly intensifying since the Russian annexation of Crimea in 2014 and experiencing a further surge in 2022 with the full-scale war. Social media platforms witness a proliferation of numerous channels disseminating partial truths and outright fabrications. These outlets frequently propagate misinformation, rumors, conspiracy theories, and launch attacks on mainstream media, concurrently establishing "hyperpartisan" or "local" news channels on social platforms. The propaganda takes various forms, encompassing text, video, audio, fake victims, sham experts, and even non-existent books falsely attributed to well-known authors. Fake news exhibits characteristics of rapid dissemination, continuous circulation, and repetition, lacking a commitment to consistency. Notably, different propaganda channels may not align on themes or messages, and they demonstrate a willingness to change stances without hesitation. Russian propagandists operate without the need for fact-checking or claim verification, enabling them to be highly responsive and agile, often being the first to broadcast "news" about events, thereby shaping initial impressions. Furthermore, they employ a strategy of repetition and recycling, allowing fake claims to resurface after a period and be applied to different events[15].

Christopher Paul and Miriam Matthews[15] emphasized that some of these characteristics directly challenge conventional notions of effective influence and communication advocated by government or defense sources, which typically emphasize the importance of truth, credibility, and consistency while avoiding contradictions. The contemporary strategy of Russia's "information warfare" deliberately seeks to sow confusion, polarize opinions, instill distrust, and construct a fundamentally distorted worldview. These tactics serve Russia's overarching goals, allowing the state to advance its political agenda, exert direct influence on post-Soviet countries, and weaken adversaries, whether on the battlefield in Ukraine or in geopolitical struggles with the USA.

Nevertheless, democratic authorities recognize Russia's engagement in "information warfare," sparking discourse on whether the effort to cleanse the media space of politically motivated fake news might inadvertently compromise freedom of speech and the cherished concept of the "marketplace of ideas"[16] as captured in social media.

Fake News on Social Media.

Social media has removed many of traditional news media's editorial norms and processes that once ensured the accuracy and credibility of information. Despite this shift, social media platforms have inherited a significant level of public trust and credibility[3]. While these technologies offer the potential to expose users to a more diverse range of viewpoints, there is also the risk of unintentionally navigating them toward harmful or extreme content. Primarily, the social media business model relies on capitalizing on attention through advertising[16]. To achieve this, complex filter and recommender algorithms have been developed to precisely target audiences and ensure the repeated visibility of advertisements, often prioritizing them over user preferences. Additionally, these recommender algorithms have given rise to concerns about the formation of "filter bubbles,"

"echo chambers," and "polarization" effects. Such effects contribute to an environment where fake news can attract a substantial audience.

Echo Chamber Effect. Social media introduces a new paradigm for information creation, circulation, and consumption. Users consume news with preselection, where algorithms tailor content based on a viewer's past behaviors, fostering the creation of informational "filter bubbles". Moreover, social media users often gravitate toward groups comprising like-minded individuals, cultivating an "echo chamber" effect. These effects make it harder for people to disprove fake news because of the psychological challenges described above.

Polarization Effect. There is a hypothesis that social media, through the effects of the "filter bubbles" and "echo chambers" reduce users' "tolerance for alternative points of view, amplify attitudinal polarization, boost their likelihood of accepting ideologically compatible news, and increase closure to new information"[3]. Russian propaganda leverages this idea, as evidenced by research on the tactics of the Internet Research Agency[17], a major Russian propaganda agency active in the 2016 US media landscape. The agency strategically fueled radical attitudes and polarization by exploiting pride, fostering feelings of disrespect, and denying the rights of specific communities. This tactic deepened divisions between communities such as the Black American and pro-police communities, Right and Left-leaning factions, Texas, Southern, and Native American cultures, and Muslim and Christian communities, etc. To escalate polarization and foster belief in inaccurate information that fuels negative group attributions, one of the approaches is to craft messages that intensify negative attitudes toward out-groups[4], which was successfully used by the Russian propaganda agency.

However, scholarly consensus on this issue is not explicit. Fletcher and Nielsen [14] have found that social media users are exposed to more diverse news and experience lower polarization compared to non-users. While the impact of fake or extreme news from social networks on traditional news outlets remains unclear, recent cases from the ongoing Israel-HAMAS conflict highlight that traditional media outlets sometimes pick up and disseminate fake news from unverified or terrorist social media accounts when reliable sources are unavailable, often omitting this detail.

Malicious Accounts. On social media, alongside legitimate users, there are malicious actors. The ease of creating accounts facilitates the presence of trolls, automated bots, or human-operated bots. Trolls, actual human users aiming to disrupt online communities, play a significant role in spreading misinformation by provoking emotional responses. Automated bots, controlled by algorithms, can be specifically designed for harm, manipulating and spreading fake news, amplifying the circulation of false information. These bots create a false impression of widespread endorsement, fostering the echo chamber effect in the propagation of fake news. Human-operated bots are "registered by humans as a camouflage and perform activities in social media"[1], executing prescribed agendas on social media.

The common goal of these malicious accounts is to evoke negative emotions like anger and fear, leading to doubt, distrust, and irrational behavior.

4 Fake News Detection

Despite recent tools to track the spread of fake news, the scientific capacity to measure human attention to identified fake news content is still limited. Fake news detection has three main directions: data-oriented, feature-oriented, and model-oriented (presented in Figure 2). The researchers usually experiment with multiple directions at the same time.

In this section, we review current work on the style-based news-content model-oriented direction for fake news on social media detection from a text analysis perspective only. We will not examine the direction of the knowledge-based model-oriented direction as it requires the use of external sources to fact-check proposed claims in news content.

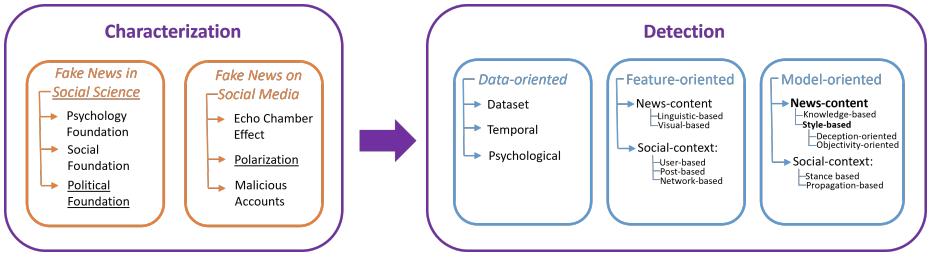

Fig. 2. Fake news on social media: from characterization to detection
Source: Kai Shu et al.[1] Our modifications are underlined.

Style-based news-content model-oriented direction

Fake news creators frequently have malicious intent, aiming to disseminate distorted and misleading information to influence their target groups. To achieve this, they employ distinct writing styles that are crafted to resonate with and persuade their audience while simultaneously mocking true news. Style-based approaches in fake news detection focus on identifying such manipulative techniques embedded in the writing style of news content. The employed methods depend on scrutinizing distinct language attributes and the structure of language, including the number of language elements (e.g., verbs, nouns, sentences, paragraphs), statistical evaluation of language complexity, uncertainty indicators (e.g., quantifiers, generalizations, question marks in the text), subjectivity, non-immediacy (e.g., rhetorical questions or passive voice count), sentiment, diversity, informality, and specificity of the analyzed text.

Binary vs Multi-classification problem

The majority of scholars classified the fake news detection problems as a binary: either the news is true or false[1]. However, fake news in the modern world, taking into account all political fundamentals discussed above, commonly mixes true statements with false claims and is rather "shadows of truth". So, it may make more sense to predict the likelihood of fake news instead of producing a

binary value. Rashkin et al.[6] proposed the approach with a graded notion of truthfulness using a 6-point scale ranging: True, Mostly True, Half-true, Mostly False, False, and Pantson-fire False. All approaches have two major questions to solve: (1) the level of detecting the use of propaganda techniques (source, outlet, article, span level, etc.); (2) transparency of the decision. Da San Martino et al.[7] tried to give their answer to these questions: annotate the propaganda fake news at the fragment level and go deeper into the types of fake news, considering 18 manipulation techniques that do not require supporting information from external resources. Therefore, one news article may consist of different manipulation techniques (span level of detection) and the reader may review them rather than gullibly believe that the article is not reliable (full transparency).

Available Datasets

Collecting online news datasets is feasible from various social media platforms such as Facebook, Twitter, Telegram, etc. However, this process comes with its set of challenges. Certain social media platforms impose restrictions on data scraping, adding complexity to the collection process. The data itself is often noisy and unstructured. Additionally, if a supervised model is chosen, it typically demands annotators with domain expertise for a meticulous analysis of claims, incurring both cost and time.

Figure 3 provides an overview of available corpora designed for supervised models, each with its own distinctions. Notably, these corpora differ in the number of articles they encompass, with QProp being the largest and PTC, collected by Da San Martino et al.[7] for granularity manipulation detection, having the lowest volume but including over 7,000 propagandist snippets. Further distinctions include the labeling methods employed, with TSHP-17 released by Rashkin et al.[6] and QProp using distant supervision, associating a document with a class based on the outlet label that published it, while PTC relies on professional annotators. Annotations also differ in their granularity, with TSHP-17 and QProp operating at the document level, and PTC at the fragment level. Moreover, these corpora vary in classification classes and balance, with QProp being unbalanced for binary classification, TSHP-17 being balanced with four classes (trusted, satire, hoax, and propaganda), and PTC being unbalanced with 18 classes for manipulation detection. All corpora include news from traditional outlets and are in English. Currently, there is no universally accepted benchmark dataset for fake news detection, as different corpora suit different tasks based on their specific features.

Corpus	Level	Sources	Classes	Articles	Fake news
QProp	document	104 (10)	2	51,294	5,737
TSHP-17	document	11 (2)	4	22,580	5,330
PTC	text span	49 (13)	18	451	7,385

Fig. 3. Textual datasets available to train supervised propaganda identification models at different granularity levels.

Text Classification with supervised models

Barrón-Cedeño et al.[11] introduced a binary model classification (propaganda vs. non-propaganda) to automatically evaluate an article's degree of propagandistic content. Their approach considered various representations, ranging from the writing style and readability level to the presence of specific keywords. They experimented on TSHP-17 and QProp corpora which have annotation on the article level. The experimentation involved the TSHP-17 and QProp corpora, both annotated at the article level. Results from their experiments demonstrated that representations emphasizing writing style and text complexity outperformed traditional word n-grams, which primarily focus on topics, in effectively identifying propaganda.

Rashkin et al.[6] tackled a text classification task with four classes— propaganda, trusted, hoax, and satire—utilizing the TSHP-17 dataset and factuality judgments on a 6-point scale. Their approach involved a linguistic analysis employing stylistic lexicons. An interesting observation from their study revealed that first-person and second-person pronouns were more prevalent in less reliable or deceptive news, in addition to that, words facilitating exaggeration, such as subjective terms, superlatives, and modal adverbs, were more commonly employed in fake news. Their findings indicated that trusted sources tended to use assertive language and avoided hedging words, suggesting a clearer and less ambiguous description of events. The researchers trained various models, including Maximum Entropy (MaxEnt), Naive Bayes, and the LSTM model, commonly used for text categorization at that time. The LSTM model demonstrated the most promising results in their study.

Da San Martino et al.[7] defined two tasks, based on annotations from the PTC dataset: (i) binary classification given a sentence in an article, predict whether any of the 18 techniques have been used in it (SLC - Sentence-level Classification); (ii) multi-label multi-class classification and span detection task —given a raw text, identify both the specific text fragments where a manipulation technique is being used as well as the type of technique (FLC - Fragment-level classification). Since manipulation is conveyed through the use of a number of techniques, such detection allows for deeper analysis at the paragraph and the sentence level that goes beyond a single document-level judgment described above.

In the table 4 is the list of 18 manipulation techniques used for classification. It only includes techniques that can be found in journalistic articles and can be judged intrinsically, without the need to retrieve supporting information from external resources.

Baseline models in the study utilized BERT for two tasks with two modifications. Additionally, the authors introduced a multi-granularity deep neural network designed to enhance the signal from the sentence-level task, thereby improving the performance of the fragment-level classifier. The best-performing models for both tasks used BERT-based contextual representations.

Other approaches to the same task used contextual representations based on fine-tuned BERT, FastText, RoBERTa, Grover, and ELMo, CNN, LSTM-

Table 1. List of the 18 propaganda techniques and their definitions.

Technique	Definition
Name calling	attack an object/subject of the propaganda with an insulting label
Repetition	repeat the same message over and over
Slogans	use a brief and memorable phrase
Appeal to fear	support an idea by instilling fear against other alternatives
Doupt	questioning the credibility of someone/something
Exaggeration/mimization	exaggerate or minimize something
Flag-Waving	appeal to patriotism or identity
Loaded language	appeal to emotions or stereotypes
Reduction ad hitlerum	disapprove an idea suggesting it is popular with groups hated by the audience
Bandwagon	appeal to the popularity of an idea
Casual oversimplification	assume a simple cause for a complex event
Obfuscation, intentional vagueness	use deliberately unclear and obscure expression to confuse the audience
Appeal to authority	use authority's support as evidence
Black and white fallacy	present only two options among many
Thought terminating cliches	phrases that discourage critical thought and meaningful discussions

CRF, a Transformer or context-independent representations based on lexical, sentiment-based, readability, and TF-IDF features or using ensembles[8].

Text Classification with weak social supervision

Another method for classifying fake news text involves weak social supervision[13], where weak labels are derived from users' engagement with the news. For instance, Shuo Yang al. et.[19] extracted users' opinions on the authenticity of news by utilizing auxiliary information from users' interactions with news tweets on social media. They aggregated these opinions in a carefully designed unsupervised manner to generate binary estimates of fake news versus true news. They used a Bayesian network model to capture the conditional dependencies among the truthfulness of news, the users' opinions, and the users' credibility.

However, weak social supervision labels may not always be well-structured, complete, or reliable. Despite this, they offer insights into understanding and detecting fake news with early-stage explainability, eliminating the need for costly and time-consuming expert-labeled corpora.

Text Classification with semi-supervised models

Detecting linguistic fake news with any supervision is a challenging task because of the nature of the text: it involves complex nuances, subtle linguistic patterns, and the potential for evolving tactics employed by purveyors of misinformation. Additionally, the success of deep multilanguage generative models introduces a new dimension to the problem: how to identify machine-generated fake news that is fluent, readable, and appealing to specific target audiences at scale. Developing robust verification techniques against generators becomes imperative. Experiments with Grover and GPT2[20] demonstrated their capability not only to produce realistic and controlled text generations but also to distinguish machine-written articles from human-written ones.

State-of-the-art Model

BERT (Bidirectional Encoder Representations from Transformers)[9], a transformer- based NLP model, has consistently delivered state-of-the-art results for various NLP tasks, outperforming other language models.

Evaluation Metrics

Most existing approaches consider the fake news problem as a classification task, predicting whether a news article is fake or not:

• True Positive (TP): Predicted fake news pieces that are actually annotated as fake news.

• True Negative (TN): Predicted true news pieces that are actually annotated as true news.

• False Negative (FN): Predicted true news pieces that are actually annotated as fake news.

• False Positive (FP): Predicted fake news pieces that are actually annotated as true news.

The Receiver Operating Characteristics (ROC) curve provides a way of comparing the performance of classifiers by looking at the trade-off between the False Positive Rate (FPR) and the True Positive Rate (TPR).

Based on the ROC curve, the Area Under the Curve (AUC) value measures the overall performance of how likely the classifier is to rank the fake news higher than any true news.

An alternative strategy is to devise specialized metrics tailored to the specific task and methodology.

5 Research Gaps

Among the various methods for style-based fake news detection like [6][7], a promising avenue is the identification of fine-grained propaganda techniques. This approach aligns with the contemporary fake news strategy of blending true and false statements or manipulating attitudes through writing style. Besides, manipulation techniques, well-documented in the literature, serve as the foundation for fake news, making them a logical focus for the research project. However, this method has certain limitations:

(1) The PTC corpora, which contains the ground true dataset labeled with 18 manipulation techniques, is biased toward the US culture and political landscape.

(2) Corpora is composed of the news from traditional outlets, and the models based on the PTC corpora were not evaluated on social media content (to the best of our knowledge), which is unstructured, contains mistakes, new words, slang, dialect, or a mix of different languages. Additional efforts may be required to address these challenges.

(3) The task of identifying 18 fine-grained propaganda techniques has been tackled in English and Arabic but not in Ukrainian and Russian. The multilingual content, hidden patterns, and distinct features of Ukrainian and Russian texts necessitate additional efforts. The recent 2023 Disinformation Detection Challenge by AI House focuses solely on the binary problem of Sentence-level

Classification for the Ukrainian language, leaving Fragment-level Classification as an open task.

(4) Although detecting manipulation techniques at the fragment level provides explainability and transparency in decision-making, it falls short of explaining the intent aspect of fake news. Considering the Russian "information warfare" against Ukraine, understanding the aim of manipulation techniques and their targets can be valuable features for manipulation detection models.

To fill these gaps, we plan to conduct a thorough study on the use of the PTC corpus to identify manipulative techniques in Telegram news in Ukrainian and Russian, taking into account the unique language features. Lastly, integrating an understanding of the intent aspect of fake news, particularly in the context of Russian "information warfare" against Ukraine, will be a focal point in enhancing the comprehensiveness of our manipulation detection model.

6 Conclusion

In this work, we presented a comprehensive review of the fake news identification challenge, delving into various disciplines such as Psychology, Sociology, Politics, and Computational Linguistics. We provided an overview of the existing computational style-based fake news detection research, defined their critical limitations and identified gaps. The motivation for tackling this challenge was explained, and we outlined our plan to fix identified gaps.

References

1. Shu, K., Sliva, A.L., Wang, S., Tang, J., Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ArXiv, abs/1708.01967.
2. Shu, K., Bhattacharjee, A., Alatawi, F.H., Nazer, T.H., Ding, K., Karami, M., Liu, H. (2020). Combating disinformation in a social media age. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10.
3. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D.M., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J. (2018). The science of fake news. Science, 359, 1094 - 1096.
4. Tucker, J.A., Guess, A.M., Barberá, P., Vaccari, C., Siegel, A.A., Sanovich, S., Stukal, D.K., Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.
5. Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Pietro, R.D., Nakov, P. (2020). A Survey on Computational Propaganda Detection. ArXiv, abs/2007.08024.
6. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. Conference on Empirical Methods in Natural Language Processing.
7. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P. (2019). Fine-Grained Analysis of Propaganda in News Article. Conference on Empirical Methods in Natural Language Processing.
8. Da San Martino, G., Barrón-Cedeño, A., Nakov, P. (2019). Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. ArXiv, abs/1910.09982.

9. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.
10. Yoosuf, S., Yang, Y. (2019). Fine-Grained Propaganda Detection with Fine-Tuned BERT. Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda.
11. Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. Inf. Process. Manag., 56, 1849-1864.
12. Barrón-Cedeño, A., Da San Martino, G., Jaradat, I., Nakov, P. (2019). Proppy: A System to Unmask Propaganda in Online News. AAAI Conference on Artificial Intelligence.
13. Shu, K., Wang, S., Lee, D., Liu, H. (2020). Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. ArXiv, abs/2001.00623.
14. Fletcher, R.J., Nielsen, R.K. (2017). Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. Journal of Communication, 67, 476-498.
15. Paul, C., Matthews, M. (2016). The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It.
16. Bakshy, E., Messing, S., Adamic, L.A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348, 1130 - 1132.
17. DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R.C., Fox, R., Albright, J., Johnson, B. (2018). The tactics tropes of the Internet Research Agency.
18. Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., Han, J. (2020). Weakly-Supervised Text Classification Using Label Names Only. Conference on Empirical Methods in Natural Language Processing.
19. Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., Liu, H. (2019). Unsupervised Fake News Detection on Social Media: A Generative Approach. AAAI Conference on Artificial Intelligence.
20. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y. (2019). Defending Against Neural Fake News. ArXiv, abs/1905.12616.
21. Gupta, P., Saxena, K., Yaseen, U., Runkler, T.A., Schütze, H. (2019). Neural Architectures for Fine-Grained Propaganda Detection in News. ArXiv, abs/1909.06162.
22. Bonet-Jover, A., Sepúlveda-Torres, R., Boró, E.S., Martínez-Barco, P. (2023). A semi-automatic annotation methodology that combines Summarization and Human-In-The-Loop to create disinformation detection resources. Knowl. Based Syst., 275, 110723.
23. Thorne, J., Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. ArXiv, abs/1806.07687.
24. Hardalov, M., Arora, A., Nakov, P., Augenstein, I. (2021). A Survey on Stance Detection for Mis- and Disinformation Identification. ArXiv, abs/2103.00242.
25. Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D.I., Da San Martino, G., Shaar, S., Firooz, H., Nakov, P. (2021). A Survey on Multimodal Disinformation Detection. International Conference on Computational Linguistics.
26. Jaiswal, J., LoSchiavo, C.E., Perlman, D.C., Perlman, D.C. (2020). Disinformation, Misinformation and Inequality-Driven Mistrust in the Time of COVID-19: Lessons Unlearned from AIDS Denialism. AIDS and Behavior, 24, 2776 - 2780.
27. Da San Martino, G., Barr'on-Cedeno, A., Wachsmuth, H., Petrov, R., Nakov, P. (2020). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. International Workshop on Semantic Evaluation.

28. Madabushi, H.T., Kochkina, E., Castelle, M. (2020). Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data. ArXiv, abs/2003.11563.
29. Elswah, M., Howard, P.N. (2020). "Anything that Causes Chaos": The Organizational Behavior of Russia Today (RT). Journal of Communication.
30. Nouh, M., Nurse, J.R., Goldsmith, M. (2019). Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter. 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), 98-103.
31. Lilly, B., Cheravitch, J. (2020). The Past, Present, and Future of Russia's Cyber Strategy and Forces. 2020 12th International Conference on Cyber Conflict (CyCon), 1300, 129-155.
32. Ventsel, A., Hansson, S., Madisson, M., Sazonov, V.Y. (2019). Discourse of fear in strategic narratives: The case of Russia's Zapad war games. Media, War Conflict, 14, 21 - 39.
33. Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A., Nakov, P. (2019). Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Conference and Labs of the Evaluation Forum.
34. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep Contextualized Word Representations. ArXiv, abs/1802.05365.
35. Baly, R., Karadzhov, G., Saleh, A., Glass, J.R., Nakov, P. (2019). Multi-Task Ordinal Regression for Jointly Predicting the Trustworthiness and the Leading Political Ideology of News Media. North American Chapter of the Association for Computational Linguistics.
36. Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., Adams, K. (2020). Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation. Social Science Computer Review, 40, 560 - 578.

Technique	Definition
Name calling	attack an object/subject of the propaganda with an insulting label
Repetition	repeat the same message over and over
Slogans	use a brief and memorable phrase
Appeal to fear	support an idea by instilling fear against other alternatives
Doubt	questioning the credibility of someone/something
Exaggeration/minimizat.	exaggerate or minimize something
Flag-Waving	appeal to patriotism or identity
Loaded Language	appeal to emotions or stereotypes
Reduction ad hitlerum	disapprove an idea suggesting it is popular with groups hated by the audience
Bandwagon	appeal to the popularity of an idea
Casual oversimplification	assume a simple cause for a complex event
Obfuscation, intentional vagueness	use deliberately unclear and obscure expressions to confuse the audience
Appeal to authority	use authority's support as evidence
Black&white fallacy	present only two options among many
Thought terminating clichés	phrases that discourage critical thought and meaningful discussions
Red herring	introduce irrelevant material to distract
Straw men	refute argument that was not presented
Whataboutism	charging an opponent with hypocrisy

Fig. 4. List of the 18 propaganda techniques and their definitions.