Multimedia Tools and Applications (2024) 83:51009–51067
https://doi.org/10.1007/s11042-023-17470-8

A comprehensive survey on machine learning approaches for fake news detection

Jawaher Alghamdi[1,2] · **Suhuai Luo**[1] · **Yuqing Lin**[1]

Received: 24 September 2022 / Revised: 16 August 2023 / Accepted: 3 October 2023 /
Published online: 9 November 2023

Abstract
The proliferation of fake news on social media platforms poses significant challenges to society and individuals, leading to negative impacts. As the tactics employed by purveyors of fake news continue to evolve, there is an urgent need for automatic fake news detection (FND) to mitigate its adverse social consequences. Machine learning (ML) and deep learning (DL) techniques have emerged as promising approaches for characterising and identifying fake news content. This paper presents an extensive review of previous studies aiming to understand and combat the dissemination of fake news. The review begins by exploring the definitions of fake news proposed in the literature and delves into related terms and psychological and scientific theories that shed light on why people believe and disseminate fake news. Subsequently, advanced ML and DL techniques for FND are dicussed in detail, focusing on three main feature categories: content-based, context-based, and hybrid-based features. Additionally, the review summarises the characteristics of fake news, commonly used datasets, and the methodologies employed in existing studies. Furthermore, the review identifies the challenges current FND studies encounter and highlights areas that require further investigation in future research. By offering a comprehensive overview of the field, this survey aims to serve as a guide for researchers working on FND, providing valuable insights for developing effective FND mechanisms in the era of technological advancements.

Keywords Fake news · Fake news detection · Misinformation

✉ Jawaher Alghamdi
jawaher.alghamdi@uon.edu.au

Suhuai Luo
suhuai.luo@newcastle.edu.au

Yuqing Lin
yuqing.lin@newcastle.edu.au

[1] School of Information and Physical Sciences, The University of Newcastle, Newcastle, Australia

[2] Department of Computer Science, King Khalid University, Abha, Saudi Arabia

1 Introduction

Traditionally, people fundamentally consume news or information from newspapers and TV channels; however, with the advent of the Internet and its intrusion into our lifestyle, the former has become less prominent [31]. Today, online social networks (OSNs) and lifestreaming platforms play a fundamental role compared to television as one of the major news sources, where in 2016, 62% of U.S. people gained news from social media, while 49% of U.S. people recorded watching news through social media in 2012 [209]. Figure 1 illustrates the level of interest in the term 'fake news' over the span of the last decade, as extracted from Google Trends [257].

Recently, the role of OSNs has significantly increased due to their convenient access, as it is no longer limited to being a window for communication between individuals, but rather it has become an important tool to exchange information and for influencing and shaping public opinion where individuals can release data in all its forms on various OSNs. Nevertheless, unfortunately, the other side of the coin is fake news dissemination, specifically on OSNs, which poses a great concern to the individual and society due to the lack of control, supervision and automatic fact-checking, leading to low-quality and fake content generation. As such, users are prone to countless disinformation and misinformation on OSNs, including fake news, i.e., news stories with intentionally false information [12, 235]. As a major source for misinformation spreaders, OSNs were developed primarily for connecting individuals who exploit the connectivity and globality of such networks [134]. According to [60], on OSNs, fake news spreads six times faster compared to true news resulting in fear, panic, and financial loss to society [299]. It is not so surprising to see such falsehood information disseminated rapidly as OSNs and the Internet, in general, give people some degree of anonymity that those fake news spreaders can harness to achieve their intent. This, in turn, is bound to result in worse and more severe consequences if not combated. For example, fake news has led to significant impacts on real-world events where a piece of fake news from Reddit causes a real shooting [247].

By the end of the 2016 U.S. presidential election, for instance, over 1 million tweets were found to be related to that piece of fake news known as PIZZAGATE[1]. Furthermore, during this period, top-20 fake news pieces were reported to be larger than the top-20 most-discussed true stories[2]. Research on fake news velocity states that tweets including falsified information on Twitter reach people six times faster than trustworthy tweets [138]. This, in fact, indicates how terribly fake news disseminates and how it can have adverse social effects. The matter of concern is the quick reactions, such as retweets, likes, and shares of a tweet (fake news story) received on Twitter without pre-thinking, aggravating the problem even more. According to [268], false news, particularly political news, on Twitter is usually retweeted by more users and spreads extremely rapidly. In fact, the major problem is that some popular sources for information that are considered to be genuine, such as Wikipedia, are also prone to fake news or false information [132]. According to a report in China, fake information constitutes more than one-third of trending events on microblogs [291]. Therefore, without news verification, fake news would spread rapidly through OSNs resulting in real consequences [73]. Fake news is not a new phenomenon [119] (e.g., the New York Sun published in 1835 a series of articles known as the Great Moon Hoax, described the discovery of life on the moon [12]); however, enquiries of why such phenomenon has attracted more attention and emerged as a hot topic of interest are specifically relevant nowadays [296]. The primary cause is that

[1] https://tinyurl.com/z38z5zh

[2] https://tinyurl.com/y8dckwhr

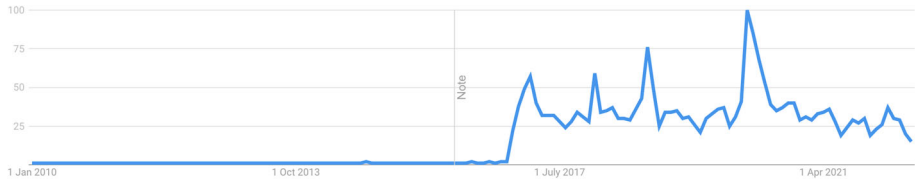

Fig. 1 Fake news trends (2010-2022) [257])

fake news online, as opposed to traditional news media (e.g., newspapers and TV), can be created and published faster and cheaper [235], which destroys the authenticity balance of the entire news ecosystem. The popularity and growth of OSNs also contribute to this raising of interest [181, 286, 296].

The OSNs have contributed to the fostering of false news that intentionally false information transmitted to the public for various purposes, including political or financial manipulation [196]. Furthermore, substantial potential political and economic benefits can be earned from the excessive activity around these online social media platforms, which often motivate spiteful entities to create and spread false information [297]. Indeed, the ulterior motive behind these (intentional) fake news creators is not necessarily "closely connected to the content of the claims they are manufacturing" [80, p. 107], where "the deception lies not in getting an audience to believe a false claim, but in getting them to believe it is worth sharing" [80, p. 108]. For example, during the U.S. presidential election, the dozen of "well-known" teenagers in the Macedonian town of Veles has become wealthy as a result of creating fake news for millions on social media, where, according to NBC, each user "has earned at least $60,000 in the past six months - far outstripping their parents' income and transforming his prospects in a town where the average annual wage is $4,800" [297]. What this example shows is that news is not necessarily created and published to manipulate others, so the intent may be for a malicious goal, such as intentionally defaming or instilling malicious goals and beliefs in people, since "purveyors of fake news deliberately engage in practices that they know, or can reasonably foresee, to lead to the likely formation of false beliefs on the part of their audience" [80, p. 107], or it may be for financial goals on purpose, but without the presence of the malicious intent [80]. Subsequently, as the valence effect theory indicates, individuals tend to overestimate the benefits of spreading fake news instead of its costs [117], and thus this might be one reason for fake news's widespread alarmingly. Therefore, the rise of falsified information (i.e., disinformation websites) is attributed to two main reasons: (i) the promotion of viral news articles generates significant advertising revenue, and (ii) it is usually the goal of false news providers to influence the public opinion on particular topics [12]. Another major factor contributing to the spread of misinformation is the presence of malicious agents such as bots and trolls [131, 229] (see Section 2.4.1 for more information on this). Fake news widespread leads to severe threats across society and makes it challenging to discover [13, 17]. This severely damaged society, including the economy, politics, health and peace. Regarding the economy, the negative implications of fake news widespread affected economies (i.e., stock markets), causing financial loss. One example is the false bankruptcy story about UAL's parent company in 2008 which led to a 76% drop in stock price [36]. In addition, an Associated Press account has released a fake tweet claiming that President Barack Obama was injured in an explosion which in turn caused stocks to drop immediately

where the Dow plunged over 140 points, and the estimated loss of market cap in S&P 500 was 136.5 billion dollars [158].

Extensive research has demonstrated that fake political news exhibits greater speed, reach, and popularity compared to fake news in domains such as terrorism, business, science, and entertainment [269]. This phenomenon was notably exemplified during the 2016 U.S. presidential election, where an overwhelming amount of fake news favouring either Hillary Clinton or Donald Trump proliferated on Facebook, amassing over 37 million shares in a relatively short three-month period before the election. Notably, the top twenty frequently discussed fake election news stories on OSNs, particularly Facebook, received a staggering 8,711,000 comments, reactions, and shares, surpassing the combined engagement of the top twenty most-discussed election stories posted by 19 major news websites [240]. Furthermore, a fake news story endorsing Donald Trump by Pope Francis garnered millions of views, shares, and likes, perpetuating its deceptive narrative [59].

The deleterious impact of fake news dissemination extends beyond politics, encompassing health-related consequences of grave magnitude. Tragic incidents have occurred where online advertisements for experimental cancer treatments, mistakenly perceived as reliable medical information, have resulted in the untimely death of cancer patients [174]. Likewise, false or misleading claims regarding the COVID-19 virus have threatened public health, as individuals are swayed to take risks by consuming harmful substances or disregarding social distancing guidelines. In recent years, the COVID-19 pandemic has witnessed the proliferation of fake news that presents attention-grabbing content, deceiving individuals into believing its purported usefulness. Shockingly, within two months, the International Fact-Checking Network (IFCN) uncovered over 3,500 false claims related to COVID-19 [201]. For instance, disseminating fake news suggesting unproven remedies or attributing the virus to 5G towers has resulted in physical harm [176]. Tragically, it is estimated that at least 800 individuals worldwide may have lost their lives during the first three months of 2020 due to coronavirus-related false claims [175]. Consequently, disseminating fake news poses a significant threat to both individuals and society, with OSNs amplifying this peril. The erosion of social confidence, credibility, and integrity within the news system, coupled with political polarization, has contributed to the rampant prevalence of fake news on OSNs [298]. The inherent structure of these networks facilitates the rapid propagation of fake news, rendering OSNs increasingly popular platforms for its dissemination. In fact, research by Gartner [34] predicts that by 2022, individuals in mature economies will consume more false information than true information, underscoring the urgency for automated FND methods to combat the exponential rise in false news. Researchers specializing in NLP have dedicated their efforts to developing a diverse array of ML and DL algorithms for detecting fake news (for further details, refer to Section 3). In this survey, we extensively review previous studies to comprehensively understand fake news dissemination to devise strategies to mitigate its impact.

The primary contributions of this paper are as follows:

- In-depth review and exploration of fake news definitions, related terms, and their manifestation in both traditional and modern online media. By delving into the underlying reasons why individuals tend to believe and disseminate fake news, we aim to enhance our comprehension of this phenomenon.
- Comprehensive survey encompassing a wide range of feature-based methods, diverse ML and DL techniques, and state-of-the-art transformer-based models employed in FND. This survey offers valuable insights into the effectiveness of these approaches and equips researchers with the necessary knowledge to navigate the evolving landscape of FND.

- Discussion of the challenges that must be addressed to effectively curb the dissemination of fake news. By highlighting these challenges, we seek to inspire further research and innovation in the field, fostering a community of dedicated scholars committed to mitigating the adverse effects of fake news dissemination.

This paper is structured as follows. Table 9 lists all the acronyms used in this paper. Section 2 reviews some preliminaries on the topic. Section 3 presents a comprehensive review of FND approaches. In Section 4, we review the commonly used methods, and in Section 5, we shed light on the current challenges. The limitations and the recommended potential directions for future research have been introduced in Sections 6 and 7, respectively. Finally, Section 8 concludes the paper. Figure 2 shows the outline of this paper.

2 Preliminaries

2.1 Fake news definition

The fake news epidemic, as a growing issue affecting the world, had existed from the time when news started to spread widely after the invention of the printing press in 1439 [157]. Efforts are dedicated primarily to identifying and detecting fake news from users' social media content. However, despite the dedicated researchers' efforts, *fake news* term (i.e., refers to a variety of terms in the literature, including (mis)disinformation, rumour, hoax, etc.) is still vague. The leading cause of such ambiguity might be the existence of related terms (discussed in detail in Section 2.2). As Axel Gelfert [80] stated, the plethora of (tentative) definitions that have been proposed have led some to worry that the term fake news, as a result of its heterogeneity, may become "a catch-all term with multiple definitions" [145, p. 1]. Facebook produced a whitepaper in 2017 that discussed potential threats to online communication and the responsibility of being one of the most popular OSNs today [276]. When exploring the growing issue of utilising the vague term fake news, they stated that "the overuse and misuse of the term "fake news" can be problematic because, without common definitions, we cannot understand or fully address these issues" [276, p. 4]. Fake news typically mimics trustworthy news in order to gain credibility where such falsified content derives its value by mimicking trustworthy content; as Axel Gelfert [80] puts it, "fakes derive their value entirely from the originals they successfully mimic, specifically from the scarcity of the latter". We acknowledge that the definition of fake news is a highly debated topic. Before listing the definitions of fake news term proposed in the literature (in their context), let us remind the reader that there is an explicit definition of fake news as "the online publication of

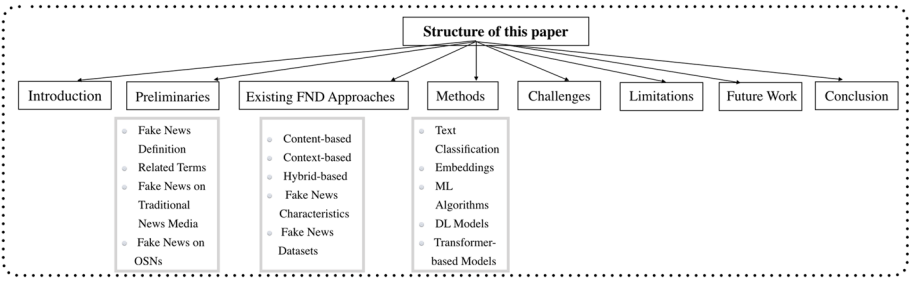

Fig. 2 Structure of the paper

intentionally or knowingly false statements of fact" [125, p. 6], since "it is widely circulated online" [19, p. 1], which, indeed, justifies the existence of "the recognition that the medium of the internet (and social media, in particular) has been especially conducive to the creation and proliferation of fake news" [80, p. 96]. Hence, why has fake news become such a powerful force in the online world? Regina Rini stated:

"A fake news story is one that purports to describe events in the real world, typically by mimicking the conventions of traditional media reportage, yet is known by its creators to be significantly false and is transmitted with the two goals of being widely re-transmitted and of deceiving at least some of its audience" [212, p. E45].

Similarly, Darren Lilleker [145, p. 2], a professor in political communication, argues that "fake news is the deliberate spread of misinformation, be it via traditional news media or through social media". Roger Plothow argues, "Fake news should be defined as a story invented entirely from thin air to entertain or mislead on purpose" [198, p. A5], which is then echoed by economists Hunt Allcott and Matthew Gentzkow as "news stories that have no factual basis but are presented as news" [12, p. 5][3].

A definition by [12], which was then adopted by most of the existing studies, including [48, 125, 170, 200], defined fake news as "news article that is intentionally and verifiably false and could mislead readers". This leaves out rumours, conspiracy theories, unintentional reporting of mistakes, and misleading reports, but of course, not necessarily false, while including intentionally fabricated pieces and satire sites [12, 125]. However, this conceptualisation excludes mainstream media misreporting from scrutiny [167]. Considering *authenticity* and *intention* as two key factors, a 2017 survey paper introduces a similar definition [235] of fake news as a news article that is intentionally and verifiably false. Based on those two key factors, fake news includes verified false information and can be created to intentionally mislead readers. This compound term can also be defined as [230] "a news article or message published and propagated through media, carrying false information regardless of the means and motives behind it," which, according to [296], overlaps with false news, misinformation [129], disinformation [128], satire news [218], or even the improper stories [83]. Defining fake news as false or inaccurate news is not fruitful since this does not exclude the occasional errors that occur in the reports from being fake news. Axel Gelfert argued that "being likely to mislead its target audience by bringing about false beliefs in them—are not yet sufficient to demarcate fake news from, say, merely accidentally false reports." Gelfert [80, p. 105], indicating that such merely accidentally false reports should not be considered fake news. This is because these "reports do mislead their audiences by instilling false beliefs in them, but they do so as the result of an unforeseen defect in the usually reliable process of news production." Gelfert [80, p. 105] while "fake news, by contrast, is misleading its target audience in a non-accidental way" [80, p. 105], and even if a putative report is misleading in a non-accidental way, it must also be deliberately in order to be counted as fake news [80]. Vian Bakir and Andrew McStay proposed to define fake news "as either wholly false or containing deliberately misleading elements incorporated within its content or context" [19, p. 1].

A widely adopted definition belongs to Cohen et al. [46], where they define fake news as everything ranging from malignant news to political propaganda. Another definition by Lazer et al. [140] fake news is "fabricated information that mimics news media content in form but not in organizational process or intent". Furthermore, the term "fake news" is defined by the Collins English dictionary [58] as "false, often sensational, information disseminated under the guise of news reporting". A definition that captures most of the fake news distinctive

[3] For more discussions about these definitions, we refer the reader to [80]

features[4] is proposed by Axel Gelfert: "(FN) Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design" [80, p. 108]. Thus far, fake news refers to any kind of content whose main purpose is to deliberately deceive and mislead the readers (by instilling false beliefs in them). Nevertheless, there has been a recent point of discussion about the definition, perception and conceptualization of fake news term [50, 235].

2.2 Related terms

Journalists and others have urged us to "stop calling everything 'fake news"' [182]. Thus, it becomes necessary to introduce related concepts and definitions to differentiate fake news from related terms such as misinformation, rumours, spam, etc. Some related terms often co-occur or overlap with fake news term reported in the literature, namely, false news [268], deceptive news [12, 139, 235], disinformation [128], satire news [119], misinformation [129, 275], clickbait [42], rumor [300] and others. A 2020 survey [296] reported that these terms could be differentiated based on three characteristics: (i) *authenticity*, which emphasizes the falsity of the information; (ii) *intention*, which emphasizes the intention to mislead readers; and (iii) *whether the information is news*. Propaganda, conspiracy theories, hoaxes, biased or one-sided stories, clickbait, and satire news, are other types of potentially false information that we can find on OSNs that contribute to information pollution [162].

However, whether satirical publications should be considered in the definition of fake news sparked controversy among many researchers where some scholars point out that satire should be excluded from the fake news definition since it is "unlikely to be misconstrued as factual" and not necessarily to inform audiences [12, p. 214], while others disagree in that it should be rather included in the definition of fake news as it could be misconstrued as telling the truth, even though it is legally protected speech [125]. For instance, an apology was issued by a satirical site run by hoaxer Christopher Blair, in 2017, for making a story "too real" since many were unable to detect its satirical nature [75, 167]. According to [289], false information can be proliferated by bots, activist or political organizations, governments, journalists, criminal/terrorist organizations, conspiracy theorists, hidden paid posters, state-sponsored trolls, and individuals that benefit from false information. There are many ways to motivate those actors, either to manipulate public opinion, to create disorder and confusion, to hurt or disrepute, to obtain financial gain by increasing site views, to promote ideological biases, or even to entertain individuals [231]. Several topics related to FND have been studied in the existing literature. This is including misinformation [109, 203, 290], rumors detection [152, 224], and spammer detection [105, 142, 159]. Following previous research [279], a series of key terms (i.e., the concept of misinformation and a list of subconcepts) related to fake news is adopted. The following definitions are provided. *Misinformation* "defined as false, mistaken, or misleading information" that is unintentionally spread due to honest reporting mistakes or incorrect interpretations [66, 98] while *Disinformation* can be understood as "the distribution, assertion, or dissemination of false, mistaken, or misleading information in an intentional, deliberate, or purposeful effort to mislead, deceive, or confuse" [70, p. 231] or promote biased agenda [265]. The differences between misinformation and disinformation have been identified conceptually in [246]. Let us remind the reader that while misinformation and disinformation terms are both referring to false (incorrect) information, they differ in terms of the intention characteristic, where misinformation is spread without the

[4] "features inherent in the design of the sources and channels through which fake news proliferates that imbue it with its novel significance" [80, p. 109].

intent to deceive, whereas disinformation is spread with the intention to deceive [67, 98, 133]. Depending on the intent of the source, rumours can fall into either of these two types [misinformation and disinformation], given that rumours are not necessarily false but may turn out to be true [300]. A *Rumor* is a story that carries truth that is sort of unverified or doubtful, circulating from one person to another. This term has been used to overlap with the term fake news and other terms of disinformation recently. Indeed, media scholars and some social epistemologists have long been concerned with demarcating such phenomena as gossip, rumour, hoaxes, and urban legends [79].

Thereupon, gossip "possesses relevance only for a specific group" and "is disseminated in a highly selective manner within a fixed social network", while rumours are "unauthorized messages that are always of universal interest and accordingly are disseminated diffusely" [27, p. 70]. On the other hand, hoaxes are "deliberately fabricated falsehoods that masquerade as the truth" and, different from fake news, "serve quite different purposes and typically intended to be found out eventually" [80]. A *Spam* is defined as unwanted messages containing irrelevant or inappropriate information sent to a large number of recipients. We remind the reader that the terms fake news and rumours, specifically, are often used interchangeably in the literature. Readers are referred to [163] for a difference between these two, where fake news was defined as false information spread through the Internet or news outlets to intentionally gain political or financial benefits, and rumours were defined as an unverified piece of information that can be true or false; when this information is false, it can be considered fake news. The conceptual differences and similarities between these terms and many other terms associated with "fake news" have been provided by previous research; we refer interested readers to [167] for more information.

2.3 Fake news on traditional news media

Traditionally, fake news, as a growing issue affecting the world, exist and is typically disseminated via traditional media ecology over time, such as newspaper and television. Nevertheless, unlike traditional newspapers, fake news on OSNs is terribly prevalent. As Alvin Goldman stated on the advantages of traditional newspapers over online blogs:

"Newspapers employ fact checkers to vet a reporter's article before it is published. They often require more than a single source before publishing an article and limit reporters' reliance on anonymous sources. These practices seem likely to raise the veristic quality of the reports newspapers publish and hence the veristic quality of their readers' resultant beliefs" [87, p. 117].

Fake news is spreading quickly due to the upbringing and development of the information environment such as OSNs, as we no longer rely on traditional news environments where people cannot express their opinions and easily share falsified information with others. Today, to create and circulate content online, it is not necessitous to be a journalist and work for a publication [167], where individuals can participate, share and react with others freely on OSNs leading to exacerbating the problem of fake news dissemination that have devastating effects on society. Furthermore, studies show that they may even be preferred over traditional professional sources [248]. This is specifically problematic given that individuals find information that agrees and matches with prior beliefs as more credible and reliable because credible information appears together with personal opinions, creating an environment that aggravates misinformation [30]. This subsection discusses several psychological and social science theories describing why people believe fake news, why they participate in spreading it, and the negative influence of fake news on individuals and society.

2.3.1 Psychological theories

In reality, fake news has the power to influence people (whom we often refer to as vulnerable users, those who were involved in fake news dissemination without recognizing the falsehood). By exchanging uninformed knowledge over networks, vulnerable users are considered major contributors to the dissemination of such knowledge. On digital platforms, different actions can be expressed by different users towards a specific piece of information (fake) where a group of users may believe and repost information blindly based on their preexisting beliefs or because a credible source received it while others may further search for other external sources for a piece of evidence in order to verify or dismiss new information. Two major psychological and cognitive factors exist to demonstrate that people are prone to fake news where they cannot, by nature, discriminate between fake and real news. (i) *Naive Realism:* individuals tend to believe that their perceptions of reality are the only accurate views, while others who disagree are regarded as uninformed, irrational, or biased [213]; (ii) *Confirmation Bias:* people tend to trust and prefer to receive information that confirms their existing views or beliefs [177]. Indeed, "citizens will rely on their beliefs when they are unable to believe alternative accounts" [145, p. 1]. According to Pennycook and Rand [192], individuals fail to think analytically when encountering misinformation, and thus they easily fall for it. This is especially true with information that agrees with their prior knowledge and beliefs [30, 210]. Owing to such cognitive biases, vulnerable users often perceive fake news as real. The matter of concern is that once such a misperception is formed, it is then very challenging to change it. Psychology studies show that true information is not only unhelpful in correcting false information (e.g., fake news) but also may sometimes increase the misperceptions [178]; Similar to confirmation bias, (iii) *Selective Exposure:* where users often tend to prefer information that confirms their preexisting attitudes, as the art historian Mark Jones stated, with some hyperbole: "Each society, each generation, fakes the thing it covets most" [118, p. 13]; and (iv) *Desirability Bias:* users are more likely to accept information that pleases them [297]. In the realm of fake news, the power to sway individuals' beliefs and behaviours is evident. Vulnerable users unknowingly spread fake news due to cognitive biases like naive realism and confirmation bias. These biases hinder discernment between real and fake news, with confirmation bias reinforcing preexisting beliefs. Additionally, encountering misinformation often curbs analytical thinking, amplifying the challenge of correcting false beliefs.

2.3.2 Social theories

Under this heading, we discuss social science theories demonstrating people's tendency to spread fake news. For example, *Prospect Theory* describes decision-making as a process through which people make choices to promote relative gains or diminish relative losses compared to their current state [121, 259]. As stated by *Social Identity Theory* [252, 253] and *Normative Influence Theory* [15, 122], this social acceptance and confirmation is a must to reflect person's identity and self-esteem. Based on that, people usually tend to choose "socially safe" options; that is, when a social group of users consumes fake news, they, as a group, are likely to disseminate news assuming that it increases social gain. Some of these theories have been categorised in Table 1; for detailed descriptions, readers are referred to [297].

2.4 Fake news on OSNs

Under this heading, we discuss some key features of fake news widespread on OSNs. Unlike traditional media ecology (e.g., television), fake news spreads very quickly on OSNs, and here we explain the characteristics that cause this phenomenon.

2.4.1 Malicious accounts on OSNs

Using social media has led to an increase in fake account creation that breeds the daily spread of fake news for specific purposes. Real humans do not necessarily manage some malicious accounts but can also be bots. Research has shown that fake news pieces will likely be created and spread by non-human accounts, such as social bots or cyborgs [228, 235] or trolls. (i) *Social bot* can be described as a social media account that is managed by a computer algorithm to automatically produce content and interact with humans (or other bot users) on OSNs [69]. Driven by benefits, social bots can distort a large amount of information on OSNs and often use the intention to spread falsified information. For example, it was found that a considerable amount of online social bots distorted the 2016 U.S. presidential election online discussions [28]. In the week before election day, a massive amount of social bots accounts on Twitter (i.e., roughly 19 million) published tweets supporting either Clinton or Trump [108]. Moreover, some current efforts discussed how social bots coordinated misinformation campaigns during the 2017 French presidential election [68]. As Howard et al. [104, p. 1] clarify, "both fake news websites and political bots are crucial tools in digital propaganda attacks—they aim to influence conversations, demobilise opposition and generate false support". A recent study, for instance, showed that social bots are considered the catalyst for fake news spreading on social media platforms where they amplify the diffusion of content coming from low-credibility sources suggesting that "curbing social bots may be an effective strategy for mitigating the spread of low credibility content" [229, p. 5-6]. Another group of users likely to spread fake news are so-called (ii) *trolls*, real human users who publish inflammatory messages that carry emotional responses on social media platforms or other newsgroups to manipulate the public. Evidence, for instance, indicates that 1,000 paid Russian trolls were disseminating fake news on Hillary Clinton [106], suggesting that these malicious users are often paid to spread false information. The trolling effect sparks people's inner negative emotions (e.g., anger and fear), leading to mistrust and irrational behaviour [235]. (iii) *Cyborg account* is another type of malicious accounts. Cyborg users are human users rather than bots who often utilise automation to spread fake news. Cyborg's account usually is registered by a human as camouflage and set automated programs to perform activities on social networks [180].

Table 1 Social sciences theories

News-related theories	User-related theories	
	Social impact	Self-impact
Undeutsch hypothesis [260]	Conservatism bias [25]	Confirmation bias [177]
Reality monitoring [116]	Semmelweis reflex [21]	Naive realism [213]
Four-factor theory [304]	Normative influence theory [55]	Desirability bias [71]
Information manipulation theory [161]	Social identity theory [16]	Selective exposure [72]

These malicious accounts promote fake news dissemination on OSNs; thus, precautionary measures are required to mitigate the effects of such accounts. Twitter, for example, deleted up to 6 per cent of all its registered suspicious accounts [107]. As social media emerges as the modern battleground for information warfare, the intricate interplay between human and non-human agents comes into focus. This subsection delves into the diverse range of actors responsible for propagating fake news, including social bots, trolls, and cyborg accounts. This exploration illuminates the multifaceted nature of fake news dissemination by scrutinising their motivations and tactics.

2.4.2 Echo chamber effect

The OSNs are one of the most powerful sources of information by which users can communicate and share their opinions. However, a disruptive new phenomenon in the news ecosystem has risen from OSNs: the so-called echo chambers [1].

Studies revealed that users on Facebook are more likely to form polarised groups, i.e., echo chambers, and choose the information that follows their belief system [53]. Users on Facebook always follow similar-minded users and therefore consume news that supports their preferred existing stories [207], leading to an echo chamber effect. Owing to the following psychological factors, the echo chamber effect simplifies the process of disseminating fake news [189]: The first factor is (i) *Social Credibility* where users frequently believe that a source is reliable if others believe it to be reliable, even when there is not enough evidence to determine whether the source is telling the truth. The second factor (ii) *Frequency Heuristic* lies in the fact that users naturally prefer information they frequently hear regardless of its truthfulness. In the echo chamber, consumers tend to share the same information. Research showed that increased exposure to an idea is sufficient to generate a positive opinion of it [287, 288]. With the flooding of information on OSNs, usually, the existence of the so-called echo chamber effect amplifies and reinforces biased information [110]. That being the case, this echo chamber effect forms segmented and homogeneous communities with a relatively limited information ecology, which, as studies stated, becomes the major factor of information dissemination that further promotes polarisation [52].

To some extent, all the factors mentioned in the previous subsections are related to the echo chamber. In turn, this leads to the emergence of homogeneous groups in which individuals share and discuss similar ideas. Groups such as these usually have polarised views since they are insulated from opposing perspectives [185, 249, 250]. This type of close-knit community is responsible for the major dissemination of misinformation [52]. Indeed, several possible interventions for preventing the spread of falsified information on OSNs have been proposed, ranging from (1) curtailing the most active (and presumably bot) users [229] to (2) harnessing the flagging activities of users in collaboration with fact-checking groups. In [264], the second intervention strategy is proposed as the first viable mitigation tool to reduce misinformation dissemination by utilising users' Facebook reporting activities. As a result, many popular organisations are now tackling the dissemination of fake news. In addition, in certain countries, Facebook works with third-party fact-checking organisations to review, rate, and identify information's accuracy [38]. On the other hand, Twitter introduced new labels and warning messages as an initiative in May 2020 to curtail the misinformation around COVID-19 [214] and to notify people about the falsified tweets, facilitating the process of fact-checking such tweets in order to make informed decisions. For instance, in

January 2021, Twitter launched BirdWatch [47], a community-driven strategy enabling users to identify tweets they perceive to be misleading. In summary, individuals are increasingly drawn to communities that share similar viewpoints, resulting in isolated information environments that strengthen their pre-existing convictions. These divided echo chambers enable the effortless spread of false information, accentuated by psychological aspects like social credibility and the frequency heuristic. Researchers and society at large need to recognise the potential consequences of these echo chambers and work towards fostering more open and diverse information ecosystems to mitigate the perpetuation of biased narratives.

3 Existing FND approaches

FND is a critical task in the field of information processing, aiming to distinguish between true and false information circulating in various media sources, particularly OSNs. It involves the development of computational methods and techniques to automatically identify and classify news articles, headlines, or social media posts that contain deceptive, misleading, or fabricated information. The proliferation of fake news has raised concerns about its detrimental impact on individuals, societies, and democratic processes, as it can influence public opinion, incite polarization, and even contribute to real-world consequences. FND involves the application of various approaches, including NLP, ML, and DL, to analyze textual content, contextual information, source credibility, and other features to discern the veracity of news items.

Trustworthiness and *veracity* analytics of online statements is a hot research topic [215] recently. This includes predicting information credibility shared in social media [166], stance classification [301] and contradiction detection [141]. Previous related studies rely heavily on textual content features to detect news veracity. Under this heading, we summarise classic and recent work on FND. Given how fast fake news is disseminated through OSNs and other websites, it would lead to real consequences if it is not paid enough attention. It is difficult for a human being to distinguish fake from real news. In one study, by a rough comparison scale, human judges achieved a success rate of only 50-63% in identifying fake news [217]. Another study found that respondents found it "'somewhat' or 'very' accurate 75% of the time" when shown a fake news article, and another discovered that 80% of high school students had a hard time determining whether an article was fake [61, 63]. Human efforts have been intensified to combat the spreading of false information. Two of the most common examples of fact-checking websites developed to reduce the effect of growing misinformation are Snopes[5], Politifact[6] and others. Previously, with more and more user-generated content (UGC) on OSNs, fact-checking websites and tools are vital to validate the information integrity [183] and reduce the effect of falsified information. These websites have been developed to verify the truthfulness of news, where annotations must be made manually by journalists and other experts who examine the article's content to determine its authenticity. These websites perform a fact-checking task (i.e., the assessment of the truthfulness of a news story or claim [262]) to verify the information veracity by comparing them with one or more reliable sources [171]. One of the shortcomings of such sites is that they require extensive expert analysis, i.e., labour-intensive and time-consuming, which results in a late response. Furthermore, due to the volume of newly generated information, particularly on OSNs, these websites do not scale well [286]. Research by Gartner [34] predicts that "By 2022, most people in mature economies will consume more false information than true information".

[5] https://www.snopes.com/

[6] https://www.politifact.com/

Table 2 Manual fact-checking websites

Fact-checking tools	Type of fact-checking	Domain
PolitiFact	Expert-based	American politics
Fiskkit	Crowed-sourced	News articles and comments
Snopes	Expert-based	Politics and social topical issues
TruthOrFiction	Expert-based	Politics, medical, religion, food nature, etc.
HoaxSlayer	Expert-based	Ambiguity
GossipCop	Expert-based	Hollywood and celebrities
FullFact	Expert-based	Economy, health, education, crime, immigration, law
Social Media sites	Crowed-sourced	Multiple domains of user posts

Thus, it becomes more obvious how it would be worth automating the FND process if you consider how much a manual detection may cost in terms of both time and human effort. It is not an easy job for sure, given that the proposed models need to accurately understand the natural language nuances of fake and real content. Table 2 presents some of these manual fact-checking websites and tools (PolitiFact[7], Fiskkit[8], Snopes[9], TruthOrFiction[10], HoaxSlayer[11], GossipCop[12], FullFact[13], Social Media sites[14]) designed for information veracity detection.

This section aims to provide a comprehensive overview of the related work on FND that utilised an assortment of ML and DL approaches based on three main categories of FND methods: content-based, context-based, and hybrid-based methods. It also provides a comprehensive review of the commonly used FND datasets. We summarise the existing approaches in Table 3. Additionally, we describe these existing representative works in the following subsections.

3.1 Content-based

Earlier research on FND mainly relied on hand-engineering relevant data that exploited linguistic features. Several ML and DL methods are employed to solve the classification problem, ranging from logistic regression to convolutional and recurrent neural networks. Text classification has traditionally used statistical ML methods such as Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbour (K-NN), Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), Gradient Boost (GB), and XGBoost (XGB). Many studies have applied the above mentioned algorithms to detect fake news content, achieving high accuracy. For example, a study by Bali et al. [20] extracted a set of features from both news headlines and news contents, such as the n-grams count feature for automatic FND. Their study—using seven different classification algorithms on three different datasets, found Gradient Boosting (XGB) classification algorithm to be the best, yielding the highest

[7] http://www.politifact.com/

[8] https://fiskkit.com/

[9] https://www.snopes.com/

[10] https://www.truthorfiction.com/

[11] http://hoax-slayer.com/

[12] https://www.gossipcop.com

[13] https://fullfact.org/

[14] Such as Twitter, Facebook, and Sina Weibo

Table 3 A summary of related work on FND

Type of features	ML	DL
Content-based	Pisarevskaya [197]	Popat et al. [199]
	Pérez-Rosas et al. [193]	Wang et al. [273]
	Potthast et al. [200]	Wu et al. [278]
	Fuller et al. [74]	Goldani et al. [84]
	Ahmed et al. [4]	Girgis et al. [82]
	Bharadwaj [29]	Shushkevich et al. [239]
	Wynne et al. [281]	Gautam et al. [78]
	Gravanis et al. [89]	Shifath et al. [232]
	Burgoon et al. [33]	Veyseh et al. [261]
	Horne et al. [100]	Kula et al. [130]
	Newman et al. [172]	Aggarwal et al. [2]
	Zhou et al. [292]	Jwa et al. [120]
	Papadopoulou et al. [184]	Baruah et al. [23]
	Rubin et al.[219]	Wani et al.[274]
	Castillo et al. [37]	Sadeghi et al. [222]
	Wang [272]	Khan et al. [123]
	Hosseini et al. [102]	
	Bandyopadhyay et el. [22]	
	Patwa et al. [187]	
	Shushkevich et al. [239]	
	Elhadad et al. [64]	
	Ahmad et al. [3]	
	Bali et al. [20]	
	Arun et al. [14]	
Context-based	Tacchini et al. [251]	Yu et al. [149]
	Volkova et al. [266]	Xinyi et al. [295]
	Ma et al. [155]	Kwon et al. [135]
	Ma et al. [153]	Ke et al. [277]
Hybrid-based	Volkova et al. [265]	Wang [272]
	O'Donovan et al. [179]	Conroy et al. [48]
	Gupta et al. [92]	Ruchansky et al. [220]
	Singh et al. [243]	Shu et al. [237]
	Khattar et al. [124]	Giachanou et al. [81]
	Dou et al. [62]	Long et al. [150]
		Alhindi et al. [10]
		Roy et al. [216]
		Shu et al. [233]
		Koloski et al. [127]
		Jin et al. [111]
		Singhal et al. [244]
		Cui et al. [49]
		Zhou et al. [294]
		Chenguang et al. [245]
		Qian et al. [205]
		Ying et al. [284]
		Wu et al. [280]

accuracy across all datasets. Perez-Rosas et al. [193] present a dataset of false and true news articles and analyse various features of the news articles (including n-grams, punctuation, grammar, and readability). Then, a linear SVM classifier is trained based on these features, with varying results depending on the feature set. In this study, computational linguistic features were shown to be useful in detecting false news automatically.

Potthast et al. [200] attempted to assess the stylistic similarity of several categories of news: hyper-partisan, mainstream, satirical, and false. A meta-learning approach originally intended for authorship verification is employed in the proposed methodology. As a result of comparing topic- and style-based features with RF classifiers, the researchers concluded that while hyper-partisan, satire and mainstream news are distinguishable, style-based analyses alone cannot detect false news. Fuller et al. [74] developed a linguistic-based method for deception detection consisting of thirty-one linguistic features where three classifiers were to be used to refine them to only eight cues. Such cues were based on the earlier proposed different clue sets in the linguistic field [190, 292]. However, their work is disadvantageous in that the cues heavily relied on the text's topic or domain, leading to generalisation issues where the model could not generalise well when tested on contents from different domains [11]. Using a relatively simple approach based on term frequency (TF) and term frequency-inverse document frequency (TF-IDF) has shown effectiveness in some previous studies. For example, Riedel et al. [211] applied a multi-layer perception (MLP) in the context of the fake news challenge dataset, and they have shown that using a relatively simple approach based on TF and TF-IDF yielded a good accuracy of 88.5%. In a study conducted by Ahmed et al. [4], the performance of the linear SVM classifier on a fake news dataset, so-called ISOT, has been tested and achieved promising results of 92% accuracy. Furthermore, Ahmed et al. [3] tested the performance of various classifiers, including but not limited to SVM, LR, and RF, to detect fake news. In their study, which uses several fake news datasets, the RF classifier also provides promising results on almost proposed datasets.

Bharadwaj [29] has experimented with different features such as TF and TF-IDF with n-gram features, and the results show that RF with bigram features achieves the best accuracy of 95.66%. Wynne et al. [281] experimented with character n-grams and word n-grams to study their effect on detecting fake news and concluded that the former contributes more towards improving FND performance compared to word n-grams. The TF-IDF and Count Vectorizer were used by [89] as feature extraction techniques. They demonstrated that their approach was more accurate than state-of-the-art approaches like Bidirectional Encoder Representations from Transformers (BERT). Linguistic approaches for FND have been applied using both supervised and unsupervised approaches. The prospective deceivers use certain language patterns, such as lots of phrasal verbs, certain tenses, small sentences, etc., which have been revealed by some experiments performed by psychologists in cooperation with linguistics experts and computer scientists [33, 96, 172, 255, 292]. For example, 16 language variables were investigated by Burgoon et al. [33] in order to see if they may help distinguish between deceptive and truthful communications. They conducted two experiments to construct a database, in which they set up face-to-face or computer-based discussions, with one volunteer acting as the deceiver and the other acting genuinely. Then such discussions were transcribed for further processing, and they came up with certain linguistic cue classes that could disclose the deceiver; refer to Table 4. They employed the C4.5 Decision Tree technique with 15-fold cross-validation to cluster and produce a hierarchical tree structure of the proposed features. In a short sample of 72 cases, their method's overall accuracy was 60.72 per cent. They concluded that noncontent words (e.g., function words) should be included when studying social and personality processes. According to the authors, linguistic style markers like articles, pronouns, and prepositions are as important as specific nouns and verbs in revealing

Table 4 Linguistic based features used in [33]

	Features
Quantity	# syllables
	Count of words
	Count of sentences
Vocabulary complexity	# big words, # syllables per word
Grammatical complexity	# short sentences
	# long sentences
	avg # of words per sentence
	sentence complexity
	number of conjunctions
Specificity and expressiveness	emotiveness index,
	rate of adjectives and adverbs,
	# affective terms

what individuals are thinking and feeling, and thus liars tend to tell stories that are less complicated, less self-relevant and more negative.

Liars were more likely than truth-tellers to utilise negative emotion (more negative feeling) terms [173]. Liars may feel guilty, either for their lie or the topic they lied about [270]. Knapp et al. [126] observed that liars were far more likely than truth-tellers to make disparaging statements about their communication partners. They observed that liars typically use more other references than truth-tellers; however, this is inconsistent with what [173] found, where they observed that liars used third-person pronouns at a lower rate than truth-tellers. According to [173], (i) liars employed fewer "exclusive" words than truth-tellers, implying a lower level of cognitive complexity. When someone employs words like but, unless, and without, they are distinguishing between what belongs in a category and what does not, and (ii) liars employ more "motion" verbs than truth-tellers, as simple, tangible descriptions are provided by motion verbs (e.g., walk, go, carry), which are more easily accessible compared to words (e.g., think, believe) that focus on evaluations and judgments. Horne et al. [100] applied an SVM classifier with a linear kernel using several linguistic clues. The authors cast the problem as a multi-class classification, attempting to determine whether an article is real, fake, or satire where classes are equally distributed. After 5-fold cross-validation with Buzfeed news (see Section 3.5) that was enriched with satire articles, they got a 78% accuracy. The feature set they employed mostly consisted of POS tags and certain Linguistic Inquiry Word Count (LIWC) word categories. As a result, they concluded that real news and false news are substantially different in their titles, while the content of satire and false news is somewhat similar.

Following the same vein, Newman et al. [172] thoroughly examined five experimental case studies, each of which had a different number of participants who were asked to be deceptive or genuine. They concluded with a collection of five out of twenty-nine linguistic cues (e.g., first-person singular pronouns, third-person pronouns, negative emotion words, exclusive words, and motion verbs) as the most significant predictors of deception. Again, the authors utilised LR to evaluate features and obtained better results than human assessors

Table 5 Linguistic based features used in [172]

	Features
Standard linguistic dimension	Word Count
	% Words captures, dictionary words
	% Words longer than six letters
	% Total Pronouns, % First Person Singular
	% Total First Person, % Total Third Person
	% Negations, % Articles
	% Prepositions
Psychological processes	Affective or emotional processes
	Positive emotions, Negative emotions
	Cognitive Processes, Causation
	Insight, Discrepancy, Tentative, Certainty
	Sensory and Perceptual Processes, Social Processes
Relativity	Space, Inclusive, Exclusive
	Motion Verbs, Time, Past tense verb
	Present tense verb, Future tense verb

(67% vs 52% of accuracy). Table 5 shows the feature set proposed in their study. A study examining fake COVID-19-related news was conducted by Bandyopadhyay et al. [22], where the authors analysed data from 150 users by extracting information from their social media accounts, such as Twitter, and their email, mobile, and Facebook for the period spanning from March 2020 to June 2020. During the pre-processing phase, unrelated and incomplete news was removed. In this case, using K-NN as a classifier, the best prediction result was for June, with 0.91 F1-score, and the worst was for March, with 0.79 F1-score. Several ML baseline models, such as DT, LR, GB and SVM, are used in [187] to detect COVID-19-related fake news. In constructing an ensemble of bidirectional Long Short-Term Memory (BiLSTM), SVM, LR, NB, and NB combined with LR, the researchers [239] achieved a 0.94 F1 score.

In addition, Zhou et al. [292] presented a collection of twenty-seven linguistic characteristics divided into nine groups. They conducted an experiment in which the players in that scenario communicated using a web-based messaging system. Participants were split into pairs, one acting as the deceiver and the other acting honestly. The authors then used statistical analysis to evaluate the features, demonstrating the feasibility of using linguistic-based cues to differentiate between true and false texts. Table 6 shows the feature set proposed in their study.

A wide range of text stylistic, morphological, grammatical, and punctuation could serve as useful cues to detect news veracity. These sets of features were adopted by Papadopoulou et al. [184] using a two-level text-based classifier to detect click baits. Similarly, Rubin et al. [219] used some of these features, such as punctuation and grammatical features. Using supervised learning, Castillo et al. [37] assessed the credibility of content on Twitter. Topics categorised as news or chat by human annotators are extracted, and a model is built that determines which topics are newsworthy based on their credibility labels. Popat et al. [199] proposed an explainable attention-based neural network framework for classifying true and

Table 6 Linguistic based features used in [292]

	Features
Quantity	Words, verbs
	Noun phrases, sentences
Complexity	avg # clauses, avg sentence length
	avg word length, avg noun phrase length, pausality
Uncertainty	Modifiers, # Modal Verbs
	# Uncertainty, # Other reference
Non immediacy	Passive voice, objectification
	Generalizing terms, self reference, group reference
Expresivity	Emotiveness lexical diversity
	Content word diversity, redundancy, typographical error ratio
Specificity	Spatio-temporal information
	Perceptual information
Affect	Positive affect
	Negative affect

false claims and providing self-evidence for the credibility assessment. The goal of Hosseini et al. [102] is to use news content to detect different (six) categories of false news (from satire to junk news). In their analysis, they analysed documents using tensor decomposition to capture latent relationships between articles and terms and spatial and contextual relations between them. They then employed an ensemble method to combine different decompositions to identify classes with higher homogeneity and lower outlier diversity yielding superior results to state-of-the-art techniques. Using Convolutional Neural Network (CNN) and pre-trained word embeddings, Goldani et al. [84] propose a capsule network model based on ISOT and LIAR datasets for (binary and multi-class) fake news classification. Their results showed that the best accuracy obtained using binary classification on ISOT is 99.8%, while multi-class classification using the LIAR dataset yielded 39.5%. Similarly, Girgis et al. [82] performed fake news classification using LIAR datasets. The authors employed three different models: a vanilla recurrent neural network (RNN), a gated recurrent unit (GRU) and an LSTM. Regarding accuracy, the GRU model results in 21.7%, slightly higher than the other two models (LSTM with 21.6% and RNN with 21.5%).

The technique of learning how to transfer knowledge is a concept in ML known as transfer learning, which stores and applies the knowledge gained from performing a specific task to another problem. Learning this way is useful for training and evaluating models with relatively small amounts of data. In recent years, pre-trained language models (PLMs) have become mainstream for downstream text classification [56], thanks to transformer-based structures. Major advances have been driven by the use of PLMs, such as ELMo [194], GPT [208], or BERT [56]. BERT and RoBERTa, as the most commonly utilised PLMs, were trained on exceptionally large corpora, such as those containing over three billion words for BERT [56]. The success of such approaches raises the question of how such models can be used for downstream text classification tasks. Over the PLMs, task-specific layers are added for each downstream task, and then the new model is trained with only those layers from scratch

[56, 147, 169] in a supervised manner. Specifically, these models use a two-step learning approach. In a self-supervised manner, they learn language representations by analysing a huge amount of text. This process is commonly called pre-training. Feature-based and fine-tuning approaches can then be used to apply these pre-trained language representations to downstream NLP tasks. The former uses pre-trained representations and includes them as additional features for learning a given task. The latter introduces minimal task-specific parameters, and all pre-trained parameters are fine-tuned on the downstream tasks. These models are advantageous in that they can learn deep context-aware word representations from large unannotated text corpora—large-scale self-supervised pre-training. This is especially useful when learning a domain-specific language with insufficient available labelled data.

Besides the fact that surface-level features cannot effectively capture semantical patterns in text, the lack of sufficient data constitutes a bottleneck for DL models. Thus, the power of BERT and its variations can be leveraged to build robust fake news predictive models. Relatively little research has been done to detect fake news using the recent pre-trained transformer-based models. The few observational studies that have been done using such models, despite the use of different methodologies and different scenarios, have shown promising results. One recent example of this is a study conducted by Kula et al. [130] presents a hybrid architecture based on a combination of BERT and RNN. Aggarwal et al. [2] showed that BERT, even with minimal text pre-processing, provided better performance compared to that of LSTM and gradient-boosted tree models. Jwa et al. [120] adopted BERT for FND by analysing the relationship between the headline and the body text of news using the FNC dataset, where they achieved an F1 score of 0.746. In an attempt to automatically detect fake news spreaders, Baruah et al. [23] proposed BERT for the classification task achieving an accuracy of 0.690. Although the BERT model has made great breakthroughs in text classification, it is computationally expensive as it contains millions of parameters (i.e., BERT base contains 110 million parameters while BERT large has 340 million parameters) [56]. Even though BERT is more complex to train (depending on how large a number of parameters are being used), a variation of BERT, so-called DistilBERT [225], provides a simpler and reasonable number of parameters compared to that of BERT (reducing BERT by 40% in size while retaining 97% of its language understanding abilities), thus, faster training (60% faster). With a larger dataset, larger batches, and more iterations, a robust BERT was developed, which is the so-called RoBERTa [147]. A benchmark study of ML models for FND has been provided by [123], where the authors formulated the problem of FND using three different datasets, including the LIAR dataset, as a binary classification. Their experimental results showed the power of advanced PLMs such as BERT and RoBERTa.

In a study conducted by Gautam et al. [78], the authors applied a pre-trained transformer model, so-called XLNet, combined with Latent Dirichlet Allocation (LDA) by integrating contextualised representations generated from the former with topical distributions produced by the latter. Their model achieved an F1 score of 0.967. In the same vine, a fine-tuned transformer-based ensemble model has been proposed by [232]. The proposed model achieved 0.979 F1 scores on the Constraint@AAAI2021-COVID19 fake news dataset. Similarly, the authors in [261] carried out several experiments on the same dataset, and they proposed a framework for detecting fake news using the BERT language model by considering content information and prior knowledge and the credibility of the source. According to the results, the highest F1 scores obtained ranged from 97.57 to 98.13. By applying several supervised ML algorithms such as CNN, LSTM, and BERT to detect COVID-19 fake news, the authors in [274] achieved the best accuracy of 98.41% using BERT cased version. Alghamdi et al. [5] conducted a comprehensive benchmark study to evaluate the effectiveness of various ML and DL techniques in detecting fake news. The study involved the use of

classical ML algorithms, advanced ML algorithms, and DL transformer-based models. The experiments were performed on four real-world fake news datasets, namely LIAR, PolitiFact, GossipCop, and COVID-19. The authors utilised different pre-trained word embedding methods and compared the performance of different techniques across the datasets. Specifically, they compared context-independent embedding methods, such as GloVe, with the effectiveness of BERT, which provides contextualised representations for FND. The results showed that the proposed approach achieved better results by solely relying on news text compared to the state-of-the-art methods across the used datasets. In a study conducted by [9], the authors focused on detecting COVID-19 fake news, considering the risks associated with disseminating false information during the pandemic. For this task, they investigated the effectiveness of various ML algorithms and transformer-based models, specifically BERT and COVID-Twitter-BERT (CT-BERT). To assess the performance of different neural network structures, the authors experimented with incorporating CNN and BiGRU layers on top of the BERT and CT-BERT models. They explored variations such as frozen or unfrozen parameters to determine the optimal configuration. The evaluation was conducted using a real-world COVID-19 fake news dataset. The experimental results revealed that incorporating BiGRU on top of the CT-BERT model yielded outstanding performance, achieving a state-of-the-art F1 score of 98%.

We believe that choosing effective features plays a crucial role in obtaining a good performance on news verification. Exploiting visual content to examine the truthfulness of social news events on OSNs is essential [113], and thus, incorporating visual cues, specifically images, with textual features is effective and would result in better performance in detecting fake from real news. The reason to justify this is that, according to several studies, visual information is easier to interact with and retain than its text-based counterparts, implying the importance of its use. This is because using images instead of 140 characters is more interesting, leads to more interaction, and is thus widespread. According to Twitter statistics, tweets with images gain roughly 35% higher retweets than text-only tweets. So, exploiting such cues in detecting fake news events on OSNs is a golden opportunity. Although the studies mentioned above that relied largely upon features derived from content perform well in identifying fake content, their major limitation is that content-based features alone cannot be used to characterise fake news content properly. The fakers mainly utilise different tactics to mimic trustworthy content. Therefore, considering other auxiliary information besides content-based features would provide a more comprehensive understanding of the phenomenon, resulting in higher detection performance.

3.2 Context-based

With the continued growth of social media platforms and the rise in falsified information that mimics trustworthy information being generated, people worldwide face a significant challenge in discriminating fake claims from real ones. The process of detecting fake news is inherently challenging because fake news is usually written intentionally to mislead the readers. One solution to this problem currently being explored is, in addition to the linguistic and lexical cues of the article, the use of auxiliary contextual information, including user engagements and activities on social media platforms. If effective, such a solution could potentially boost detection performance. However, high-quality data, particularly online social media data, poses another challenge to the process, including misspellings and others. Adding to the problem is the access restriction posed by Twitter API. As such, developing an automated solution with high accuracy is, therefore, challenging. Besides content-based information,

additional contextual features about the context—derived from user social engagements on microblogging can be used for successful FND. User social engagements represent the diffusion of news over time, giving useful auxiliary information to infer the news article's veracity [235]. Researchers relied heavily on using news content to model user behaviour and interests to detect fake news based on the assumption that users' posts and activities on microblogs and social networks often reflect their behaviour. Context-based features refer to users' interactions and engagements through social media platforms. This includes social context features such as the number of friends (followers), number of posts, replies [136, 302], and others.

These features have been investigated by the existing work to characterise fake content. For example, Tacchini et al. [251] present a framework wherein the authors used conspiracy theories and scientific pages as sources to build a dataset by collecting a set of posts and users for the goal of detecting fake news based on users who liked them on Facebook. They compare LR to a harmonic method that demonstrates the effectiveness of their model using only a small percentage of the training data.

Volkova et al. [266] proposed a fusion neural-based model using a combination of tweet text, linguistic cues such as moral foundations, and user interactions to detect four types of suspicious news: satire, hoaxes, clickbait, and propaganda. They compared their approach with state-of-the-art techniques. The authors discovered that adding syntax and grammar features does not affect performance, whereas incorporating linguistic features improves classification results, with social interaction features being the most informative for finer-grained separation of such suspicious news posts. Propagation of news items on OSNs has proven effective in uncovering useful patterns to help discriminate fake content from real ones. For example, Yu et al. [149] performed an early detection of false news by modelling diffusion pathways as multivariate time series using a hybrid model of CNN and GRU. Their approach is tested on two real-world datasets from Twitter and Sine Weibo, outperforming other state-of-the-art algorithms.

The goal of Wu et al. [278] is to investigate the propagation of falsified messages in social networks. As a result, they used the Twitter API and the fact-checking website Snopes to create a custom dataset that included both genuine and fake news. Furthermore, they employ a neural network model to classify news after inferring embeddings for users from the social graph. To this end, they developed a new model for embedding a social network graph in a low-dimensional space and built a sequence classifier by analysing message propagation pathways using Long Short-Term Memory (LSTM) networks. Another study conducted by Wu et al. [277] detected fake news/rumours by examining high-order propagation patterns using a graph kernel-based model. Similarly, Ma et al. [153] evaluated the similarities between propagation tree structures of news by applying the same classifier and features for their verification. Later, they suggested a top-down/bottom-up tree-structured neural network for rumour detection; in other words, they made use of a non-sequential propagation structure for identifying different types of rumours [155]. Kwon et al. [135] used the network and temporal features to detect rumours, and their findings demonstrated the importance of such features in detecting and identifying rumours over longer periods.

Finally, based on the assumption that fake news has a different propagation pattern than other types of news, one study examined the propagation pattern among news publishers and subscribers using a propagation network [295]. However, the downside of the approaches that solely relied on context-based features is that these approaches are unable to effectively detect fake news content as early as possible, i.e., early fake news detection. This is because such information is insufficient or often unavailable at the early stage of fake news dissemination. Therefore, this calls for building approaches that can effectively harness news content to detect fake content as quickly as possible before such content goes viral.

3.3 Hybrid-based features

Incorporating heterogeneous features, often called hybrid features, has demonstrated favourable results across a range of tasks in diverse domains [202]. By combining multiple types of information or characteristics, such as textual, visual, or contextual features, the integration of hybrid features allows for a more comprehensive representation of the underlying data. Several related studies leverage the advantage of exploiting both content- and context-based features for FND. For example, Shu et al. [237] detected fake news by considering the tri-relationship between publishers, news items, and users. A non-negative matrix factorisation [188] and users' credibility scores were used to analyse user-news interactions and publisher-news relations. The performance of several classifiers was evaluated on the FakeNewsNet dataset, and the findings suggested that the social context could be leveraged effectively to improve the detection of fake news. A natural language inference approach (i.e., inferring the veracity of the news item) using BiLSTM and BERT embeddings is proposed by Sadeghi et al. [222] using the PolitiFact dataset. The authors aimed to use NLI methods to improve several classical ML and DL models, such as DT, NB, RF, LR, k-NN, SVM, BiGRU and BiLSTM, using various word embedding methods, including context-independent word embedding methods such as Word2vec, GloVe, fastText, and context-aware embedding models such as BERT. The experimental results show the effectiveness of using such methods for FND, achieving an accuracy 85.58% on the PolitiFact dataset.

To classify suspicious and trusted news, Volkova et al. [265] built models based on linguistic features. The authors aimed to classify 130 thousand news posts as either verified or suspicious and to predict four sub-types of suspicious news: hoaxes, satire, propaganda, and click baits, using different predictive models. Using tweet content and social network interactions, the author demonstrates that neural network-based models surpass lexicon-based models. Additionally, in contrast to earlier studies on deception detection, they discovered that incorporating grammar and syntax features into the models does not affect performance.

Although including linguistic features showed promising classification results, a finer-grained separation between the classes is achieved with social interaction features. Various features, according to O'Donovan et al. [179], can be used to predict content credibility. To begin, they defined a set of features that included content-based features, user profile features, and others that reflected the dynamics of information flow based on Twitter data. Then, after examining the distribution of each feature category across Twitter topics, they concluded that their efficacy varies greatly with context, both in terms of the occurrence of a particular feature and how it is used. Considering tweet-based textual features and Twitter account features, Gupta et al. [92] applied decision tree-based classification models that achieved a 97 per cent detection accuracy to detect fake and real images shared on Twitter using the Hurricane Sandy dataset. To identify fake news, Conroy et al. [48]—as one of the first researchers to apply network analysis in FND, reviewed linguistic and network approaches as two major categories of methods to uncover fake news characteristics. The former is in charge of revealing language patterns such as n-grams and syntactic structures, semantic similarity, and rhetoric relations between linguistic elements and their associations with deception. As the name implies, the latter deals with network information and propagation patterns that can be harnessed to measure mass deception.

Elhadad et al. [64] experimented with an assortment of ML algorithms such as DT, LR, SVM, multinomial NB, and neural networks using hybrid sets of features extracted from online news content and textual metadata on three publicly available datasets (ISOT, FA-KES and LIAR). In their study, both SVM and LR classifiers were shown to be the best-performing models on the LIAR dataset, while the best accuracy of 100% was obtained when using

Decision Trees on the ISOT dataset, and the best result of 58% was obtained when applying Multinomial Naive Bayes model on FA-KES dataset. Ruchansky et al. [220] proposed a hybrid DL model using two real-world datasets of Twitter and Weibo for FND. Their approach comprised three modules, namely capture, score and integrate. Given an article, the first module adopted RNN LSTM to capture the temporal pattern of user activity based on the response and text, while the second module, based on the behaviour of users, learns source characteristics, and those modules are combined to classify fake articles from real ones. Their experimental results demonstarted the importance of capturing the articles' temporal behaviour and the users' behaviour for FND, where they achieved high accuracies of 89% and 95%, respectively, on Twitter and Weibo datasets.

Alghamdi et al. [6] developed a deep 6-way multi-class classifier using the BERT model to classify statements in the LIAR dataset into fine-grained categories of fake news. The framework employed three main components: BERT_{base} was utilised for encoding and representing the text data, followed by passing it through a CNN and a max-pooling layer for feature map reduction. The metadata associated with the statements was encoded using another CNN to capture local patterns. The output from this CNN was then passed through a BiLSTM network to extract contextual features. The outputs from the two components were concatenated and fed into a fully connected layer for classification. The authors emphasised the importance of feature selection, particularly in the pre-processing stage, to improve the classifier's performance. They found that selecting relevant features, such as credit history, was crucial, as some other features were found to confuse the classifier and degrade its performance. The authors in [8] focused on detecting fake news by examining both the news content and users' posting behaviour. They employed DL techniques, specifically BERT, CNN and a BiGRU with a self-attention mechanism, to capture rich and contextual representations of news texts. By combining natural language understanding with transfer learning and context-based features, the proposed architectures aimed to enhance the detection of fake news. The experiments were conducted using the FakeNewsNet dataset. The results demonstrated that incorporating information about users' posting behaviours, in addition to textual news data, improved the performance of the models in detecting fake news.

Incorporating visual cues has also been investigated. For example, Wang et al. [273] proposed an event adversarial neural network (EANN) approach using textual and visual features from multi-modal data for FND. Their architecture comprises three modules: (1) multi-modal feature extractor (given news article, this module is in charge of extracting both textual and visual features using neural networks); (2) event discriminator (given news article, this component uses a min-max game in order to further captures event-invariant features; and (3) fake news detector (for news classification as either true or false). In addition, deep neural network (DNN)-based models, including Convolutional Neural Networks (CNN), BiLSTM, and ubiquitous transformers, have increasingly been the focus of research over the past few years. For example, Yang et al. [283] proposed a hybrid model using a CNN-based model by including textual and visual (images) features for FND. While their model was found to be effective, nevertheless, relying on the CNN model would result in neglecting contextual semantical relations existing in the context of the news content. Thus, when fusing CNN with other state-of-the-art techniques, such as the pre-trained BERT model, we expect the local and global contextual and semantical relations to be effectively captured, leading to fine-grained performance.

Using the attention mechanism, Jin et al. [111] built a hybrid framework by incorporating text, images, and context-based features. Their framework used a pre-trained VGG-19 network to extract visual features and an LSTM network to concatenate text- and context-based information. Then, these components are fused using an attention mechanism. To detect

fake news, Khattar et al. [124] utilised text- and visual-based information in a variational autoencoder model coupled with a binary classifier. More recent work by Singhal et al. [244] designed an architecture called Spotfake by conducting a survey in order to detect fake news based on text and images. Text feature representations are obtained using BERT, while image feature representations are obtained using a pre-trained VGG-19 network. Predictions are then made by combining the two modalities. Despite the success of their model's performance, their architecture failed to take advantage of the correlation between modalities, which is crucial when detecting fake news. They found that using the multimodal approach, 81.4% of people could discriminate fake news from real ones while relying on the unimodal approach (only text or only images), the discrimination rate is 38.4% if only and 32.6%, respectively.

Researchers have also focused on using auxiliary information beyond visual and textual information to detect fake news. As part of their deep framework, Cui et al. [49] incorporated user sentiment extracted from comments into the multimodal framework. According to their experimental results on PolitiFact and GossipCop datasets, they have achieved better F1 scores than baseline methods (77% and 80%, respectively). The similarity between the text and image features is also measured to evaluate whether the news is credible [294]. Fusing the relevant information between different modalities while maintaining the unique properties of each is a challenge. In addition, for some news, a fusion between different modalities may result in noise information that has a negative impact on the performance of the model. To handle these challenges, a multimodal FND framework based on Crossmodal Attention Residual and Multichannel convolutional neural Networks (CARMN) [245] is proposed. While maintaining the unique information about the target modality, this framework can selectively extract relevant information about a target modality from another source modality.

Multiple co-attention layers [205], hierarchical multimodal contextual attention networks by Ying et al. [284], and multimodal attention networks by Wu et al. [280] are employed to research complementary interactions among textual and visual features. Singh et al. [243] explored several classical ML algorithms (e.g., LR, SVM, linear discriminant analysis (LDA), K-NN, NB, and RF) for FND using multimodal data (i.e., text+image). Their experimental results showed the superiority of RF over other algorithms with an accuracy of 95.18%. Giachanou et al. [81] proposed a multimodal framework by incorporating BERT for text-based modelling content and the VGG-19 network for encoding visual information. The authors conducted their experiments using the FakeNewsNet dataset, and their approach achieved an F1 score of 79.55%. Using the LIAR dataset, W.Yang [272] applied a hybrid CNN approach using speakers' metadata and statements to detect fake news, and he found that the CNN model achieved the best accuracy of 0.27. Using the attention mechanism, Long et al. [150] applied a hybrid model based on LSTM coupled with attention on the LIAR dataset, where they achieved 0.41 in accuracy. Alhindi et al. [10] proposed adding justification to label the existing LIAR dataset, and they released LIAR-Plus; their approach is based on BiLSTM achieved 0.37 accuracy. Roy et al. [216] proposed an ensemble architecture based on CNN and BiLSTM using the LIAR dataset, and they achieved state-of-the-art performance on this dataset of 0.44 accuracy. Although these studies showed significant performance, however, they tend to encode the input unidirectionally using context-independent word embedding techniques.

Shu et al. [233] propose a sentence-comment co-attention subnetwork by jointly deriving explainable top-k check-worthy sentences and comments. The framework consists of (i) a module responsible for encoding news content by learning the news sentence representation in a hierarchical structure to capture semantic and syntactic clues using a hierarchical neural network and (ii) a module for encoding user comments through the word-level attention sub-

network in a way to learn the latent representations of the given comments, (iii) a component that is a catalyst for capturing the correlation between the former and the latter and selectively process top-k explainable sentences and comments, and (iv) a fake news classification component. According to the authors, the underlying assumption is that both news content and user comments are inherently related to each other and thus can provide useful clues for the interpretability of such a model. They concluded that user comments relevant to the content of original news pieces are useful for detecting fake news and explaining prediction results. However, their framework cannot be adapted to other real-world fake news datasets since not all datasets contain user social comments, for which their approach requires a long with news content. Koloski et al. [127] applied a deeper neural network model with various representations both for textual patterns and for the embeddings of concepts appearing in the given input text, achieving an accuracy of 0.88 on the PolitiFact dataset. A review by [7] provides a succinct overview and analysis of the FND phenomenon, offering valuable insights into this important area of research. Table 7 shows different features used in related work, where a unified DL framework that is not only capable of leveraging all these features but, most importantly, is explainable is urgently needed to advance FND.

The following section will provide an exposition of the fundamental characteristics that distinguish fake news, shedding light on the key attributes and distinctive traits commonly observed in the realm of fake news dissemination.

3.4 Fake news characteristics

Earlier work has conducted empirical research into the characteristics that distinguish real from fake content. Examining the text-based characteristics of articles, Horne et al. [100] analysed features that distinguish fake from real content. Fake news article titles were found to be longer, with more capitalised words, fewer stop words, repetition, and fewer nouns. According to Perez et al. [193], fake news articles featured more social, verbal, and temporal words, implying that falsified information was more focused on the present and future. Previous research has shown that there is a relationship between news diffusion and user activity on OSNs, where it is possible to analyse what is circulated on such microblogs of opinions, comments, and others to characterise and distinguish false from true news, and existing studies have tackled this. The responses to false and true news were analysed by [204], who discovered that compared to real news, fake news received more negative and doubtful comments.

False news reached more people than the truth, according to [268], and falsified information spread faster and deeper than the truth. They investigated true and false rumours on Twitter and found that false rumours evoked surprise, disgust and fear in their replies, whereas true rumours elicited joy, sadness, trust, and anticipation. Another study by Zuckerman et al. [304] found negative remarks as a key signal of deceit in an early meta-analysis. According to [173], (i) liars employed fewer "exclusive" words than truth-tellers, implying a lower level of cognitive complexity. When someone employs words like but, unless, and without, they are distinguishing between what belongs in a category and what does not, and (ii) liars employ more "motion" verbs than truth-tellers, as simple, tangible descriptions are provided by motion verbs (e.g., walk, go, carry), which are more easily accessible compared to words that emphasise evaluations and judgments (e.g., think, believe). A study by Li et al. [143] investigates users' beliefs in a massive amount of falsified information and observes how they evolve and define the roles of different types of users in such information dissemination. When there is no evidence to verify false information, people tend to spread them

Table 7 Previous studies on fake news and rumours detection using various features

Study	Cues					Approach
	TF	*VF*	*UF*	*PF*	*ME*	
Castillo et al. [37]	✓		✓	✓		J48 decision tree
Chang et al. [39]	✓	✓	✓			Clustering
Vosoughi et al. [267]	✓		✓	✓		Hidden Markov Models
Ma et al. [151]	✓			✓		RNN
Chen et al. [41]	✓		✓	✓		Anomaly detection, KNN
Wu et al. [278]	✓			✓		LSTM-RNN
Gupta et al. [94]	✓		✓	✓		Graph-based method
Gupta et al. [93]		✓	✓	✓		Graph-based method, DT
Qazvinian et al. [203]	✓			✓		L_1-regularized log-linear model
Zhao et al. [291]	✓					DT ranking method
Chua et al. [45]	✓					Linear Regression (LR)
Ma et al. [154]	✓			✓		Kernel-based method
Kwon et al. [137]	✓		✓	✓		Random Forest
Kwon et al. [135]	✓		✓	✓		SpikeM
Zubiaga et al. [303]	✓		✓	✓		Conditional Random Fields
Enayet et al. [65]	✓		✓			SVM
Yang et al. [282]	✓		✓			SVM
Qin et al. [206]	✓					SVM
Shu et al. [236]	✓		✓	✓		Neural Network
Jin et al. [112]	✓		✓	✓		LDA,Graph
Li et al. [143]	✓		✓	✓		SVM
Li et al. [144]	✓		✓	✓		LSTM
Shu et al. [233]	✓			✓	✓	Hierarchical Attention Network

Note that TF - Textual Features, VF - Visual Features, UF - User Features, PF - Propagation Features, ME - Model Explainability

without expressing their beliefs [32]. The role of user profile features for FND has been investigated in [238]. The authors found that user profile characteristics have a key role in characterising fake news. In an attempt to attract the audience's attention and spread extensively, fake news creators tend to write content that triggers readers' emotions in order to promote the success of their creations. That is, fake news usually has a strong positive or negative sentiment of hate, anger or resentment [242]. Furthermore, the role of multimodal information (textual and visual features) in FND has also been examined in the literature. Recent research [113, 293] found that some fake news content patterns, i.e., text and image style, are different from those of true news. The authors found that fake news text is more informal, diverse, subjective, and emotional than true news text. Furthermore, they found that when compared to real news images, fake news images often show higher coherence and clarity while lower clustering and diversity scores.

In summary, prior research has extensively examined the text-based characteristics that differentiate real from fake content. Findings indicate that fake news articles exhibit longer titles with more capitalization, fewer stop words, repetition, and fewer nouns. Additionally, fake news employs more social, verbal, and temporal words, indicating a focus on the

present and future. Studies have also explored the correlation between news diffusion and user engagement on OSNs, revealing patterns in user responses to fake news. The accelerated spread of false information compared to truth and its evocative impact on emotions have been highlighted. Linguistic analyses have identified linguistic cues of deceit, including the use of "exclusive" words and "motion" verbs. User profiles and emotional content have been investigated, along with the role of multimodal information in distinguishing fake news.

The following section will present an overview of the prevalent real-world fake news datasets extensively utilised in prior studies to detect and address the issue of fake news.

3.5 Commonly used fake news datasets

Handling unstructured data is a non-trivial task, especially when applying [deep] neural network-based models on such data where some noisy and sparsity issues are presented. These previously discussed studies used various datasets. The datasets used in the literature vary largely, where different datasets are related to different domains, such as politics. In the following paragraph, we summarise some popular fake news datasets. Data insufficiency refers to the lack of training data which in turn constitutes a major obstacle resulting in poor performance. As a pipeline of various aspects of data components, several previous studies applied ML and DL approaches using Twitter-based data (short text with or without propagation patterns such as retweets, likes, and conversational threads) or articles. According to [261], several studies rely heavily on the main tweet or post, while other studies take into account additional features of the news (e.g., replies and comments). Vlachos et al. [263] cast the task of fact-checking as a binary classification task using the k-Nearest Neighbour classifier and constructed a dataset from two popular fact-checking websites. Wang [272] presented the LIAR[15] publicly available dataset, which comprises 12.8K and contains two main components: user profile and short political statements. User profile features include the speaker's name, job, party affiliation, state, credit history, and context. The statements (reported during the time interval from 2007 to 2016) have been labelled by the editors of Politifact.com using six fine-grained categories, namely, true, mostly-true, half-true, barely-true, false and pants-fire. These six labels are relatively balanced in size. Overall, each statement has its associated label and information about the speaker of that statement.

Another dataset named PoliticalNews has been collected. Based on the fact that "a classifier trained using content from articles published at a given time is likely to become ineffective in the future" [35], the authors of this work collected a dataset spanning from 2013 to 2018 by crawling news websites in order to evaluate the performance of their model on different years. Fact Extraction and Verification (FEVER) is a dataset that comprises 185,445 claims formed by extracting data from Wikipedia. Without prior knowledge of the sentences of origin, these claims were verified and classified into three classes: supported, refuted or not enough info [256]. Several fake news datasets based on OSNs have been created, including BuzzFeedNews[16] published using Facebook by nine news agencies one week before the 2016 U.S. elections, which comprises 2282 samples. Five BuzzFeed journalists checked and verified every post or link in the data [241]. Similarly, the some-like-it-hoax dataset created based on the Facebook platform involves 15,500 posts and 909,236 users classified as either hoax or not hoax [251]. Ma et al. [151] used Twitter and Sine Weibo microblogs to collect five million posts comprising 778 reported events for fake news and rumour detection. In a study by Tanushree et al. [165], the authors released a large-scale social media corpus

[15] https://www.cs.ucsb.edu/william/data/liardataset.zip

[16] https://github.com/BuzzFeedNews/2016-10-facebookfact-check/tree/master/data

Table 8 A summary of fake news datasets used in the literature

Dataset	Number of instances	Number of classes
LIAR [272]	12.8K	6
CREDBANK [165]	4856	2
FakeNewsNet [234]	varied	2
PHEME [54]	6425	2
FEVER [256]	185,445	3
PolitiFact [100]	488	2
Weibo [151]	816	2
BuzzfeedNews [241]	2282	4
YelpChi [168]	67K	2
Twitter [156]	1111	2
COVID-19 [187]	10,700	2
Fake News Challenge (FNC)	75,385	4

comprising 37 million event-related tweets and 60 million event-related tweets, grouped into over 1049 events. PHEME dataset [54] contains several rumour source tweets associated with their replies. This dataset contains 6425 tweets that can be rumours and non-rumours. The CREDBANK[17] dataset is a collection of tweets that consists of tweet content and topics classified as events or non-events that are annotated with ratings stating their credibility [165]. The COVID-19 dataset is a collection of COVID-19-related social media posts, comments and news, classified as real or fake based on their truthfulness. The data set [187] is collected from various social media platforms, such as Twitter and YouTube. The challenge organisers collected 10,700 social media posts and news articles about COVID-19 in the form of an annotated dataset in English.

Fake News Challenge (FNC)[18] data set covers a different range of topics, including but not limited to topics like politics, health, environment, lifestyle, etc. Such data set involves the headlines or content of news articles and aims at labelling the stance associated with it into categories such as “agree”, “disagree”, “discuss”, and “unrelated”. The training set contains around 49,972 records, while the test set contains around 25,413 records. FakeNewsNet[19] is a comprehensive dataset[20] contains various rich information including textual, visual, spatiotemporal, and contextual information. The dataset consists of full-text news articles collected from politifact.com and gossipcop.com websites. Each of these includes tagged news content (e.g., news articles) and social context information (e.g., relevant social user interactions for news articles). Table 8 summarises the datasets in addition to other datasets.

4 Methods

In this section, we introduce the methods for FND, starting with the embeddings.

17 https://compsocial.github.io/CREDBANK-data/

18 https://github.com/FakeNewsChallenge/fnc-1

19 https://github.com/KaiDMML/FakeNewsNet

20 https://github.com/KaiDMML/FakeNewsNet

4.1 Text classification

Text manipulation has become effortless with the emergence of computers nowadays. Nevertheless, the numerical representation of the underlying texts is required for computers in order to manipulate them. Typically, several preprocessing steps are involved in the text classification framework required to convert such text-based information (e.g., words) into informative representations while preserving most linguistic features, namely, general preprocessing, feature extraction, feature selection, and finally, the classification phase. These steps, especially feature extraction (a.k.a feature encoding), are crucial, and the model performance would suffer if paid insufficient attention. Modelling text is challenging because it is messy, and techniques like ML algorithms require well-defined fixed-length inputs and outputs. ML algorithms require the text to be converted into numbers (specifically, vectors of numbers) as they cannot directly work with the raw text. These vectors capture more linguistic properties of the text; "in language processing, the vectors x are derived from textual data, in order to reflect various linguistic properties of the text" [86, p. 65]. Textual data requires a numerical representation and, most importantly, an efficient representation for computation and manipulation. Various statistical- and contextual-based methods have been developed to represent text numerically. The former is based only on statistical metrics, which typically generate sparse vector space, while the latter is based on the concept of word context, which rather produces dense vector space. Several approaches have applied such methods (e.g., bag-of-words [271], n-gram [51], and TF-IDF) methods [114]) as input for text classification using ML (e.g., Naïve Bayes (NB) classifiers [160], K-nearest neighbour (KNN) algorithms [258], and SVM [115]). Nevertheless, the contextualised representation that allows for efficient computation and captures the underlying patterns is urgently needed, especially with the incredibly massive amounts of textual data. Even though these statistical-based representation methods are computationally efficient and have achieved promising results where they are traditionally considered the centre of any text classification task, they are not free from one limitation. These methods focus entirely on capturing the frequency features of a particular word, and the contextual information of a text is fully disregarded, making it difficult to capture semantic-based information. To capture more semantics, PLMs (a.k.a context-independent models) are developed, such as Word2Vec and GloVe, which basically captures those semantics patterns but do little to capture context information. A significant amount of attention is devoted to developing context-aware representations of textual data, e.g., transformer-based models such as BERT, which has led to outstanding results in most NLP mainstream tasks. Figure 3 shows the general text classification process using ML and DL models. In the next subsections, we will introduce these representation models.

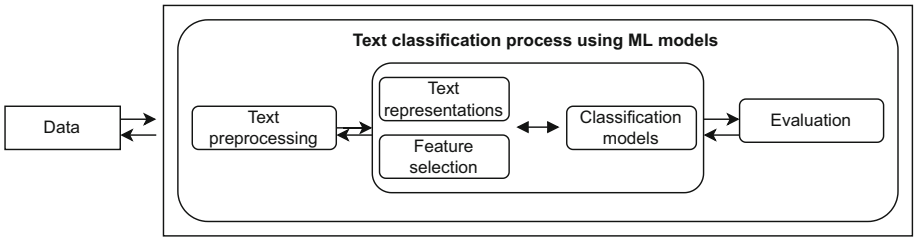

Fig. 3 Text classification process using ML and DL models

4.2 Embeddings

Word embedding is considered the centre of any NLP task, basically a form of word representation that bridges the human understanding of language to that of a machine. Word embeddings are typically learned from a large corpus of text. The distributed representational vector that captures the semantics of the input text can be obtained using two main alternatives that have been used for several text classification tasks in NLP. As mentioned previously, although statistical-based representation methods are computationally efficient and have shown promise as the foundation of text classification tasks, they are not without limitations. These methods focus solely on capturing frequency features of individual words, disregarding the contextual information of the text, thereby making it challenging to capture semantic-based information. To address this limitation, context-independent pre-trained models like Word2Vec and GloVe have been developed, which capture semantic patterns but overlook contextual information. Consequently, significant attention has been devoted to developing context-aware representations of textual data, particularly transformer-based models such as BERT. These models have achieved remarkable results across various mainstream NLP tasks.

Overall, these techniques are trying to model the following problem:

Assume we have a corpus of n documents $S = \{d_1, d_2, ..., d_n\}$ each of which consists of m words $W = \{w_1, w_2, ..., w_m\}$, and a vocabulary V; the embedding vector representation $\vec{w}_i$ is defined as mapping each word w_i in a specific document d_i into a continuous space $\mathbb{R}^d$ where d is the dimension of the vector space. Mathematically, words in a document can be mapped as follows:

$$w_i \rightarrow \vec{w}_i, \vec{w}_i \in \mathbb{R}^d \tag{1}$$

4.2.1 Non-contextualised embeddings–sparse vector representation-based

This subsection presents two statistical methods that generate sparse vector representations of documents. The first one is the popular and simple feature extraction method with text data, the bag-of-words (BoWs) method. The second one is the TF-IDF which overcomes the problem of the former.

Bag-of-words (BoWs) A bag-of-words model, or BoWs for short, is very popular, simple and flexible and can be used in a myriad of ways for extracting features from the text in order to be used for modelling using ML algorithms. A bag of words is a representation of text that describes the occurrence of words within a specific document. The idea of this distributional representation of features was investigated and proposed by Harris [97]. "A very common feature extraction procedure for sentences and documents is the BoWs approach. In this approach, we look at the histogram of the words within the text, i.e. considering each word count as a feature" [86, p. 69]. To illustrate, each word of a document is a feature (i.e., presents an embedding of that document) derived from the number of occurrences of each word in the document. A bag-of-words model, as the name implies, discards any information about the order or structure of words in the document and is only concerned with whether known words occur in the document, not where in the document, thus, failing to capture semantic patterns. To illustrate, given two sentences (the NLP is difficult, not easy at all, the NLP is easy, not difficult at all) with completely opposite semantic meanings, the BoWs model would give them the exact same representation just because they have the exact same words but in a different order, which is not effective. Another issue of such a model is that as the vocabulary size increases, so does the vector space dimension since, in this model, the number of words in the vocabulary forms the dimension of the vector representation of a document resulting

in what is called sparse representations. Such large dimensions, however, are bound to result in so-called *curse of dimensionality*, making it easy to fall into overfitting resulting in terrible out-of-sample performance.

Term frequency - inverse document frequency (TF-IDF) The problem with the BoWs method is that it treats all words equally important, and this is attributed to the fact that BoWs score each word based on its frequency in a document; thus, highly frequently occurring words dominate others in the document with a larger score which is problematic especially when such words are not as informative to the model as rarer occurring words. Rescaling the frequency of words by penalising the scores of the most frequent words across all documents is one approach to dealing with this issue. This approach is the so-called TF-IDF metric which has been proposed by [223] and has been widely used in many NLP tasks. TF-IDF method allows for quantifying words by reflecting how important a word is to a document in a corpus of documents. This method is premised on the idea that each word is assigned its own weight w_{ij} based on its appearance in the document and across all of the documents. These weights highlight words that are distinct and contain useful information in a given document. "Thus the idf of a rare term is high, whereas the idf of a frequent term is likely to be low" [227, p. 118]. For each word in a document j, the TF-IDF value is calculated by first calculating Term Frequency TF, which counts the number of occurrences of words in a document, then inverse document frequency (IDF), which is the catalyst for ensuring that words appear less commonly are assigned more weights compared to those occurring more frequently (e.g., stop words) which is calculated as follows:

$$\log(\frac{|D|}{df_i}) \tag{2}$$

where df_i denotes the number of documents that contains word i and $|D|$ refers to the number of documents in the corpus. TF-IDF metric is calculated as follows:

$$w_i = tf_{ij} \cdot \log(\frac{|D|}{df_i} \tag{3}$$

where tf_{ij}, df_i, and $|D|$, respectively, refer to the number of appearances of word i in the document j, the number of documents containing word i and the number of documents. However, this method also is unable to capture semantic patterns, making it only useful for lexical features.

4.2.2 Non-contextualized embeddings–dense vector representation-based

Word embedding—perhaps one of the key breakthroughs for the remarkable performance of DL methods in a suit of NLP tasks, is a way of representing words (i.e., a learned representation) in a given text by allowing words with similar meaning to have a similar representation. It is this approach that generates a dense vector that carries more informative information. This can have many advantages; "one of the benefits of using dense and low-dimensional vectors is computational: the majority of neural network toolkits do not play well with very high-dimensional, sparse vectors" [86, p. 92]. And this also allows the model to generalise well. "The main benefit of the dense representations is generalisation power: if we believe some features may provide similar clues, it is worthwhile to provide a representation that is able to capture these similarities" [86, p. 92]. Contrary to these classical word representation

methods, such as BoWs that generate sparse word representations using thousands or millions of dimensions, the rationale behind the word embedding approach is premised on the idea of assigning each word a densely distributed representation (i.e., a real-valued vector), often tens or hundreds of dimensions.

Word2Vec What is clear is that methods that can capture the context of a word in a document, semantic and syntactic similarity, and relation with other words are needed. Unfortunately, those previously mentioned methods fail to capture these properties producing a sparse vector representation. To solve these issues, methods based on neural networks are proposed. Word2Vec, developed by Tomas Mikolov et al. at Google in 2013 [164] is a statistical method that leverages the use of neural networks for efficiently learning a standalone word embedding from a given text corpus. This approach is considered a de facto standard for developing pre-trained word embedding. The learned vectors by such an approach can be analysed, and interesting results can be found. "We find that these representations are surprisingly good at capturing syntactic and semantic regularities in language and that each relationship is characterised by a relation-specific vector offset. This allows vector-oriented reasoning based on the offsets between words. For example, the male/female relationship is automatically learned, and with the induced vector representations, King - Man + Woman results in a vector very close to Queen" [164]. Two different learning approaches are proposed to model the algorithm architecture: the continuous bag-of-words (CBOW) and the Skip-Gram model. The CBOW model uses the context of a current word in order to predict that word. In other words, it learns the embedding by predicting the current word based on its context. Alternatively, the continuous skip-gram model learns by predicting the surrounding words given a current word. Word2Vec approach is advantageous in that efficient and high-quality word embeddings can be learned with thankfully less space and time complexity, and this shows the key benefit of this approach where it can handle larger corpora of a text by allowing larger dimensional embeddings to be learned (more dimensions) from such corpora.

GloVe By extending the previous word embedding approach (Word2Vec), the Global Vectors for Word Representation, or GloVe for short, is developed by Pennington et al. [191] in order to learn word vectors more efficiently. This approach results in generally better word embeddings. This is because GloVe combined global statistics from matrix factorisation techniques like Latent Semantic Analysis (LSA) and local context-based learning like Word2Vec. In short, "GloVe, is a new global log-bilinear regression model for the unsupervised learning of word representations that outperforms other models on word analogy, word similarity, and named entity recognition tasks." [191].

4.2.3 Contextualized embeddings–context-aware embeddings

Bi-directional Encoder Representations from Transformers (BERT) PLMs (vector representations of words and embeddings) trained on massive amounts of textual data have formed the basis of many language-based tasks nowadays. As context-independent neural embeddings, Word2Vec and Glove extracted from shallow neural networks are examples of the most frequently pre-trained word embedding techniques prior to the advent of recent trends (PLMs shined in 2018). Yet, nevertheless, these techniques failed to capture deeper contextual relations since they mostly model indirect relationships by capturing only short-range context based on a specific co-occurrence window. In fact, since 2018, the interest of the NLP community in these kinds of pre-trained word embedding techniques have constantly been fading in favour of the most recent trend of transfer learning. Examples are Universal

Language Model Fine-Tuning (ULMFiT) [103], Embedding from Language Models (ELMo) [195], OpenAI Generative Pre-trained Transformer (GPT) [208], and Googles BERT model [56]. ULMFiT [103] pre-trained on a universal language model on a general-domain corpus which then can be fine-tuned on target task data. Radford et al. [208] generated transformer-based language models, the so-called OpenAI GPT, a unidirectional language model.

Unlike OpenAI GPT, BERT generated by Devlin et al. [56] is the first deeply bidirectional and unsupervised language representation that plays a multi-layer bidirectional transformer encoder that jointly conditions both the left and the right contexts in all layers. The transformer architecture comprises two blocks, an encoder and a decoder, to read the text and produce a prediction, respectively. BERT uses only the encoder portion of the transformer. BERT stands for Bidirectional Encoder Representation from Transformer [56] is a language representation model which Google AI introduced. Before discussing how BERT works, we first discuss some key points behind its prominent success. As the first-of-its-kind language representation, BERT contains a bunch of Transformer encoders stacked together that can be used to pre-train deep bidirectional representations. The concept of bi-directionality in BERT allows it to consider left and right contexts. In other words, BERT is based on a self-attention layer (a multi-layer bidirectional transformer encoder that (performs self-attention in both directions) jointly conditions both the left and right contexts in all layers; thus, BERT generates context-aware embeddings. This is the key differentiator between BERT and its predecessor OpenAI GPT where the former is deeply bidirectional, whereas the latter is a unidirectional pre-trained model (left-to-right language model pre-training). More details on its architecture will be introduced further next.

4.3 ML algorithms

In this subsection, we mainly describe classification models used in the literature for FND.

4.3.1 Classical ML algorithms

A plethora of ML algorithms has been explored and tested in the literature for FND.

- **Logistic Regression (LR):** LR is a statistical model applied as a great baseline algorithm on a wide range of text classification tasks.
- **Support Vector Machine (SVM):** SVM classifier is a strong classifier that yields promising results on a suit of NLP tasks.
- **Multinomial Naive Bayes (MNB):** MNB is a kind of probabilistic algorithm (a Bayesian learning approach) that is also popular and yields great results on different NLP tasks.
- **Random Forest (RF)** An ensemble of decision trees with labels predicted using a tree-like model.
- **XGBoost (XGB)** An ensemble ML algorithm. Using this method, the first model constructed using training data is used to construct a strong classifier, followed by a second model that attempts to correct the errors of the first model. The XGB algorithm uses a gradient boosting framework whose algorithm is based on decision trees.
- **Artificial Neural Network (ANN)** ANN, as the name implies, has some resemblance to the human brain. ML algorithms that have been trained using supervised learning can learn to recognise various factors that are difficult to specify using logic systems. For example, by giving the ANN a piece of text (along with other features relevant to determining the correct answer) and telling it whether or not it is fake, the ANN learns to represent the problem by updating its internal state. To begin with, ANN has a hierarchical

structure and models a problem using a learning algorithm—a concept inspired by the biological brain. The algorithm is a simplified representation of neural processing in the mammalian central nervous system: a neuron is stimulated, the weighted sum is passed through the activation function (the neuron is activated), and the neuron's output is then fed to neurons in the next layer. Although ANN can be set up in a variety of ways, they all have an input layer (a collection of input nodes or neurons), one or more hidden layers, and an output layer. An ANN is trained through two phases: (1) the feed-forward phase and (2) the backpropagation phase. The feed-forward phase is when a weighted sum of the inputs is propagated through the network and goes through the activation function of each neuron. Once the output error between the predicted output and the true output is calculated at the end of the feed-forward phase, the backpropagation phase begins by backward-propagating such an error through the network, and consequently, the weights are updated to reduce the error. The neural network is composed of a set of layers with a finite number of interconnected nodes, h, that are associated with an activation function, $a_h(\cdot)$. In the finite set of edges E, each edge connecting a node h to another node h' is associated with a weight $w_{hh'}$ that reflects the importance of the input from the previous node. The activation function a_h is then applied to a weighted sum of the values of its input nodes, according to the weights $w_{hh'}$, to determine the value v_h output by each node h.

$$v_h = a_h(\sum_{h'} w_{hh'} \cdot v'_h) \tag{4}$$

The model behaviour and capabilities are determined by the function $a_h(\cdot)$. Below is a list of some of the most commonly used ones.

Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

Tanh

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{6}$$

Rectified Linear Unit (ReLU)

$$Relu(x) = \max(0, x) \tag{7}$$

Softmax

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad for\ i = 1, 2, \ldots, K \tag{8}$$

The first activation function and the latter one [Sigmoid and Softmax]—where the latter is a generalisation of the former, are usually used in the final layer of neural network-based classifiers for modelling binary and multi-class classification tasks, respectively. These functions are used to transform the scores assigned by the model for each output class into probabilities. A Feedforward Neural Network with L layers has the following neural network architecture:

- Each input node receives an input value, which is then multiplied by, and added to, their respective weights and then passed along to nodes in the next layer in the network.
- An output layer outputs the network's decision once the hidden layers have completed their calculations. Finally, an epoch of training is completed after processing and passing the information through the neural network once, and propagating the errors back to update the weights.

– The network's output is then collected from the output layer and compared to the training dataset's ground truth. A loss function $\mathcal{L}(\hat{y}, y)$ is then used to define the difference between predicted and actual outputs, which must be minimised through multiple training epochs to obtain the most accurate set of weights.
– Following the calculation of the error, it is transmitted back through the ANN, beginning at the output node. The partial derivative of the error function with respect to each weight which is the so-called gradient is computed for each weight in each layer so as to adjust the weights $w_{hh'}$ so that the loss function is minimised. Each node's weight is modified based on the gradient's value. A learning rate determines the magnitude of the change to the internal state of the network for each learning iteration. The partial derivative is calculated according to (9). Next, a gradient is computed for each node in the ANN (starting at the output node or neuron) using the chain rule. In other words, a derivative value is calculated for a single weight by propagating it back through the activations and outputs of the network. That is, the amount by which weight in the ANN needs to be adjusted is dependent upon the gradient of the weight with respect to the node at the output of the chain whose error is to be minimised—the chain rule applied sequentially to each neuron along the way.

$$\frac{\delta \mathcal{L}(\hat{y}, y)}{\delta w_{hh'}} \tag{9}$$

To recap, ANN is a learning algorithm built for information processing through mathematical or computational models. Each connection in an ANN has a weight associated with it, so the ANN is basically a group of interconnected neurons. In order to obtain a correct prediction result, these weights can be modified iteratively throughout the classification process. ANN generally consists of three layers: an input layer that includes historical dataset information, a hidden layer that can vary depending on the application; however, only one layer is typically used, and an output layer that displays the class label.

4.3.2 DL models

Developing an automatic FND model is more important than ever, given how much data a person can curate daily. Just think of solutions for detecting and categorising social users' tweets on social media to understand the characteristics of fake and real content. Detecting fake content from text has already shown its importance in literature, where adding features extracted from text is essential for good performance in fake news classification. Automatic feature extraction by the deep neural network was the key breakthrough of impressive performance in FND in recent years. Given a text input consisting of x words, $x = \{x_1, x_2, \ldots, x_T\}$, then the main objective of the deep feed-forward neural network is to approximate some function f^* in a way so as to map an input x to a category y, $y = f^*(x)$. A feed-forward model defines the mapping $y = f(x, \theta)$ where x denotes the text input while θ refers to the learned network weights. These weights are learned in a way that achieves a good function approximation of the unknown function f^* [88, p. 164]. A deep neural network, as the name suggests, consists of many hidden layers; think of four functions $f(x)$ and a four-fold composition $f(x) = f4(f3(f2(f1(x))))$ where f_1, f_2, f_3, and f_4, respectively, refer to the first, second, third, and fourth layer. Here, all these layers are so-called hidden layers, except the last layer, which is usually about the class probabilities calculation and is often named the classification layer; in a deep neural network, simply learn some useful representations for a certain classifier. As such, the choice of representation used is not important because the essential purpose is to make the next learning task easier [88, p. 524]. Different parameters

affect the performance of a neural network, such as the initialisation of weights and biases, the activation functions used at each layer, the optimiser and a loss function. DL methods have proven their effectiveness in the field of NLP.

- **Convolutional Neural Networks (CNN)** The promising performance obtained by deep neural networks in recent years has attracted the attention of many researchers in various fields, particularly the NLP field. Among these neural network types, CNN is found to be a good candidate for processing textual data while automatically capturing local features. Traditionally used for computer vision, CNN is also being used for NLP. In computer vision, CNN identifies image features with the help of sliding windows of learned filters, while in NLP, CNN is designed to process segments of words so as to determine the most relevant word combinations for a particular task. As the name suggests, a CNN provides an output that can be used for further training by sort of matrices multiplication operation, so-called convolution. In the context of NLP, one can imagine the document or news article as a sequence of words, each of which is then represented as a real-valued word vector using any word representation method such as BERT, with which such vectors are fed as input for training a CNN model by specifying a kernel size and a number of filters. CNN has been proven effective in a suite of NLP tasks where a one-dimensional CNN (Conv1D) is usually used to generate predictions. In this case, Conv1D works by specifying a filter with a fixed size window and using that filter sliding window (each filter cell is initialised with weight) to iterate over training data, where at each step, the given input (word vectors) is multiplied by such filter weights resulting to so-called a feature map (filter output array) that encodes informative features form input training data. CNN is well known as a good candidate for automatic feature extraction and capturing local features more faithfully.
  CNN is used for learning how to distinguish documents on a suit of classification problems. As Yoav Goldberg [85] mentions in his primer on DL for NLP, neural networks, in general, are more effective than classical linear classifiers, especially when used with pre-trained word embeddings; that is, superior classification accuracy can be obtained as a result of using the nonlinearity of the network, as well as the ability to integrate pre-trained word embeddings easily. He pointed out that CNN is effective at document classification and is found strong at extracting useful local clues (salient features) in a way invariant to their places in the input sequences. Yoav Goldberg [86, p. 152], in his book, emphasises the role of CNN as a feature extractor model:
  "the CNN is, in essence, a feature-extracting architecture. It does not constitute a standalone, useful network on its own but is meant to be integrated into a larger network and trained to work in tandem with it to produce an end result. The CNN layer's responsibility is to extract meaningful sub-structures that are useful for the overall prediction task at hand".
  Thus, generally speaking, CNN is more capable of extracting features, which can successfully cut down the dimensionality of input data and increase robustness [146].
- **Recurrent Neural Networks (RNN)** A recurrent Neural Network (RNN) is a type of neural network in which the input and output include sequential data. RNN is a method widely utilised as a general paradigm for sequence modelling problems. Thanks to the specialised architecture. Different from traditional feed-forward neural networks, where data is propagated in one direction from input to output (thus cannot handle sequential data and are not well suited for sequence modelling), RNN is particularly designed to take in a sequence of inputs over time. RNN is particularly well adapted to handle events with a sequence of inputs rather than a single input, as these sequences can be propagated,

generating a single output. For example, we train a model that accepts a word sequence and predicts whether it is fake or real. RNN allows for the information to be processed over time, where the output of the RNN at each timestep relies on not only the input at the current timestep but also inputs at previous (hidden states) timesteps. RNN works as follows:
Given an input sequence $\{x_1, x_2, ..., x_t\}$ with length T, a basic RNN predicts the output sequence $\{y_1, y_2, ..., y_t\}$, and calculates the hidden layer $\{h_1, h_2, ..., h_t\}$ with a recurrent unit as follows:

$$h_t = \mathcal{H}(h_{t-1}, x_t) \tag{10}$$

where $\mathcal{H}(\cdot)$ can be thought of as an activation function or other hidden layer function that produces current hidden state h_t by taking h_{t-1} (the last hidden state) and x_t (the current input) as inputs. In the context of NLP tasks, RNN processes data by taking one word at a time and then learning linguistic-based patterns based on different series of words. Unlike regular RNN, which uses the last hidden state at the last timestep to make a decision, the attention mechanism exploits each output generated at each timestep and then processes and selectively incorporates the most salient and informative outputs according to their relevance scores. This way, RNN is able to store (retain) only useful information at each timestep while ignoring irrelevant information and then later select which outputs to use for the final decision. It is noteworthy to mention that while such regular RNN architectures were found to be effective for short sentences in most tasks, accurately remembering long sentences is still challenging.
In RNN, gradients are typically computed by applying back-propagation through time [221]. However, due to the vanishing or exploding gradients [26], with gradient-based optimisation, the basic RNN is unable to learn long-distance temporal dependencies. To solve this problem, it is possible to make an extension that includes a "memory" unit that can store information over long periods of time, commonly known as the Long Short-Term Memory (LSTM) unit [91, 99]. Another simpler RNN model is the Gated Recurrent Unit (GRU) [44].

- **Long Short-Term Memory (LSTM)** Driven by RNN architectures with long short-term memory (LSTM) units, deep neural networks have achieved state-of-the-art results in several prediction tasks. Despite RNN's remarkable success, it experiences some problems related to gradient vanishing and exploding, making it difficult to handle long sequences (fail to retain information)—meaning they break down during training and fail to model the problem well [186]. This can be attributed to the fact that basic RNN is unable to find relationships over a large number of timesteps where the input data to RNN is fed at each timestep even if they are irrelevant to the task in question, and that is a well-known problem in NLP when semantic meaning is distributed over a long sequence of words. Later, LSTM was introduced and explicitly designed to overcome this limitation [99]. LSTM cells and gated recurrent unit (GRU) [forward and backward] are RNN variations (architectures) that explicitly address this problem in a way by utilising gating functions that control the flow of information into and out of the RNN, thanks to the remarkable design. LSTM [99] is an RNN that improves on RNN flaws by altering the recurrence formula to include multiplicative and additive interactions, as well as a distinct memory state. LSTM layers can also be stacked to increase the model's complexity. To process the information flow through the cells and eliminate gradient vanishing and explosion concerns brought by RNN, LSTM has three gates: an input gate, a forget gate, and an output gate. Because of its capacity to capture long-term dependency, LSTM performs

very well. LSTM is proven effective for long sentences [254]. LSTM components can be formulated mathematically as follows [90]:

$$f_t = \sigma(W_{xf} \cdot x_t + W_{hf} \cdot h_{t-1} + b_f), \tag{11}$$

$$i_t = \sigma(W_{xi} \cdot x_t + W_{hi} \cdot h_{t-1} + b_i), \tag{12}$$

$$\tilde{C}_t = tanh(W_{xc} \cdot x_t + W_{hc} \cdot h_{t-1} + b_i), \tag{13}$$

$$C_t = f_t \odot C_{t-1} + i \odot \tilde{C}_t, \tag{14}$$

$$o_t = \sigma(W_{xo} \cdot x_t + W_{ho} \cdot h_{h-1} + b_o), \tag{15}$$

$$h_t = o_t \odot tanh(C_t). \tag{16}$$

In the formulas above, σ represents the logistic sigmoid activation function. W, b, and C_t, respectively, represent the weight matrix, the bias, and the state of the memory unit at time t.

- **Gated Recurrent Units (GRU):** A GRU variant consists of only two gates, an update gate which is a catalyst for combining forget and input gates, which decides the amount of information to be passed to the current state and a reset gate which is responsible for deciding when to ignore the previously hidden state [43]. Similar to LSTM, the update and reset gates are computed as follows [43]:

$$r_t = \delta(W_r h_{t-1} + U_r x_t + b_r), \tag{17}$$

$$z_t = \delta(W_z h_{t-1} + U_z x_t + b_z), \tag{18}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h_t}, \tag{19}$$

$$\tilde{h_t} = tanh(W_{\tilde{h_t}}(h_{t-1} \odot r_t) + U_{\tilde{h_t}} x_t). \tag{20}$$

In the formulas above, $\delta(.)$ denotes logistic sigmoid function, W and U show weight matrices of gates, h_t and b, respectively, refer to the hidden state and bias vectors. The basic RNN only considers the previous context but cannot capture future context. As such, to account for the future and preceding contexts, bidirectional LSTM (BiLSTM) and bidirectional GRU (BiGRU) are good candidates, thanks to the breakthrough design. To achieve bi-directionality, the forward and backward hidden layers are combined, which consequently controls the temporal information flow in both directions leading to better learning. Figure 4 shows the subtle difference between LSTM and GRU units. GRU and BiLSTM are proven effective and outperform CNN in many NLP tasks. For example, one study is [18], proving that classifiers based on BiLSTM and BiGRU yielded better performance than CNN. As can be noted, the major difference between LSTM and GRU is as follows: 1) unlike the LSTM unit, which encompasses three gates, the GRU unit constitutes only two gates; 2) the LSTM unit contains an internal memory unit whereas the GRU unit does not; 3) because GRUs have fewer parameters, they are easier to train than LSTMs. However, on large datasets, LSTMs produce better results. Even though BiLSTM and BiGRU have shown their superiority in a suit of NLP problems, they are not free from two shortcomings. (i) As high-dimensional input space increases, so does the complexity of these models leading to further complexity in optimising such models, and (ii) as these models can capture succeeding and proceeding contextual information (bidirectionality concept), they are not able to focus on the most salient parts of the contextual information of the text. Therefore, to overcome the former issue, the feature dimensionality can be reduced using feature selection techniques. Also, CNN can be

used to reduce the dimensionality of feature space while retaining informative features from the text. In addition, CNN can capture and extract local patterns. An attention mechanism can be used to solve the latter limitation by assigning weights [importance score] to different parts of the context of a given input text.
More importantly, such architecture selectively picks the most salient information, processes the input, and outputs the most relevant information by retaining (storing) the important and relevant input while neglecting the irrelevant and noisy data (filtering the input data). The output generated at the final timestep is typically used to form a decision in basic RNN, where the entire input sequence is encoded and captured (compressed) in a single output vector. This, of course, would lead to information loss. To capture more and more salient and informative information, an attention mechanism, as a key breakthrough and an inflection point that has led to great performance, has been introduced, which can be used to significantly boost the performance of RNN. RNN coupled with an attention mechanism results in a structure (model) that pays attention to the timesteps most critical to the task in question by focusing on certain parts of the input sequence, leading to improved performance.

- **Attention mechanism** Inspired by human biological systems, the attention paradigm has recently seen a growing interest in NLP fields. The NLP community is preparing for the paradigm shift by designing models that can assign different weights to various parts of a given input text, capturing more relevant information for further processing. The attention model aims to mimic humans' biological systems, where, given a piece of text, humans can selectively identify what is most vital and relevant in a given context while ignoring irrelevant information. The intuition behind the assumption that not all parts of input text are relevant is manifested by the so-called attention mechanism. For instance, in a machine translation task, some words may be irrelevant when translating each word. Thus, considering irrelevant information during the modelling process would cause performance degradation. Furthermore, the knowledge gathered by neural networks is stored in numerical elements, which cannot be interpreted by themselves since they are subsymbolic in nature; that is, when neural architectures produce incorrect output, it becomes difficult, if not impossible, to pinpoint the reasons [76].
Consequently, certain parts of the input can be considered adaptively by the model [40]. The attention mechanism enables the model to learn to attend to the relevant parts of the given input text (e.g., hidden states of the BiGRU network) in order to generate a single salient vector representation. As such, by attending to specific parts of the input when processing the data, the model can recognise the words that are being concentrated on for each input text. As described below, weighting different words in a sentence combines all hidden states and generates a single vector representation (See Fig. 5).

$$\alpha_t = \frac{\exp(v^T \cdot \tilde{h_t})}{\sum_t \exp(v \cdot \tilde{h_t})} \tag{21}$$

$$S_{Aw} = \sum_t \alpha_t h_t \tag{22}$$

where v is a trainable parameter [226] and the hidden states are computed in (19) and (20).
Interestingly, neuronal network behaviour can be partially explained and interpreted by attention [76], where by using attentional weights, we could observe the highlighted irrelevant parts that have been overlooked by the neural network or relevant parts of the

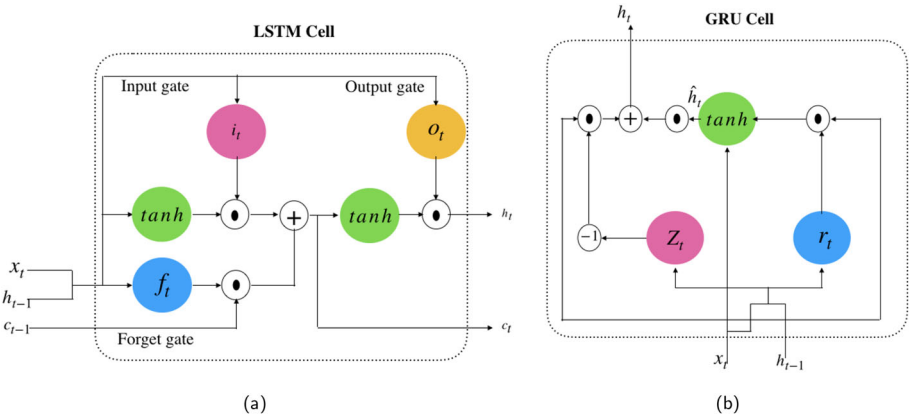

Fig. 4 Comparison of RNN units (a) LSTM cell and (b) GRU cell. Source: Adopted from [24]

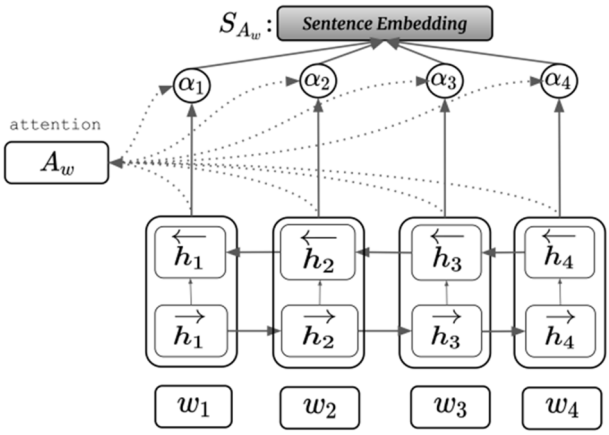

Fig. 5 Attention mechanism in a bidirectional network, directly from [226]

input text that have been focused on. All these reasons make attention a crucial component of neural architectures for NLP [77, 285].

4.3.3 Transformer-based models

Analysing existing related work, only a few studies have used PLMs to detect fake news, and little research has explored how to best harness such PLMs to detect such fake news. It becomes prohibitively challenging to process massive amounts of UGC manually. Therefore, automated systems capable of detecting fake content are essential. However, fake news on social media is a non-trivial task since fake news is written deliberately to mislead readers, and UGC is typically of poor quality. To address these challenges, researchers proposed various methods for interpreting the meaning of a word through embedding vectors. Neural network-based methods such as Word2Vec and GloVe are commonly used to learn word embeddings from large word corpora. These embedding models have the disadvantage of being context-free since context is neglected, and static embeddings for words are generated regardless

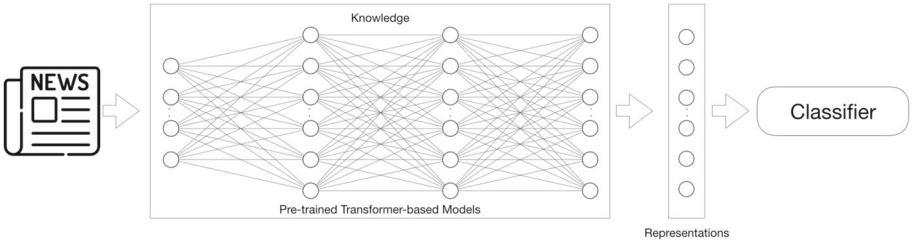

Fig. 6 The structure of PLMs models

of their contexts. To achieve finer-grained performance, a model must be able to capture semantic and contextual patterns. Moreover, an ML or DL model can automatically extract semantic information from a given input to detect fake content, but they cannot accurately recognise fake content without a deep understanding of the text. There has been a growing interest in the attention paradigm in recent years.

There is an overall paradigm shift taking place in the NLP community, which aims to develop a set of models that not only improve accuracy but also address the problem of lacking labelled data, which has been a long-standing issue in the scientific community. In addition, there is an urgent need to detect fake news automatically; however, this is a challenging task since existing ML and DL models (prior to the advent of transformer models) fail to provide a deeper semantic understanding of text input. This has caused NLP research to make great strides by introducing pre-trained transformer-based language models. Using PLMs trained on massive unlabeled data for text classification tasks is becoming increasingly popular. To adapt for the downstream task, new neural network layers are layered on top of the pre-trained layers in the PLM [95]. As seen in Fig. 6, a fully contented layer (FC) is added on top of the PLMs for classification. The adoption of PLMs within a transfer learning framework facilitates the utilisation of their acquired knowledge (referred to as knowledge in Fig. 6). This knowledge can be effectively leveraged to enhance performance on specific tasks by employing techniques such as fine-tuning or feature extraction (represented as Representations). Subsequently, a classifier (depicted as Classifier) can be applied to these representations to accomplish the desired task objectives. A sophisticated approach is needed to detect fake news since it has become increasingly difficult to distinguish between fake and real content.

This section introduces three PLMs: BERT [57], DistilBERT [225] and RoBERTa [148] that have been considered the key breakthroughs of the impressive performance on a suite of NLP tasks largely due to their powerful language representations being learned from massive amounts of a text corpus. Such models can be easily fine-tuned on a specific downstream task through what is so-called transfer learning.

- **BERT:** BERT—stands for Bidirectional Encoder Representation from Transformer, introduced by Devlin et al. [56], is the first deeply bidirectional and unsupervised language representation that plays a multi-layer bidirectional transformer encoder (performs self-attention in both directions) that jointly conditions both the left and the right contexts in all layers. Thus, BERT generates context-aware embeddings. Furthermore, to remove the unidirectionality constraint, BERT performs pre-training using an unsupervised prediction task, including a masked language model (MLM) that is responsible for understanding context and making predictions (of words). Thus, the model can produce a vector representation that can capture the general information of the input text.

These semantic representations of each word in the input text can be improved using an attention mechanism in the sense that different words in a context show different effects in boosting semantical representation. As a core component of transformer architecture, the attention mechanism's underlying role is to assign less or more weights to different parts of text towards the output (i.e., differentiate the contribution of different parts of the input on the output). Attention can be considered as a function that maps queries and follows key-value and output vector pairs; the scaled dot-product attention formula can be seen in (23).

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{23}$$

where Q, K, and V, respectively, denote the query, key, and value itself. $\sqrt{d_k}$ denotes the dimension of the key vector k and query vector q. Attention uses a Softmax activation function that normalises the inputs to a value between 0 and 1. BERT uses a multi-head attention mechanism (since BERT uses a transformer's encoder), which can be seen in (24) where each specific head and the associated weight matrices are denoted with the subscript i.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{24}$$

where each $head_i$ is calculated as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{25}$$

Due to the incapability of vector space methods such as CV and TF-IDF to take context into account, using these representations with ML classifiers rely on the appearance of tokens in making final decisions, regardless of their context. These vector space models are ineffective at capturing deeper semantics and contextual patterns, specifically those contained in UGC (e.g., tweets). A major advantage of BERT (and its variations) in the case of Twitter (where UGC often contains misspellings, noise, and abbreviations) is the use of sub-tokens rather than fixed tokens; it is thus ideal for use with such data [101] instead of standard context-independent word embeddings. Although the BERT model has made a great breakthrough in text classification, it is computationally expensive as it contains millions of parameters (i.e., BERT_{BASE} contains 110 million parameters while BERT_{LARGE} has 340 million parameters) [56].

- **DistilBERT** [225]**:** Even though BERT is more complex to train (depending on how large a number of parameters are being used), a variation of BERT, so-called DistilBERT, provides a simpler and reasonable number of parameters compared to that of BERT (reducing BERT by 40% in size while retaining 97% of its language understanding abilities), thus, faster training (60% faster).
- **RoBERTa** [147]**:** Stands for the Robustly optimized BERT approach, which Facebook introduced. It is simply retraining of BERT with improved training methodology (i) by removing the Next Sentence Prediction task from the pre-training process; (ii) RoBERTa was trained over ten times more data, and (iii) by introducing dynamic masking using larger batch sizes so that the masked token changes during the training rather than the static masking pattern used in BERT. Thus, RoBERTa introduces a different pre-training approach to BERT.

The following section will elucidate the common evaluation metrics extensively employed in previous research endeavours to assess the efficacy and performance of FND approaches.

4.4 Evaluation metrics

The last stage of training a text classification model is performance evaluation. The five evaluation criteria are extensively used in text classification tasks, namely, accuracy, precision, recall, and F1 measure (calculated as in Equations below), to assess the performance of the models.

- Accuracy (A): a measure of the classifier's ability to correctly classify a piece of information as either fake or real. The Accuracy can be estimated using (26).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

- Precsion (P): is a measure for the classifier exactness such that a low value indicates large number of False Positives. The precsion represents the number of positive predictions divided by the total number of positive class values predicted and is calculated using (27).

$$Precision = \frac{TP}{TP + FP} \tag{27}$$

- Recall (R): is considered as a measure of a classifier completeness (i.e. a low value of recall indicates many False Negatives) where the number of True Positives is divided by the number of True Positives and the number of False Negatives, as can be clearly seen in (28).

$$Recall = \frac{TP}{TP + FN} \tag{28}$$

- F1 score (F1): is calculated as the weighted harmonic mean of the precision and recall measures of the classifier using (29).

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{29}$$

where TP,TN,FP, and FN, respectively, are True Positive, True Negative, False Positive, and False Negative.

5 Current challenges

5.1 ML algorithms

Classical ML algorithms, such as LR, SVM, etc., are easy to comprehend and perform well on small datasets, but they (i) require complex feature engineering, (ii) fail to capture substantial semantical contextual knowledge for a specific input text and (iii) limited generalisation abilities. DL techniques perform better when processing large-scale (and complex) datasets than traditional ML algorithms, which plateau as datasets grow. DL techniques can generally reduce the time and effort required for feature extraction, leading to more accurate prediction results. Classical ML algorithms, including shallow neural networks, lack an in-depth understanding of a given input text, which increases the risk of overfitting. DL techniques can uncover complex patterns in large datasets by composing more complex and abstract

representations from simpler ones. By building high-level features from low-level ones, DL models can efficiently and hierarchically model complex functions or learn representations. Therefore, the ability to capture deep representations is more robust and helps achieve good generalisation.

5.2 DL models

As opposed to traditional ML models, which require human experts to (manually) encode domain knowledge through feature engineering, which is inefficient and impractical, DL models are able to learn relevant and important feature representations automatically, making them particularly well suited for NLP tasks. More specifically, DL-based techniques such as CNN and RNN-based methods are well suited for complicated classification problems, powered by massive data and can learn more complicated (latent) features. However, CNN typically struggles with capturing long-term contextual dependencies, while RNN-based methods perform sub-optimally in handling such dependencies. As such, a combination of these two architectures may be able to overcome some of their inherent limitations. However, besides the fact that surface-level features cannot effectively capture semantical patterns in text, the lack of a sufficient amount of data constitutes a bottleneck for DL models. Furthermore, these models are impotent towards capturing deep semantical contextualised understanding of a given input text. Thus, to address this, the power of the advanced transformer-based models, such as BERT and its variations, have proven effective in capturing deep contextualised patterns of a given input text and can be effectively leveraged to build robust fake news predictive models, even with small amounts of data. Moreover, efforts are dedicated by researchers in the field of NLP to detect and combat fake news using an assortment of ML and DL algorithms. However, research on FND has been mostly restricted to limited comparisons of deep and ML models using particular untested combinations of word representation methods on specific datasets in order to obtain better detection results. The question of which ML, DL, and other advanced transformer-based methods are the most effective for FND remains unanswered. A benchmark study that tests different models with different sets of text representation methods using different datasets is needed.

5.3 Multimodality

Despite the success of ML and DL models that relied largely on textual content for FND, textual content is insufficient on its own. Besides content-based clues, little research on exploiting visual information, which of course, shows the superiority of multimodal approaches over unimodal approaches, has been established to combat fake news dissemination. News on social media is disseminated in various forms, and the originators of fake news usually create and manipulate their (fake) content by incorporating visual information (e.g., images and videos) to attract readers' attention. This indicates that multimedia content on social media plays an important role in discriminating fake news from real. Inspired by the intuition behind the idea that fusing various modalities would help uncover different useful aspects of news where such modalities could contain complementary information for detecting the news authenticity [244], developing an effective multimodal framework that is capable of not only harnessing visual information but also capturing semantical contextualised relations

between different modalities (i.e., text, image, user behavioural information) is needed for a finer-detection performance.

5.4 Transparency

Driven by the high performance achieved by most of today's models, they often lack transparency as they seemingly follow a black box nature that provides results that are obscure to humans. Thus, models are required to not only deliver the highest possible performance but also provide interpretations and explanations of how the decisions are made.

6 Limitations

This study has some limitations that should be acknowledged:

- Lack of empirical analysis: The study primarily focuses on providing an extensive review of previous studies in the field of FND. However, it does not include original empirical analysis or experiments to validate the effectiveness of the discussed methods and techniques.
- Limited discussion on real-world implementation: Although the study extensively discusses various methods and models used in the literature, there is a lack of in-depth analysis of the practical implementation and deployment of these techniques in real-world scenarios. Further exploration of the challenges and considerations associated with applying these methods to actual systems is necessary.

Despite these limitations, the study provides a comprehensive overview of FND research, including definitions, related terms, psychological insights, feature-based methods, ML and DL techniques, and transformer-based models. It serves as a valuable resource for researchers in the field and offers insights into the current state of FND.

7 Potential directions for future research

Potential directions for future research based on the current challenges are as follows:

- Exploration of novel ML algorithms designed specifically for FND that address the limitations of classical ML algorithms. This can involve developing algorithms that reduce the reliance on complex feature engineering and can effectively capture substantial semantical contextual knowledge for a specific input text.
- Investigation of techniques that enhance the generalization abilities of ML algorithms, especially when processing large-scale and complex datasets. This can include incorporating transfer learning, ensemble methods, or meta-learning approaches to improve the performance and scalability of ML-based FND models.
- Development of hybrid architectures that combine the strengths of CNN and RNN-based models to overcome their individual limitations. This can involve exploring techniques that effectively capture both short-term and long-term contextual dependencies in text, leading to improved performance in FND tasks.

- Further research on transformer-based models, such as BERT and its variations, to leverage their deep contextualized understanding of the text. This includes investigating the application of pre-trained transformer models for FND and exploring methods to effectively fine-tune these models with limited amounts of data.
- Conducting benchmark studies to compare the effectiveness of ML, DL, and other advanced transformer-based methods for FND. These studies should consider different text representation methods, utilise diverse datasets, and provide comprehensive evaluations to determine the most effective approaches.
- Development of effective multimodal frameworks that leverage visual information, in addition to textual content, for FND. This involves exploring techniques to integrate and model the relationships between different modalities, such as text, images, and user behavioural information, to achieve better detection performance.
- Investigation of feature fusion techniques and DL architectures that can effectively leverage multimodal information to identify and distinguish between real and fake news. This research can focus on developing models that capture complementary information from different modalities and exploit their combined power for more accurate detection.
- Exploration of methods to balance performance and transparency ensures that FND models deliver high accuracy and provide understandable justifications for their predictions. This can include the development of post-hoc explanation techniques or model-agnostic interpretability methods that can be applied to various FND models.

8 Conclusion

In conclusion, this paper presents a comprehensive survey and analysis of the research efforts in the field of automatic FND. The study encompasses a wide range of topics, including the definitions and related terms of fake news, psychological and scientific theories explaining its dissemination, advanced ML and DL techniques used in FND, characteristics of fake news, commonly used datasets, methodologies employed in existing studies, and identified research challenges.

The key contributions of this paper are as follows:

1. In-depth understanding: By summarizing different fake news definitions, related terms, and psychological and scientific theories, this study enhances our understanding of the reasons behind the belief in and dissemination of fake news. It provides a comprehensive overview of the factors contributing to its proliferation.
2. ML and DL techniques: Through an extensive survey, this paper highlights the advancements in ML and DL techniques used in FND. It covers various feature-based methods and transformer-based models, shedding light on their effectiveness in characterizing and identifying fake news content.
3. Fake news characteristics and datasets: The paper summarizes the characteristics of fake news and commonly used datasets, providing researchers with valuable insights into the nature of fake news and facilitating the evaluation and comparison of different detection methods.
4. Methodologies and challenges: This paper outlines the current landscape of FND research by discussing the methodologies employed in existing studies. It also identifies the challenges faced in the field, such as the rapid evolution of fake news tactics and the need for more robust and scalable detection mechanisms.

Appendix

Table 9 Acronyms and their explanation

Acronym	Explanation
FND	Fake News Detection
ML	Machine Learning
DL	Deep Learning
DNN	Deep Neural Networks
MLP	Multi-Layer Perception
NLP	Natural Language Processing
SVM	Support Vector Machines
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
XGB	eXtreme Gradient Boosting
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
LSTM	Long Shor-Term Memory
BiLSTM	Bidirectional LSTM
GRU	Gated Recurrent Units
BiGRU	Bidirectional GRUs
SOTA	State-Of-The-Art
PLMs	Pre-trained Language Models
UGC	User-generated Content
OSNs	Online Social Network
TF-IDF	Term Frequency Inverse Term Frequency
CV	Count Vectorizer
BERT	Bidirectional Representations from Transformers
ELMo	Embeddings from Language Models
POS	Part of Speech
LDA	Latent Dirichlet Allocation

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This study was not funded by any organization.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study. All discussed datasets are referenced in this article.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. AGE JITD (2022) Journalism in the digital age, the echo chamber effect. https://cs181journalism2015.weebly.com/the-echo-chamber-effect.html. Accessed 12 Feb 2022
2. Aggarwal A, Chauhan A, Kumar D, Mittal M, Verma S (2020) Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Trans Scalable Inform Syst 7
3. Ahmad I, Yousaf M, Yousaf S, Ahmad MO (2020) Fake news detection using machine learning ensemble methods. Complexity 2020
4. Ahmed H, Traore I, Saad S (2017) Detection of online fake news using n-gram analysis and machine learning techniques. In: International conference on intelligent, secure, and dependable systems in distributed and cloud environments. Springer, pp 127–138
5. Alghamdi J, Lin Y, Luo S (2022a) A comparative study of machine learning and deep learning techniques for fake news detection. Information 13. https://www.mdpi.com/2078-2489/13/12/576, https://doi.org/10.3390/info13120576
6. Alghamdi J, Lin Y, Luo S (2022b) Modeling fake news detection using bert-cnn-bilstm architecture. In: 2022 IEEE 5th international conference on multimedia information processing and retrieval (MIPR), pp 354–357. https://doi.org/10.1109/MIPR54900.2022.00069
7. Alghamdi J, Lin Y, Luo S (2022c) Towards fake news detection on social media. In: 2022 21st IEEE international conference on machine learning and applications (ICMLA), pp 148–153. https://doi.org/10.1109/ICMLA55696.2022.00028
8. Alghamdi J, Lin Y, Luo S (2023a) Does context matter? effective deep learning approaches to curb fake news dissemination on social media. App Sci 13. https://www.mdpi.com/2076-3417/13/5/3345, https://doi.org/10.3390/app13053345
9. Alghamdi J, Lin Y, Luo S (2023b) Towards covid-19 fake news detection using transformer-based models. Knowledge-Based Syst 274:110642. https://www.sciencedirect.com/science/article/pii/S0950705123003921, https://doi.org/10.1016/j.knosys.2023.110642
10. Alhindi T, Petridis S, Muresan S (2018) Where is your evidence: improving fact-checking by justification modeling. In: Proceedings of the first workshop on fact extraction and VERification (FEVER), pp 85–90
11. Ali M, Levine T (2008) The language of truthful and deceptive denials and confessions. Commu Rep 21:82–91
12. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31:211–36
13. Allcott H, Gentzkow M, Yu C (2019) Trends in the diffusion of misinformation on social media. Res Polit 6:2053168019848554. https://doi.org/10.1177/2053168019848554
14. Arun AS, Subhash VG Shridevi S (2022) Fake news detection in mainstream media using bert, In: Computational methods and data engineering: proceedings of ICCMDE 2021. Springer, pp 379–390
15. Asch SE (1961) Effects of group pressure upon the modification and distortion of judgments. In: Documents of gestalt psychology. University of California Press, pp 222–236
16. Ashforth BE, Mael F (1989) Social identity theory and the organization. Acad Manag Rev 14:20–39
17. Azzimonti M, Fernandes M (2018) Social media networks, fake news, and polarization. NBER Working Papers 24462. National Bureau of Economic Research Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:24462
18. Bajaj S (2017) "The pope has a new baby! " fake news detection using deep learning
19. Bakir V, McStay A (2018) Fake news and the economy of emotions. Digit J 6:154–175. https://doi.org/10.1080/21670811.2017.1345645
20. Bali APS, Fernandes M, Choubey S, Goel M (2019) Comparative performance of machine learning algorithms for fake news detection. In: Singh M, Gupta P, Tyagi V, Flusser J, Ören T, Kashyap R (eds) Advances in computing and data sciences. Springer, Singapore, pp 420–430
21. Bálint P, Bálint G (2009) The semmelweis-reflex. Orv Hetil 150:1430–1430

22. Bandyopadhyay S, Dutta S (2020) The analysis of fake news in social medias for four months during lockdown in covid-19-a study: biostatistical analysis of covid-19. Xeno J Biomed Sci 1:1–6
23. Baruah A, Das KA, Barbhuiya FA, Dey K (2020) Automatic detection of fake news spreaders using bert. In: CLEF (working notes)
24. Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR (2021) Abcdm: an attention-based bidirectional cnn-rnn deep model for sentiment analysis. Futur Gener Comput Syst 115:279–294. https://www.sciencedirect.com/science/article/pii/S0167739X20309195, https://doi.org/10.1016/j.future.2020.08.005
25. Basu S (1997) The conservatism principle and the asymmetric timeliness of earnings1. J Account Econ 24:3–37
26. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5:157–166
27. Bergmann JR (1993) Discreet indiscretions: the social organization of gossip. HOEPLI EDITORE
28. Bessi A, Ferrara E (2016) Social bots distort the 2016 us presidential election online discussion. First Monday 21
29. Bharadwaj P, Shao Z (2019) Fake news detection with semantic features and text mining. Int J Nat Language Comput (IJNLC) 8
30. Bode L, Vraga EK (2015) In related news, that was wrong: the correction of misinformation through related stories functionality in social media. J Commun 65:619–638
31. Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inf Sci 497:38–55
32. Buckner HT (1965) A theory of rumor transmission. Public Opin Q 29:54–70
33. Burgoon JK, Blair JP, Qin T, Nunamaker JF (2003) Detecting deception through linguistic analysis. In: International conference on intelligence and security informatics. Springer, pp 91–101
34. Carson J (2018) Fake news: what exactly is it–and how can you spot it? Telegraph 28
35. Castelo S, Almeida T, Elghafari A, Santos A, Pham K, Nakamura E, Freire J (2019) A topic-agnostic approach for identifying fake news pages. In: Companion proceedings of the 2019 world wide web conference, pp 975–980
36. Castillo C, Mendoza M, Poblete B (2011a) Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp 675–684
37. Castillo C, Mendoza M, Poblete B (2011b) Information credibility on twitter. In: Proceedings of the 20th international conference on world wide web. Association for Computing Machinery, New York, pp 675–684. https://doi.org/10.1145/1963405.1963500
38. Centre FH (2022) How is facebook addressing false information through independent fact-checkers? | facebook help centre. https://www.facebook.com/help/1952307158131536. Accessed 2 March 2022
39. Chang C, Zhang Y, Szabo C, Sheng QZ (2016) Extreme user and political rumor detection on twitter. In: International conference on advanced data mining and applications. Springer, pp 751–763
40. Chaudhari S, Mithal V, Polatkan G, Ramanath R (2021) An attentive survey of attention models. ACM Trans Intell Sys Techno (TIST) 12:1–32
41. Chen W, Yeo CK, Lau CT, Lee BS (2016) Behavior deviation: an anomaly detection view of rumor preemption. In: 2016 IEEE 7th annual information technology, electronics and mobile communication conference (IEMCON). IEEE, pp 1–7
42. Chen Y, Conroy NJ, Rubin VL (2015) Misleading online content: recognizing clickbait as “false news". In: Proceedings of the 2015 ACM on workshop on multimodal deception detection, pp 15–19
43. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014a) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078
44. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014b) On the properties of neural machine translation: encoder-decoder approaches. arXiv:1409.1259
45. Chua AY, Banerjee S (2016) Linguistic predictors of rumor veracity on the internet. In: Proceedings of the international multiconference of engineers and computer scientists, pp 387–391
46. Cohen M (2017) Fake news and manipulated data, the new gdpr, and the future of information. Bus Inf Rev 34:81–85
47. Coleman K (2022) Introducing birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation. Accessed 15 Feb 2022
48. Conroy NK, Rubin VL, Chen Y (2015) Automatic deception detection: methods for finding fake news. Proc Ass Inform Sci Tech 52:1–4
49. Cui L, Wang S, Lee D (2019) Same: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. Association for Computing Machinery, New York, pp 41–48

50. Cunha E, Magno G, Caetano J, Teixeira D, Almeida V (2018) Fake news as we feel it: perception and conceptualization of the term "fake news" in the media. In: International conference on social informatics. Springer, pp 151–166
51. Damashek M (1995) Gauging similarity with n-grams: language-independent categorization of text. Sci 267:843–848
52. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci 113:554–559
53. Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W (2016) Echo chambers: emotional contagion and group polarization on facebook. Sci Rep 6:1–12
54. Derczynski L, Bontcheva K, Lukasik M, Declerck T, Scharl A, Georgiev G, Osenova P, Lobo TP, Kolliakou A, Stewart R, Terp SJ, Wong G, Burger C, Zubiaga A, Procter R, Liakata M (2014) Pheme: computing veracity: the fourth challenge of big social data. In: European semantic web conference ESWC. http://wrap.warwick.ac.uk/71304/
55. Deutsch M, Gerard HB (1955) A study of normative and informational social influences upon individual judgment. J Abnorm Soc Psychol 51:629
56. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
57. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
58. Dictionary C (2022) Collins 2017 word of the year shortlist. https://blog.collinsdictionary.com/language-lovers/collins-2017-word-of-the-year-shortlist/. Accessed 6 Feb 2022
59. DiFranzo D, Gloria-Garcia K (2017) Filter bubbles and fake news. XRDS 23:32–35. https://doi.org/10.1145/3055153
60. Dizikes P (2023) Study: on twitter, false news travels faster than true stories. http://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308,2018
61. Domonoske C (2016) Students have 'dismaying' inability to tell fake news from real, study finds
62. Dou Y, Shu K, Xia C, Yu PS, Sun L (2021) User preference-aware fake news detection. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pp 2051–2055
63. Edkins B (2016) Americans believe they can detect fake news. Studies show they can't
64. Elhadad MK, Li KF, Gebali F (2019) A novel approach for selecting hybrid features from online news textual metadata for fake news detection. In: International conference on P2P, parallel, grid, cloud and internet computing. Springer, pp 914–925
65. Enayet O, El-Beltagy SR (2017) Niletmrg at semeval-2017 task 8: determining rumour and veracity support for rumours on twitter. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp 470–474
66. Fallis D (2009) A conceptual analysis of disinformation, iConference
67. Fallis D (2014) A functional analysis of disinformation. iConference 2014 Proceedings
68. Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 french presidential election. First Monday. https://doi.org/10.5210/fm.v22i8.8005
69. Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59:96–104
70. Fetzer JH (2004) Disinformation: the use of false information. Minds Mach 14:231–240
71. Fisher RJ (1993) Social desirability bias and the validity of indirect questioning. J Consum Res 20:303–315
72. Freedman JL, Sears DO (1965) Selective exposure. In: Advances in experimental social psychology. Elsevier, vol 2, pp 57–97
73. Friggeri A, Adamic LA, Eckles D, Cheng J (2014) Rumor cascades. In: ICWSM
74. Fuller CM, Biros DP, Wilson RL (2009) Decision support for determining veracity via linguistic-based cues. Decis Support Syst 46:695–703
75. Funke D (2017) A satirical fake news site apologized for making a story too real. Poynter. Retrieved from https://www.poynter.org/news/satirical-fake-news-site-apologized-making-story-too-real
76. Galassi A, Lippi M, Torroni P (2021) Attention in natural language processing. IEEE Trans Neural Netw Learn Syst 32:4291–4308. https://doi.org/10.1109/TNNLS.2020.3019893
77. Gatt A, Krahmer E (2018) Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J Artif Intell Res 61:65–170
78. Gautam AVV, Masud S (2021) Fake news detection system using xlnet model with topic distributions: constraint@aaai2021 shared task. arXiv:2101.11425, https://doi.org/10.48550/ARXIV.2101.11425
79. Gelfert A (2013) Rumor, gossip, and conspiracy theories: pathologies of testimony and the principle of publicity. In: Rumor and communication in Asia in the internet age. Routledge, pp 34–59

80. Gelfert A (2018) Fake news: a definition. Inform Log 38:84–117
81. Giachanou A, Zhang G, Rosso P (2020) Multimodal multi-image fake news detection. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). IEEE, pp 647–654
82. Girgis S, Amer E, Gadallah M (2018) Deep learning algorithms for detecting fake news in online text. In: 2018 13th international conference on computer engineering and systems (ICCES). IEEE, pp 93–97
83. Golbeck J, Mauriello M, Auxier B, Bhanushali KH, Bonk C, Bouzaghrane MA, Buntain C, Chanduka R, Cheakalos P, Everett JB et al (2018) Fake news vs satire: a dataset and analysis. In: Proceedings of the 10th ACM conference on web science, pp 17–21
84. Goldani MH, Momtazi S, Safabakhsh R (2021) Detecting fake news with capsule neural networks. Appl Soft Comput 101:106991
85. Goldberg Y (2016) A primer on neural network models for natural language processing. J Artif Intell Res 57:345–420
86. Goldberg Y (2017) Neural network methods for natural language processing. Synth Lect Human Lang Technol 10:1–309
87. Goldman AI (2008) The social epistemology of blogging. Cambridge University Press. Cambridge Studies in Philosophy and Public Policy, pp 111–122. https://doi.org/10.1017/CBO9780511498725.007
88. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT press
89. Gravanis G, Vakali A, Diamantaras K, Karadais P (2019) Behind the cues: a benchmarking study for fake news detection. Expert Syst Appl 128:201–213
90. Graves A (2012) Supervised sequence labelling. In: Supervised sequence labelling with recurrent neural networks. Springer, pp 5–13
91. Graves A (2013) Generating sequences with recurrent neural networks. arXiv:1308.0850
92. Gupta A, Kumaraguru P (2012) Twitter explodes with activity in Mumbai blasts! a lifeline or an unmonitored daemon in the lurking? Technical Report
93. Gupta A, Lamba H, Kumaraguru P, Joshi A (2013) Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on world wide web, pp 729–736
94. Gupta M, Zhao P, Han J (2012) Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM international conference on data mining. SIAM, pp 153–164
95. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: adapt language models to domains and tasks. arXiv:2004.10964
96. Hancock JT, Curry LE, Goorha S, Woodworth M (2007) On lying and being lied to: s linguistic analysis of deception in computer-mediated communication. Discourse Process 45:1–23
97. Harris ZS (1954) Distributional structure. Word 10:146–162
98. Hernon P (1995) Disinformation and misinformation through the internet: findings of an exploratory study. Gov Inf Q 12:133–139
99. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
100. Horne B, Adali S (2017) This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Proceedings of the international AAAI conference on web and social media, pp 759–766
101. Horne L, Matti M, Pourjafar P, Wang Z (2020) GRUBERT: a GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis. In: Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing: student research workshop. Association for Computational Linguistics, Suzhou, pp 130–138. https://aclanthology.org/2020.aacl-srw.19
102. Hosseinimotlagh S, Papalexakis EE (2018) Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In: Proceedings of the workshop on misinformation and misbehavior mining on the web (MIS2)
103. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv:1801.06146
104. Howard PN, Kollanyi B, Bradshaw S, Neudert LM (2018) Social media, news and political information during the us election: was polarizing content concentrated in swing states? arXiv:1802.03573
105. Hu X, Tang J, Zhang Y, Liu H (2013) Social spammer detection in microblogging. In: IJCAI 2013 - proceedings of the 23rd international joint conference on artificial intelligence, pp 2633–2639. 23rd international joint conference on artificial intelligence, IJCAI 2013; conference date: 03-08-2013 through 09-08-2013
106. HUFFPOST (2022) 1,000 paid russian trolls spread fake news on hillary clinton, senate intelligence heads told. https://www.huffpost.com/entry/russian-trolls-fake-news_n_58dde6bae4b08194e3b8d5c4. Accessed 1 March 2022

107. INDEPENDENT (2022) Twitter to delete 6% of all accounts in huge cull. https://www.independent.co.uk/tech/twitter-fake-followers-lost-delete-accounts-cull-a8444236.html. Accessed 5 March 2022
108. Institute OI (2022) Resource for understanding political bots. https://www.oii.ox.ac.uk/news-events/news/resource-for-understanding-political-bots/. Accessed 22 Feb 2022
109. Jain S, Sharma V, Kaushal R (2016) Towards automated real-time detection of misinformation on twitter. 2016 international conference on advances in computing, communications and informatics (ICACCI), pp 2015–2020
110. Jamieson KH, Cappella JN (2008) Echo chamber: rush Limbaugh and the conservative media establishment. Oxford University Press
111. Jin Z, Cao J, Guo H, Zhang Y, Luo J (2017a) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. Proceedings of the 25th ACM international conference on multimedia
112. Jin Z, Cao J, Zhang Y, Luo J (2016) News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the AAAI conference on artificial intelligence
113. Jin Z, Cao J, Zhang Y, Zhou J, Tian Q (2017) Novel visual and statistical image features for microblogs news verification. IEEE Trans Multimed 19:598–608
114. Joachims T (1996) A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Technical Report. Carnegie-mellon univ pittsburgh pa dept of computer science
115. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: European conference on machine learning. Springer, pp 137–142
116. Johnson MK, Raye CL (1981) Reality monitoring. Psychol Rev 88:67
117. Jones EE, McGillis D (1976) Correspondent inferences and the attribution cube: a comparative reappraisal. New directions in attribution research 1:389–420
118. Jones M, Craddock PT, Barker N et al (1990) Fake?: the art of deception. Univ of California Press
119. Jr ECT, Lim ZW, Ling R (2018) Defining "fake news". Digit J 6:137–153. https://doi.org/10.1080/21670811.2017.1360143
120. Jwa H, Oh D, Park K, Kang JM, Lim H (2019) exbake: automatic fake news detection model based on bidirectional encoder representations from transformers (bert). App Sci 9:4062
121. Kahneman D, Tversky A (2013) Prospect theory: an analysis of decision under risk, In: Handbook of the fundamentals of financial decision making: Part I. World Scientific, pp 99–127
122. Kapferer JN (2017) Rumors: uses, interpretation and necessity. Routledge
123. Khan JY, Khondaker MTI, Afroz S, Uddin G, Iqbal A (2021) A benchmark study of machine learning models for online fake news detection. Mach Learn App 4:100032
124. Khattar D, Goud JS, Gupta M, Varma V (2019) Mvae: multimodal variational autoencoder for fake news detection. In: The world wide web conference, pp 2915–2921
125. Klein DO, Wueller JR (2018) Fake news: a legal perspective. Australasian Policing 10
126. Knapp ML, Hart RP, Dennis HS (1974) An exploration of deception as a communication construct. Hum Commun Res 1:15–29
127. Koloski B, Stepišnik-Perdih T, Robnik-Šikonja M, Pollak S, Škrlj B (2021) Knowledge graph informed fake news classification via heterogeneous representation ensembles. arXiv:2110.10457
128. Kshetri N, Voas J (2017) The economics of "fake news". IT Professional 19:8–12
129. Kucharski A (2016) Study epidemiology of fake news. Nature 540:525–525
130. Kula S, Choraś M, Kozik R (2019) Application of the bert-based architecture in fake news detection. In: Computational intelligence in security for information systems conference. Springer, pp 239–249
131. Kumar S, Shah N (2018) False information on web and social media: a survey. arXiv:1804.08559
132. Kumar S, West R, Leskovec J (2016a) Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes. International World Wide Web Conferences Steering Committee. Republic and Canton of Geneva, CHE. pp 591–602. https://doi.org/10.1145/2872427.2883085
133. Kumar S, West R, Leskovec J (2016b) Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes. In: Proceedings of the 25th international conference on world wide web, pp 591–602
134. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web. Association for Computing Machinery, New York, pp 591–600. https://doi.org/10.1145/1772690.1772751
135. Kwon S, Cha M, Jung K (2017) Rumor detection over varying time windows. PLOS ONE 12:1–19. https://doi.org/10.1371/journal.pone.0168344
136. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013a) Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th international conference on data mining, pp 1103–1108. https://doi.org/10.1109/ICDM.2013.61
137. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013b) Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th international conference on data mining. IEEE, pp 1103–1108

138. Langin K (2018) Fake news spreads faster than true news on twitter—thanks to people, not bots. Science Magazine
139. Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D et al (2018) The science of fake news. Science 359:1094–1096
140. Lazer D, Baum MA, BYea (2018) The science of fake news. arXiv:1708.01967
141. Lendvai P, Reichel UD (2016) Contradiction detection for rumorous claims. arXiv:1611.02588
142. Li C, Liu S (2018) A comparative study of the class imbalance problem in twitter spam detection. Concurr Comput Pract Experience 30:e4281. https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4281, https://doi.org/10.1002/cpe.4281, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4281. e4281 cpe.4281
143. Li Q, Liu X, Fang R, Nourbakhsh A, Shah S (2016) User behaviors in newsworthy rumors: a case study of twitter. In: Proceedings of the international AAAI conference on web and social media
144. Li Q, Zhang Q, Si L (2019) eventai at semeval-2019 task 7: rumor detection on social media by exploiting content, user credibility and propagation information. In: Proceedings of the 13th international workshop on semantic evaluation, pp 855–859
145. Lilleker D (2017) Evidence to the culture, media and sport committee'fake news' inquiry presented by the faculty for media & communication, bournemouth university
146. Liu G, Guo J (2019) Bidirectional lstm with attention mechanism and convolutional layer for text classification. Neurocomputing 337:325–338
147. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019a) Roberta: a robustly optimized bert pretraining approach. arXiv:1907.11692
148. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019b) Roberta: a robustly optimized bert pretraining approach
149. Liu Y, Wu YF (2018) Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Proceedings of the AAAI conference on artificial intelligence
150. Long Y, Lu Q, Xiang R, Li M, Huang CR (2017) Fake news detection through multi-perspective speaker profiles. In: Proceedings of the eighth international joint conference on natural language processing, pp 252–256
151. Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong KF, Cha M (2016) Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence. AAAI Press, pp 3818–3824
152. Ma J, Gao W, Wei Z, Lu Y, Wong KF (2015) Detect rumors using time series of social context information on microblogging websites. In: Proceedings of the 24th ACM international on conference on information and knowledge management. Association for Computing Machinery, New York, pp 1751–1754. https://doi.org/10.1145/2806416.2806607
153. Ma J, Gao W, Wong KF (2017a) Detect rumors in microblog posts using propagation structure via kernel learning. In: ACL
154. Ma J, Gao W, Wong KF (2017b) Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics
155. Ma J, Gao W, Wong KF (2018a) Rumor detection on Twitter with tree-structured recursive neural networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp 1980–1989
156. Ma J, Gao W, Wong KF (2018b) Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics
157. MAGAZINE P (2022) The long and brutal history of fake news. https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/. Accessed 23 March 2022
158. MarketWatch (2022) This day in history: Hacked ap tweet about white house explosions triggers panic. https://www.marketwatch.com/story/this-day-in-history-hacked-ap-tweet-about-white-house-explosions-triggers-panic-2018-04-23. Accessed 13 March 2022
159. Markines B, Cattuto C, Menczer F (2009) Social spam detection. In: Proceedings of the 5th international workshop on adversarial information retrieval on the web. Association for Computing Machinery, New York, pp 41–48. https://doi.org/10.1145/1531914.1531924
160. McCallum A, Nigam K et al (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. Citeseer, pp 41–48
161. McCornack SA, Morrison K, Paik JE, Wisner AM, Zhu X (2014) Information manipulation theory 2: a propositional theory of deceptive discourse production. J Lang Soc Psychol 33:348–377
162. Meel P, Vishwakarma DK (2020) Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst Appl 153:112986

163. Meel P, Vishwakarma DK (2020b) Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst Appl 153:112986. https://www.sciencedirect.com/science/article/pii/S0957417419307043, https://doi.org/10.1016/j.eswa.2019.112986
164. Mikolov T, Yih Wt, Zweig G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: human language technologies, pp 746–751
165. Mitra T, Gilbert E (2015) Credbank: a large-scale social media corpus with associated credibility annotations. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10582
166. Mitra T, Wright GP, Gilbert E (2017) A parsimonious language model of social media credibility across disparate events. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, pp 126–145
167. Molina MD, Sundar SS, Le T, Lee D (2021) fake news is not simply false information: a concept explication and taxonomy of online content. Am Behav Sci 65:180–212
168. Mukherjee A, Venkataraman V, Liu B, Glance N (2013) What yelp fake review filter might be doing? In: Seventh international AAAI conference on weblogs and social media
169. Müller M, Salathé M, Kummervold PE (2020) Covid-twitter-bert: a natural language processing model to analyse covid-19 content on twitter. arXiv:2005.07503
170. Mustafaraj E, Metaxas PT (2017) The fake news spreading plague: was it preventable? In: Proceedings of the 2017 ACM on web science conference, pp 235–239
171. Myslinski LJ (2012) Fact checking method and system. Google Patents, US Patent 8,185,448
172. Newman ML, Pennebaker JW, Berry DS, Richards JM (2003) Lying words: predicting deception from linguistic styles. Personal Soc Psychol Bull 29:665–675
173. Newman ML, Pennebaker JW, Berry DS, Richards JM (2003) Lying words: predicting deception from linguistic styles. Personal Soc Psychol Bull 29:665–675. https://doi.org/10.1177/0146167203029005010 pMID: 15272998
174. News B (2022a) China investigates search engine baidu after student's death. https://www.bbc.com/news/business-36189252. Accessed 13 March 2022
175. News B (2022b) 'hundreds dead' because of covid-19 misinformation. https://www.bbc.com/news/world-53755067. Accessed 20 March 2022
176. News U (2022c) During this coronavirus pandemic, 'fake news' is putting lives at risk: Unesco. https://news.un.org/en/story/2020/04/1061592. Accessed 20 March 2022
177. Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2:175–220
178. Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Polit Behav 32:303–330
179. ODonovan J, Kang B, Meyer G, Höllerer T, Adalii S (2012) Credibility in context: an analysis of feature distributions in twitter. In: 2012 International conference on privacy, security, risk and trust and 2012 international conference on social computing. IEEE, pp 293–301
180. Oliveira N, Pisa PS, Lopez MA, Medeiros DSV, Mattos DMF (2021) Identifying fake news on social networks based on natural language processing: trends and challenges. Inf 12:38
181. Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social data: biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2:13. https://www.frontiersin.org/article/10.3389/fdata.2019.00013, https://doi.org/10.3389/fdata.2019.00013
182. Oremus W (2016) Stop calling everything fake news
183. Pal A, Loke C (2019) Communicating fact to combat fake: analysis of fact-checking websites. In: Proceedings of the 2019 international conference on information technology and computer communications, pp 66–73
184. Papadopoulou O, Zampoglou M, Papadopoulos S, Kompatsiaris I (2017) A two-level classification approach for detecting clickbait posts using text-based features. arXiv:1710.08528
185. Pariser E (2011) The filter bubble: what the Internet is hiding from you. Penguin UK
186. Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: International conference on machine learning. PMLR, pp 1310–1318
187. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar MS, Ekbal A, Das A, Chakraborty T (2021) Fighting an infodemic: Covid-19 fake news dataset. In: International workshop on combating on line hostile posts in regional languages during emergency situation. Springer, pp 21–29
188. Pauca VP, Shahnaz F, Berry MW, Plemmons RJ (2004) Text mining using non-negative matrix factorizations. In: Proceedings of the 2004 SIAM international conference on data mining. SIAM, pp 452–456
189. Paul C, Matthews M (2016) The russian "firehose of falsehood" propaganda model. Rand Corp 2:1–10

190. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71:2001
191. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
192. Pennycook G, Rand DG (2017) Who falls for fake news? The roles of analytic thinking, motivated reasoning, political ideology, and bullshit receptivity. SSRN Electron J 88:1–63
193. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. arXiv:1708.07104
194. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018a) Deep contextualized word representations
195. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018b) Deep contextualized word representations. arXiv:1802.05365
196. Pierri F, Ceri S (2019) False news on social media: a data-driven survey. SIGMOD Rec 48:18–27. https://doi.org/10.1145/3377330.3377334
197. Pisarevskaya D (2017) Deception detection in news reports in the russian language: lexics and discourse. In: Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism, pp 74–79
198. Plothow R (2017) Defining fake news... again. Post-Register (1 August 2017) A 5
199. Popat K, Mukherjee S, Yates A, Weikum G (2018) Declare: debunking fake news and false claims using evidence-aware deep learning. arXiv:1809.06416
200. Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2017) A stylometric inquiry into hyperpartisan and fake news. arXiv:1702.05638
201. Poynter (2022) Fighting the infodemic: The #coronavirusfacts alliance. https://www.poynter.org/coronavirusfactsalliance/. Accessed 18 March 2022
202. Praveena HD, Guptha NS, Kazemzadeh A, Parameshachari B, Hemalatha K (2022) Effective cbmir system using hybrid features-based independent condensed nearest neighbor model. J Healthcare Eng 2022
203. Qazvinian V, Rosengren E, Radev DR, Mei Q (2011) Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the 2011 conference on empirical methods in natural language processing. Association for Computational Linguistics, Edinburgh, pp 1589–1599. https://aclanthology.org/D11-1147
204. Qian F, Gong C, Sharma K, Liu Y (2018) Neural user response generator: fake news detection with collective user intelligence. In: IJCAI, pp 3834–3840
205. Qian S, Wang J, Hu J, Fang Q, Xu C (2021) Hierarchical multi-modal contextual attention network for fake news detection. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. Association for Computing Machinery, New York, pp 153–162
206. Qin Y, Wurzer D, Lavrenko V, Tang C (2016) Spotting rumors via novelty detection. arXiv:1611.06322
207. Quattrociocchi W, Scala A, Sunstein CR (2016) Echo chambers on facebook. Available at SSRN 2795110
208. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
209. Research P (2022) News use across social media platforms 2016. https://www.pewresearch.org/journalism/2016/05/26/news-use-across-social-media-platforms-2016/. Accessed 11 March 2022
210. Rich PR, Zaragoza MS (2016) The continued influence of implied and explicitly stated misinformation in news reports. J Exp Psychol Learn Mem Cogn 42:62
211. Riedel B, Augenstein I, Spithourakis GP, Riedel S (2018) A simple but tough-to-beat baseline for the fake news challenge stance detection task. arXiv:1707.03264
212. Rini R (2017) Fake news and partisan epistemology. Kennedy Inst Ethics J 27:E–43
213. Ross L, Ward A et al (1996) Naive realism in everyday life: implications for social conflict and misunderstanding. Values Knowl 103:135
214. Roth Y, Pickles N (2022) Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information. Accessed 26 Feb 2022
215. Rowe M, Butters J (2009) Assessing trust: contextual accountability. In: SPOT@ ESWC
216. Roy A, Basak K, Ekbal A, Bhattacharyya P (2018) A deep ensemble framework for fake news detection and classification. arXiv:1811.04670
217. Rubin VL (2017) Deception detection and rumor debunking for social media
218. Rubin VL, Chen Y, Conroy NK (2015) Deception detection for news: three types of fakes. Proc Ass Inf Sci Tech 52:1–4
219. Rubin VL, Conroy N, Chen Y, Cornwell S (2016) Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of the second workshop on computational approaches to deception detection, pp 7–17

220. Ruchansky N, Seo S, Liu, Y (2017) CSI: a hybrid deep model for fake news detection. Association for Computing Machinery, New York, pp 797–806. https://doi.org/10.1145/3132847.3132877
221. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
222. Sadeghi F, Jalaly Bidgoly A, Amirkhani H (2020) Fake news detection on social media using a natural language inference approach. https://doi.org/10.21203/rs.3.rs-107893/v1
223. Salton G (1986) Mcgill, Michael. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York
224. Sampson J, Morstatter F, Wu L, Liu H (2016) Leveraging the implicit structure within social media for emergent rumor detection. In: CIKM 2016 - proceedings of the 2016 ACM conference on information and knowledge management. Association for Computing Machinery, pp 2377–2382. https://doi.org/10.1145/2983323.2983697
225. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108
226. Sardelich M, Manandhar S (2018) Multimodal deep learning for short-term stock volatility prediction. arXiv:1812.10479, https://doi.org/10.48550/ARXIV.1812.10479
227. Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval. Cambridge University Press Cambridge, vol 39
228. Shao C, Ciampaglia GL, Varol O, Flammini A, Menczer F (2017) The spread of fake news by social bots, vol 96, p 104. arXiv:1707.07592
229. Shao C, Ciampaglia GL, Varol O, Yang KC, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nat Commun 9:1–9
230. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y (2019a) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol 10
231. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y (2019) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Sys Tech (TIST) 10:1–42
232. Shifath SMSUR, Khan MF, Islam MS (2021) A transformer based approach for fighting covid-19 fake news. arXiv:2101.12027, https://doi.org/10.48550/ARXIV.2101.12027
233. Shu K, Cui L, Wang S, Lee D, Liu H (2019a) Defend: explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery amp; data mining. Association for Computing Machinery, New York, pp 395–405. https://doi.org/10.1145/3292500.3330935
234. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2018) Fakenewsnet: a data repository with news content, social context and dynamic information for studying fake news on social media
235. Shu K, Sliva A, Wang S, Tang J, Liu H (2017a) Fake news detection on social media: a data mining perspective. arXiv:1708.01967
236. Shu K, Wang S, Liu H (2017b) Exploiting tri-relationship for fake news detection. 8. arXiv:1712.07709
237. Shu K, Wang S, Liu, H (2019b) Beyond news contents: the role of social context for fake news detection. In: Proceedings of the twelfth ACM international conference on web search and data mining, pp 312–320
238. Shu K, Zhou X, Wang S, Zafarani R, Liu H (2019c) The role of user profiles for fake news detection. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. Association for Computing Machinery, New York, pp 436–439. https://doi.org/10.1145/3341161.3342927
239. Shushkevich E, Cardiff J (2021) Tudublin team at constraint@ aaai2021–covid19 fake news detection. arXiv:2101.05701
240. Silverman C (2016a) This analysis shows how viral fake election news stories outperformed real news on facebook. BuzzFeed News 16
241. Silverman C (2016b) Viral fake election news outperformed real news on facebook in final months of the us election. BuzzFeed News 16
242. Singh V, Dasgupta R, Sonagra D, Raman K, Ghosh I (2017) Automated fake news detection using linguistic analysis and machine learning. In: International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation (SBP-BRiMS), pp 1–3
243. Singh VK, Ghosh I, Sonagara D (2021) Detecting fake news stories via multimodal analysis. J Ass Inf Sci Tech 72:3–17
244. Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh S (2019) Spotfake: a multi-modal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, pp 39–47
245. Song C, Ning N, Zhang Y, Wu B (2021) A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf Process Manag 58:102437

246. Southwell BG, Thorson EA, Sheble L (2017) The persistence and peril of misinformation. Am Sci 105:372–375
247. Stone R (2022) Anatomy of a fake news scandal. https://www.rollingstone.com/feature/anatomy-of-a-fake-news-scandal-125877/. Accessed 12 March 2022
248. Sundar SS, Nass C (2001) Conceptualizing sources in online news. J Commun 51:52–72
249. Sunstein CR (2001) Echo chambers: Bush v. Gore, impeachment, and beyond. Princeton University Press, Princeton
250. Sunstein CR (2014) On rumors. In: On rumors. Princeton University Press
251. Tacchini E, Ballarin G, Della Vedova ML, Moret S, de Alfaro L (2017) Some like it hoax: automated fake news detection in social networks. arXiv:1704.07506
252. Tajfel H, Turner JC (2004) The social identity theory of intergroup behavior. in jt jost & j. sidanius (eds) key readings in social psychology. Political Psychology: Key Readings 276–293
253. Tajfel H, Turner JC, Austin WG, Worchel S (1979) An integrative theory of intergroup conflict. Organizational identity: A reader 56:9780203505984–16
254. Tang D, Qin B, Liu T (2015) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1422–1432
255. Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. J Lang Soc Psychol 29:24–54
256. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2018) Fever: a large-scale dataset for fact extraction and verification. arXiv:1803.05355
257. Trends G (2022) fake news - explore - google trends. https://trends.google.com/trends/explore?date=2010-01-01%202022-07-14&q=%22fake%20news%22. Accessed 20 Jul 2022
258. Trstenjak B, Mikac S, Donko D (2014) Knn with tf-idf based framework for text categorization. Procedia Eng 69:1356–1364
259. Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain 5:297–323
260. Undeutsch U (1967) Beurteilung der glaubhaftigkeit von aussagen. Handbuch Der Psychologie 11:26–181
261. Veyseh APB, Thai MT, Nguyen TH, Dou D (2019) Rumor detection in social networks via deep contextual modeling. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, pp 113–120
262. Vlachos A, Riedel S, (2014a) Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 workshop on language technologies and computational social science, pp 18–22
263. Vlachos A, Riedel S, (2014b) Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 workshop on language technologies and computational social science. Association for Computational Linguistics, Baltimore, pp 18–22. https://aclanthology.org/W14-2508 https://doi.org/10.3115/v1/W14-2508
264. Vo N, Lee K (2018) The rise of guardians: fact-checking url recommendation to combat fake news. In: The 41st international ACM SIGIR conference on research & development in information retrieval, pp 275–284
265. Volkova S, Shaffer K, Jang JY, Hodas N (2017a) Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the 55th annual meeting of the association for computational linguistics, (vol 2: Short papers), pp 647–653
266. Volkova S, Shaffer K, Jang JY, Hodas N (2017b) Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on Twitter. In: Proceedings of the 55th annual meeting of the association for computational linguistics, (vol 2: Short Papers). Association for Computational Linguistics, Vancouver, pp 647–653. https://aclanthology.org/P17-2102, https://doi.org/10.18653/v1/P17-2102
267. Vosoughi S (2015) Automatic detection and verification of rumors on Twitter. Ph.D. thesis. Massachusetts Institute of Technology
268. Vosoughi S, Roy D, Aral S (2018a) The spread of true and false news online. Sci 359:1146–1151. https://www.science.org/doi/abs/10.1126/science.aap9559, https://doi.org/10.1126/science.aap9559
269. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Sci 359:1146–1151
270. Vrij A (2000) Detecting lies and deceit: the psychology of lying and implications for professional practice. Wiley
271. Wallach HM (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on machine learning, pp 977–984
272. Wang WY (2017) "liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv:1705.00648

273. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 849–857
274. Wani A, Joshi I, Khandve S, Wagh V, Joshi R (2021) Evaluating deep learning approaches for covid19 fake news detection. In: Combating online hostile posts in regional languages during emergency situation. Springer International Publishing, pp 153–163. https://doi.org/10.1007/978-3-030-73696-5_15
275. Wardle C (2017) Fake news. it's complicated. First Draft 16:1–11
276. Weedon J, Nuland W Stamos A (2017) Information operations and facebook. Retrieved from Facebook: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf
277. Wu K, Yang S, Zhu KQ (2015) False rumors detection on sina weibo by propagation structures. 2015 IEEE 31st international conference on data engineering, pp 651–662
278. Wu L, Liu H (2018) Tracing fake-news footprints: characterizing social media messages by how they propagate. In: Proceedings of the eleventh ACM international conference on web search and data mining. Association for Computing Machinery, New York, pp 637–645
279. Wu L, Morstatter F, Hu X, Liu H (2016) Mining misinformation in social media. In: Big data in complex and social networks. Chapman and Hall/CRC, pp 135–162
280. Wu Y, Zhan P, Zhang Y, Wang L, Xu Z (2021) Multimodal fusion with co-attention networks for fake news detection, In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, pp 2560–2569
281. Wynne HE, Wint ZZ (2019) Content based fake news detection using n-gram models. In: Proceedings of the 21st international conference on information integration and web-based applications & services, pp 669–673
282. Yang F, Liu Y, Yu X, Yang M (2012) Automatic detection of rumor on sina weibo. In: Proceedings of the ACM SIGKDD workshop on mining data semantics. Association for Computing Machinery, New York. https://doi.org/10.1145/2350190.2350203
283. Yang Y, Zheng L, Zhang J, Cui Q, Li Z, TI-CNN PSY (2018) Convolutional neural networks for fake news detection 2. arXiv:1806.00749
284. Ying L, Yu H, Wang J, Ji Y, Qian S (2021) Multi-level multi-modal cross-attention network for fake news detection. IEEE Access 9:132363–132373
285. Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag 13:55–75
286. Zafarani R, Abbasi MA, Liu H (2014) Social media mining: an introduction. Cambridge University Press. https://doi.org/10.1017/CBO9781139088510
287. Zajonc RB (1968) Attitudinal effects of mere exposure. J Pers Soc Psychol 9:1
288. Zajonc RB (2001) Mere exposure: a gateway to the subliminal. Curr Dir Psychol Sci 10:224–228
289. Zannettou S, Sirivianos M, Blackburn J, Kourtellis N (2019) The web of false information. J Data Inf Qual 11:1–37. https://doi.org/10.1145/3309699
290. Zhang H, Alim MA, Li X, Thai MT, Nguyen HT (2016) Misinformation in online social networks: detect them all with a limited budget. ACM Trans Inf Syst 34. https://doi.org/10.1145/2885494
291. Zhao Z, Resnick P, Mei Q (2015) Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th international conference on world wide web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp 1395–1405. https://doi.org/10.1145/2736277.2741637
292. Zhou L, Burgoon JK, Nunamaker JF, Twitchell D (2004) Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group Decis Negot 13:81–106
293. Zhou X, Jain A, Phoha, VV, Zafarani R (2020a) Fake news early detection: an interdisciplinary study. arXiv:1904.11679
294. Zhou X, Wu J, Zafarani R (2020b) Safe: similarity-aware multi-modal fake news detection, In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 354–367
295. Zhou X, Zafarani R (2019) Network-based fake news detection: a pattern-driven approach. arXiv:1906.04210
296. Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 53. https://doi.org/10.1145/3395046
297. Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv (CSUR) 53:1–40
298. Zhou X, Zafarani R, Shu K, Liu H (2019) Fake news: fundamental theories, detection strategies and challenges. In: WSDM 2019 - Proceedings of the 12th ACM international conference on web search and data mining. Association for Computing Machinery, Inc., pp 836–837. 12th ACM International

conference on Web Search and Data Mining, WSDM 2019; Conference date: 11-02-2019 Through 15-02-2019. https://doi.org/10.1145/3289600.3291382
299. Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: a survey. ACM Comput Surv 51. https://doi.org/10.1145/3161603
300. Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: a survey. ACM Comput Surv (CSUR) 51:1–36
301. Zubiaga A, Kochkina E, Liakata M, Procter R, Lukasik M (2016a) Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. arXiv:1609.09028
302. Zubiaga A, Liakata M, Procter R (2016b) Learning reporting dynamics during breaking news for rumour detection in social media. arXiv:1610.07363
303. Zubiaga A, Liakata M, Procter R (2017) Exploiting context for rumour detection in social media. In: International conference on social informatics. Springer, pp 109–123
304. Zuckerman M, DePaulo BM, Rosenthal R (1981) Verbal and nonverbal communication of deception. In: Advances in experimental social psychology. Elsevier, vol 14, pp 1–59