Literature Review
Predicting Users Likely to Repost Unreliable News on X (formerly Twitter)

1. INTRODUCTION

With millions of users depending on feeds for information updates, social media platforms have emerged as a significant news consumption channel [1–3]. Regrettably, incorrect or misleading material can also proliferate on these platforms [2, 4, 5]. Online, fake news, pseudoscience, and propaganda spread quickly and extensively, frequently outpacing reliable information [2, 4, 6, 7]. The following are key terms in this domain: **rumours** (unverified claims that may later prove true or false [8, 9]), **fake news** (fabricated news articles styled to appear legitimate), **misinformation** (inaccurate or misleading information shared regardless of intent) [2, 3], and **disinformation** (deliberately false information spread with the intent to deceive [3, 6]). For clarity, information from reputable, fact-checked sources is referred to as **reliable news**, whereas information from sources with a reputation for partisan misrepresentation or poor factual accuracy is referred to as **unreliable news** [1, 5].

Unchecked misinformation has significant negative effects on society. Online falsehoods have increased conflict in the real world and damaged public confidence in institutions [2, 7]. For instance, false information about health during pandemics has jeopardised public safety, and conspiracy theories regarding electoral fraud have incited violence [11,12]. Individual user behaviour and psychology can also be impacted by exposure to false information. According to studies, people may become desensitized and alter their posting behaviour (e.g., using fewer analytical or emotional words and more profanity) after coming across fake news [8]. A vicious cycle of amplification can be created when deceptive content elicits negative emotions and moral outrage, which can boost its virality [12, 13]. On sites like Twitter, it has been demonstrated that false information spreads “farther, faster, deeper, and more broadly” than real information (as initially measured by Vosoughi et al. [4]). It is noteworthy that a tiny minority of very active users frequently drive the propagation of misinformation: previous investigations of fake news connected to elections revealed that approximately 80% of misinformation disseminated on Twitter came from just 0.1% of users [5]. These low-credibility material “superspreaders” typically have sizable fan bases and exert a disproportionate amount of network influence [5, 16]. The urgent need for efficient techniques to identify users likely to spread false information is highlighted by the combination of broad reach and nuanced behavioural consequences. In the sections that follow, we examine pertinent fields of study and cutting-edge methods that guide our inquiry into predicting these users.

2. REVIEW OF RESEARCH FOCUS AREAS

2.1 Misinformation Detection and Characterization

The capacity to identify and define fake news itself is absolutely fundamental for research on spreaders of false information. Aiming to automatically separate fraudulent or low-credibility material from authentic news, fake news detection on social media has been the subject of much study. Recent thorough studies track the development and future trends of this domain [2, 7]. These works underline the development from early methods based on basic content heuristics to more complex machine learning models including text cues, user cues, images, and network interactions. Modern fake information detection techniques use deep learning and natural language processing (NLP) to seize linguistic patterns suggesting fraud or low credibility [17]. Classifiers, for instance, have been taught on social media postings and news articles to spot false material using characteristics such writing style, semantic discrepancies, and even multi-modal content (pictures, videos) when accessible [17, 18]. Especially when several feature kinds (text, social context, etc.) are combined in a hybrid model, results demonstrate persistent high accuracy given sufficient training data for known misinformation [18].

The description of misinformation and rumours ties to material detection. Though both can be detrimental, scholars differentiate “rumours” (unverified information that might eventually be proven or discredited) from clear erroneous “misinformation” [8]. Research has shown that although rumours and accurate information follow roughly the same diffusion patterns, rumours and fake news are frequently designed to be more viral; they also tend to be more unexpected or contentious, which could speed its spread [4]. False articles sometimes attract attention with clickbait titles, emotionally charged language, or startling claims. For example, a study by Vosoughi et al. showed that bogus news articles on Twitter spread far more quickly, deeply, and broadly than accurate stories across several subject areas [4]. Such results highlight the need for quick detection: by the time a fake narrative is disproved, it could have already spread widely.

2.2 Diffusion Patterns and Social Dynamics

Another focal point investigates the dissemination patterns of misinformation inside social networks. Researchers in this domain examine the dissemination of incorrect information through user interactions (retweets, responses, mentions) and the network structures that either promote or obstruct its propagation. Misinformation propagation is consistently observed to be focused inside specific communities and echo chambers rather than being uniformly distributed across the network. Case studies on specific subjects, such as COVID-19 misinformation or political conspiracies, indicate that users who often interact with low-credibility content tend to create densely interconnected clusters, facilitating the rapid circulation of disinformation. Vilella et al. assigned individuals a “Untrustworthiness” score based on their interactions with unreliable news sources during an Italian Twitter debate, revealing that users with high scores were significantly clustered together [19]. These

clusters aligned with specific political inclinations and had a higher prevalence of bot-like accounts, demonstrating a synergy between human disinformation networks and automated agents.

Emotional and ethical content has become a vital element in dissemination. Misinformation often exploits outrage, fear, or other intense emotions to encourage dissemination. Recent studies offer actual evidence that messages invoking moral emotions, such as rage or indignation, exhibit increased virality, particularly with health and political disinformation [12]. Solovev and Pröllochs examined COVID-19 misinformation on social media and discovered that content with moral-emotional language disseminated more extensively than content with a neutral tone [12]. Likewise, messages characterised by negative feelings or an unsettling tone have been associated with heightened retweet rates [13]. A 2023 study by Robertson et al. concluded that negativity significantly influences engagement and sharing behaviour in online news consumption [14]. These findings correspond with theoretical expectations that emotionally charged misinformation (e.g., hoaxes that elicit wrath or worry) attracts attention and motivates users to disseminate the information, hence expediting dissemination. Conversely, correcting information and factual rebuttals generally disseminate more slowly and reach limited audiences, complicating efforts to address viral misinformation [15].

An additional significant aspect is the distinction between influential users and ordinary users in the dissemination of disinformation. Studies demonstrate that a limited quantity of extensively connected or well followed accounts can function as “hubs” that disseminate disinformation throughout the network, propagating it through numerous reshares [5, 16]. These prominent disseminators encompass political commentators, marginal media platforms, or charismatic individuals who exert significant influence on the dissemination of information [16]. Average users, although lacking substantial followings, can nonetheless participate in cascades; however, their influence is significantly enhanced when integrated inside receptive groups and when promoting material from high-influence accounts. Misinformation diffusion fundamentally involves a complicated interaction between content attributes (e.g., emotional appeal) and network dynamics (who distributes information with whom). This interaction establishes the circumstances for sporadic explosive surges of false information, while numerous other instances of disinformation remain confined inside small groups. Comprehending these patterns is essential for predicting which users are prone to launch or propagate such cascades.

2.3 User Characteristics and Susceptibility

To predict individuals likely to disseminate inaccurate news, one must evaluate the attributes of the users themselves. A burgeoning field of study, frequently situated at the confluence of computational social science and psychology, investigates the characteristics, attributes, and actions of those disseminating disinformation. A significant discovery from survey-based and observational studies is the impact of demographic variables. Guess et al. examined the dissemination of bogus news on Facebook and discovered that elderly persons (65 years and above) shared much more fabricated stories

than younger demographics, even after adjusting for additional characteristics [1]. The “age effect” is posited to be associated with variations in digital media literacy or cognitive susceptibility among different generations. Political ideology is an additional factor: the study by Guess et al. and others noted that conservatives (within the U.S. context) were marginally more inclined to disseminate recognised fake news sources than liberals, although the age variable was more pronounced [1]. These demographic relationships indicate that specific groups of the population may be more susceptible to believing or disseminating misinformation.

In addition to demography, personality and cognitive characteristics have also been examined. In a collaborative challenge focused on identifying fake news disseminators, teams utilised psycho-linguistic attributes to extract personality indicators from users’ tweets [20]. The rationale is that characteristics like impulsivity, openness, or conscientiousness may affect an individual’s propensity to critically assess information prior to dissemination. Shrestha et al. indicated that the integration of Big Five personality factors, derived from linguistic analysis, enhanced the identification of fake news disseminators, suggesting that personality influences user sharing behaviour [20]. A separate study analysed user language following exposure to misinformation and identified subsequent behavioural alterations: users exhibited an increase in tweets containing anger or conflict-oriented vocabulary and a decrease in positive emotional language, indicating that exposure to misinformation may elicit negative emotional reactions and foster a defensive mentality [8]. These observations suggest a feedback cycle in which vulnerability to misinformation may originate from and strengthen specific emotional or personality-driven communication behaviours.

Significantly, not all individuals disseminating misinformation do so with malevolent intent. Some individuals may inadvertently disseminate misinformation. Recent study has sought to identify the intentions of disseminators. Zhou et al. examined Twitter users who disseminated false information, seeking evidence of whether such actions were inadvertent (e.g., subsequent post deletion or expressions of remorse) or deliberate [22]. A substantial segment of users seemed to disseminate false material inadvertently, underscoring the distinction between intentional misinformation purveyors and casual users who may be naive or negligent in verifying facts. Individuals that deliberately disseminate false information exhibit consistent behaviour by regularly spreading questionable content, while unintentional disseminators are more erratic and occasionally rectify their mistakes [22]. This differentiation is essential for predicting: characteristics or patterns that indicate a habitual disseminator of misinformation may vary from those that recognise a singular sharer. A user’s information environment, specifically the sources they follow and encounter, is another factor associated with the dissemination of disinformation. Mosleh and Rand assessed users’ exposure to disinformation on Twitter by monitoring the frequency of tweets from recognised political elites that promote misinformation [23]. Research indicated that individuals who followed a greater number of accounts that regularly disseminate disinformation were more susceptible to encountering and even sharing such content, hence reinforcing the notion that “you are what you read.” This focus area underscores the necessity of considering user demographics, personality, social feed exposure, and historical sharing

patterns of low-credibility content to predict spreading behaviour. These variables jointly delineate user vulnerability to disseminating inaccurate news.

2.4 Predictive Modeling of Misinformation Spreaders

Building on the aforementioned fields, an increasing amount of research directly addresses the challenge of predicting which users are most likely to provide false information. From the standpoint of predictive modelling, the issue is one of user classification or risk prediction. Mu and Aletras (2020), who provided a dataset of Twitter users marked by whether they had ever reposted material from dubious news sources [24], were among the first research to clearly define this objective. Approximately 6,200 users were collected, divided between those who shared material from known fake/low-credibility news sources and those who only shared from reputable sources. Models were then trained to predict the category based only on the users' own tweet history, excluding the unreliable news posts themselves [24]. Using linguistic characteristics of their past posts, this investigation showed the viability of the task, attaining around 80% F1 accuracy in separating probable fake news spreaders from others. It verified that individuals had noticeable linguistic signatures, in terms of subjects, mood, and style, that correspond with their tendency to disseminate false information.

Separately, the PAN@CLEF 2020 contest had a task on characterising fake news spreaders, which helped to drive creation of predictive methods [20]. Several research groups created algorithms to automatically determine if a certain user, represented by a sample of their tweets, was a "fake news spreader" or not. Reported evaluation indicators for several methodologies and a benchmark dataset in English and Spanish this common effort gave Usually, the best-performing techniques integrated a rich collection of characteristics: for instance, Shrestha et al.'s method employed lexical elements (e.g. character and word *n*-grams), stylistic aspects, sentiment, and personality traits deduced from language, fed into ensemble classifiers [20]. Such initiatives indicated that predicting spreaders is tractable using supervised learning since they could categorise users with more than 70% accuracy. They also mentioned difficulties, though, such as differentiating bogus news spreaders from generic spam/bot accounts and making sure models weren't only picking up topic-specific signals that could not generalise.

Another branch of predictive research examines information diffusion models to anticipate who would disseminate false information and how. Rather than considering users in isolation, these models simulate or predict the cascade of reshares an original post could produce. Some researchers treat this as a cascade prediction or link prediction challenge: given an initial set of sharers of a piece of material, predict which other people will rebroadcast it in the future [9]. Tian et al. (2020) suggested a model in the context of rumours to predict which users will repeat a rumour during public crises using characteristics such user influence, proximity to the source in the network, and topic relevance [9].

Often using graph-based methods or perhaps deep learning on social networks, for example, graph neural networks, such models evaluate the likelihood of every candidate user joining the cascade.

Although diffusion prediction has historically been content-agnostic, current research is beginning to include material credibility into these models, combining cascade prediction with false information detection. For instance, one could use a diffusion model to predict a tweet's distribution and therefore find high-risk individuals in its route if it was flagged as probably fake. Predictive modelling emphasis, thus, closes the gap between knowledge of historical patterns and projection of future behaviour. It proactively flags users who are likely to propagate the next piece of false news that comes along using knowledge from detection, diffusion, and user profiling. Since 2020, new datasets and benchmarks have made this field rapidly develop; it is still an active research frontier with clear relevance for content control and social platform policy (e.g., directing which users or communities could merit fact-checking attention before significant events).

3. DETECTION AND IDENTIFICATION TECHNIQUES

Researchers have employed a variety of techniques to detect likely misinformation spreaders and identify them before or as they engage in reposting unreliable news. These techniques span feature engineering approaches, network analysis, and modern machine learning models.

3.1 Feature-Based User Classification

A prevalent method is to define the task as a binary classification problem: based on a user's data, categorise them as a possible "unreliable news disseminator" or not. The efficacy of this method is largely contingent upon the selection of traits that identify indicators of behaviour susceptible to disinformation. Previous research has developed features from several domains:

3.1.1 User Profile and Behavior Features. These elements encapsulate user characteristics and online behaviours, including profile metadata (e.g., account age, follower count, verification status) and behavioural patterns (posting frequency, interaction networks, etc.). Previous research indicates that user-centric characteristics can serve as significant predictors of an individual's likelihood to disseminate disinformation. Lin et al. [18] emphasise that the integration of social context elements, user statistics, characteristics of publishers, and their distribution networks yields essential indicators for the early detection of fake news, hence markedly enhancing model efficacy. The hybrid system demonstrated that user profile information, in conjunction with text and photos, enhanced accuracy, highlighting the significance of profile and behavior-based indicators in detecting malicious disseminators [18].

Other research indicates that demographic or political characteristics are associated with increased dissemination of disinformation. Older users and those with strong partisan beliefs, especially

conservative ones, have been observed to disseminate much more misinformation [1, 5]. Profile attributes, including account lifetime, follower-to-followee ratios, and verification status, can function as indicators of authenticity; ephemeral accounts with disproportionately high follower counts may suggest inauthentic or bot-assisted expansion. Unverified accounts or those employing misleading or generic profile descriptions are more prone to disseminating incorrect information. Furthermore, behavioural indicators, particularly posting frequency, retweet ratios, and temporal activity clusters, can identify habitual disseminators of misinformation. Research assessing "untrustworthiness" or the frequency of sharing low-credibility sources indicates that individuals with high scores typically engage with bot-like accounts, implying a relationship between automated agents and susceptible human disseminators [19].

Similarly, Mu and Aletras [24] note that individuals who subsequently disseminate inaccurate news display identifiable behaviour in their previous posts, such as an excessive emphasis on polarising political subjects and partisan rhetoric. These findings suggest that specific profile-driven characteristics (e.g., a user's subject interests or engagement style) can predict susceptibility to disinformation [24]. By utilising user metadata and behavioural patterns, classifiers can profile individuals and identify those exhibiting traits linked to disseminating false or incorrect information, independent of the precise content of a message.

3.1.2 Content and Linguistic Features. Content-based features examine the textual material generated or shared by users, emphasising both the substance and the manner of expression. This encompasses lexical attributes (e.g., keywords or *n*-grams indicative of disinformation subjects), syntactic and stylistic indicators (writing complexity, utilisation of emotive language or particular pronouns), and semantic or topical characteristics (sentiment, psycho-linguistic categories, etc.). Linguistic traits are fundamental to the classification of misinformation users, as the language employed by individuals disseminating misleading information frequently diverges from that of other users.

Mu and Aletras [24] present compelling evidence for the effectiveness of linguistic features: utilising a comprehensive array of textual indicators (bag-of-words, topic modelling, and LIWC metrics) derived from users' tweet histories, their model achieved nearly 80% macro-F1 accuracy in predicting which users would share content from unreliable news sources. Distinct stylistic and thematic differences were identified between "unreliable" and "reliable" users; for instance, the former's tweets predominantly focus on polarised political discourse, while the latter's content is more aligned with everyday-life subjects, thereby substantiating the correlation between language patterns and misinformation engagement [24].

Additional studies yield comparable findings: lexicon-based and stylometric attributes (e.g., word and character *n*-grams, punctuation utilisation, vocabulary diversity) are proficient in delineating writing

style [10, 20, 32]. Psycholinguistic indicators, assessed using LIWC categories or sentiment polarity, might identify emotionally charged or negatively biased language prevalent among disseminators of misinformation [33]. Research indicates that the integration of psycholinguistic signals and style markers (e.g., all-caps, exaggerated exclamation) frequently enhances effectiveness in user-level false news identification [20, 32]. As noted by Lin et al. [18], text-based analysis has traditionally been the predominant modality for disinformation detection, underscoring the critical significance of content features in this field.

3.1.3 Image-Centric Features. In addition to text, image-centric features utilise the visual content shared by users (such as photographs, memes, and connected media), which may contain substantial misleading indicators. Visual attributes may encompass picture metadata, descriptions of objects or scenes, and potential forensic indicators of alteration. Despite being traditionally underexplored compared to textual aspects, recent research highlights their increasing significance. Lin et al. [18] observe that a limited number of user classification methods integrate visuals, even though images frequently accompany bogus news. Their research indicates that using picture features alongside textual and social (user) context variables significantly improves detection, attaining approximately 92.5% accuracy, surpassing models that utilise solely text or user data [18].

Comparable evidence is derived from case studies on crisis disinformation, including the dissemination of fraudulent photographs during Hurricane Sandy [26], and contemporary methodologies employing deep learning to identify altered or low-credibility media [27]. A user who frequently disseminates altered or inaccurately labelled photographs may be identified as a habitual disseminator. Cross-modal inconsistencies, such as discrepancies between text captions and image content, suggest dishonest intent and enhance detection [28]. The integration of image-centric signals with textual data and user traits has demonstrated advantages in multimodal frameworks, resulting in enhanced detection of users disseminating misinformation. The efficacy of multimodal techniques demonstrates that visuals provide further signals; a photo taken out of context can expose disinformation, even when the accompanying text appears innocuous. As a result, image-based traits are widely acknowledged as a primary category for user-level categorisation, enhancing the capacity to identify and characterise disinformation disseminators [18].

Standard classification techniques, including logistic regression, support vector machines, random forests, and gradient boosting machines, have been utilised with these features and demonstrated satisfactory performance in differentiating probable spreaders from non-spreaders [20, 24]. Recently, deep learning methodologies utilise the concatenation of user features as input for multilayer perceptrons or employ sequence models on a user's postings (e.g., an LSTM analysing a sequence of tweets) to generate predictions [24]. Deep models have the capacity to capture non-linear feature interactions and nuanced patterns; yet, they necessitate larger datasets and are susceptible to overfitting. In the trials conducted by Mu and Aletras, simpler classifiers such as logistic regression,

when supplemented with meticulously designed language features, demonstrated competitive performance against neural models for their dataset, underscoring the importance of a comprehensive feature collection [24]. Ensemble approaches that integrate several classifiers have triumphed in competitions, exemplified by an ensemble of SVM and neural networks in the PAN 2020 challenge, by utilising various facets of the feature space [20]. Feature-based categorisation continues to be a foundational technique, experiencing consistent enhancements as novel feature types (e.g., images in posts, interaction patterns) and representation learning methodologies (e.g., user embedding through graph representation) are incorporated.

3.2 Network-Based and Graph Techniques

In addition to addressing each user individually, another category of methods directly leverages the social network architecture and information dissemination patterns to identify probable spreaders. These methodologies are based on network analysis and often examine a propagation scenario (for a specific piece of disinformation) to deduce which users would be implicated. One method involves analysing propagation graphs for established misinformation cascades. Through the analysis of historical misinformation dissemination events, researchers ascertain which users regularly emerge early or centrally within these cascades. For example, if specific individuals were among the initial retweeters of numerous misleading narratives, those users may be classified as chronic disseminators [5]. This retroactive identification can inform predictive rules: presuming that future disinformation will be similarly propagated by the same cohort of superspreaders. DeVerna et al. implemented this by presenting straightforward metrics to predict leading superspreaders months ahead [16]. They discovered that ranking individuals based on the volume of low-credibility content they share within a specific time frame enables a relatively accurate prediction of who will lead the rankings in a subsequent period [16]. This persistence indicates that high-risk spreaders sustain their behaviour over time. This strategy, albeit simple, necessitates the oversight and revision of a roster of common disseminators.

Graph-based algorithms are capable of identifying structural signatures of spreaders. Community discovery methods, such as modularity-based clustering, frequently uncover communities associated with interest or belief groups. If a certain community is recognised for its significant involvement in misinformation (e.g., an anti-vaccine organisation), then its members, regardless of prior observation in disseminating erroneous information, may be at an elevated risk. Consequently, community membership emerges as a characteristic. Furthermore, algorithms designed for influence maximisation or information centrality in networks, commonly employed in marketing to identify influencers, can be adapted: individuals who occupy prominent positions in the disinformation propagation network are probable subjects for monitoring. Techniques such as *k*-core decomposition can discern the “inner core” of enduring disseminators in misinformation cascades. Research indicates that disinformation networks typically have a core-periphery structure, wherein a central group of interconnected active disseminators propels virality [5]. Utilising graph algorithms to identify users within that core

effectively designates them as probable spreaders.

A different array of methodologies encompasses cascade prediction models. These algorithms, sometimes employing deep learning, utilise a partial cascade of reshares as input to estimate the future reshare count or identify the subsequent nodes to be infected. Although frequently employed for content-agnostic viral marketing or general retweet predicting, many models have been specifically adapted for rumours. For instance, certain graph neural network algorithms encode the diffusion tree of a post and predict which followers of the current sharers will retweet subsequently, occasionally incorporating user attributes and tweet content into the graph node representations [31]. Within a disinformation framework, one may model the propagation of a fictitious narrative originating from a certain user and identify which people the model predicts would participate in the cascade; those consistently identified could be deemed prospective disseminators. This process is computationally demanding and generally necessitates an understanding of the social network and prior cascades.

A more straightforward graph-based metric encompasses bot identification and interaction with people. Numerous programs, such as Botometer, are available to evaluate accounts based on their probability of being bots. Upon identifying an account as a bot disseminating misinformation, one can examine its human retweeters or mentioners to ascertain those amplifying the bot's remarks. Those individuals may be acting either unconsciously or in a coordinated manner. Network patterns, such as a human account retweeting numerous tweets from recognised bots that disseminate misinformation, may indicate that the human is functioning as a conduit, propagating misleading content further into the human network. Although bot detection is an independent domain, using it as a preprocessing measure (to filter bots or categorise bot-human interactions) improves the identification of authentic at-risk consumers [4]. It is important to acknowledge that network-based methodologies are potent but may falter when the social graph is partially observed or when addressing new individuals or events where historical network behaviour is unavailable. They may also elicit heightened privacy and ethical concerns, as they entail the mapping of social connections. A synthesis of user-level attributes and network context typically produces optimal outcomes. For instance, certain methods for detecting false information utilise a graph comprising users and material, where an edge signifies that a user has shared a specific piece of content, then employing graph convolution or propagation to concurrently ascertain which users and content are likely to be disinformation [29]. Within those frameworks, user nodes with elevated misinformation scores are successfully recognised as disseminators as a consequence of content detection. The joint modelling of content and users through graphs is a promising approach, as it reflects the reality that specific users and particular pieces of content mutually reinforce each other's classification (a user disseminating numerous fake stories likely indicates both the user's untrustworthiness and the untrustworthiness of those stories).

3.3 Advanced Machine Learning and Hybrid Approaches

Recent research has commenced investigating sophisticated machine learning methodologies and hybrid models that integrate several dimensions (content, user behaviour, network) to predict the dissemination of disinformation. A significant development is the utilisation of deep neural networks specifically designed for user prediction tasks. For instance, instead of manually engineering features, several studies employ representation learning to autonomously extract features. A neural network can be trained end-to-end by inputting raw data, such as a user's tweet sequence, allowing the network to develop an internal representation that effectively classifies the user as a spreader or not. Methods employing convolutional neural networks (CNNs) on a user's aggregated tweets, or recurrent neural networks (RNNs) that analyse a user's tweet history as a temporal sequence, have demonstrated competitive efficacy [24]. These algorithms may identify intricate patterns, such as temporal rhythms or contextual nuances in language, that manual features may overlook. Nevertheless, they necessitate more extensive training datasets to prevent overfitting and frequently gain advantages from pre-training (e.g., initiating with a language model).

Another approach is multi-task learning, in which the model is concurrently taught on interconnected tasks. A model might be trained to simultaneously recognise fake news content and identify disseminators of fake news inside a unified framework, sharing some layers between the two objectives [17]. The premise is that by acquiring the ability to identify disinformation, the model recognises content trends that may also indicate which individuals are predisposed to disseminate it. Qing Liao et al. developed a cohesive model that addressed both false news identification and user stance categorisation, facilitating the exchange of representations between content analysis and user analysis [17]. These multi-task models have demonstrated enhanced performance on both tasks compared to training individual models, owing to the complementary nature of the signals. Ensemble and hybrid methodologies continue to be prevalent. An initial strategy could involve employing a content-based fake news detector to ascertain which information is likely fraudulent, followed by utilising a diffusion model to predict the individuals who would disseminate such content. The two stages collectively predict the disseminators of misinformation regarding a particular narrative. Alternatively, an ensemble may comprise a network-based predictor and a user-feature-based predictor, with their outputs amalgamated (e.g., through weighted voting or a meta-classifier). This is advantageous as certain disseminators may be constrained by their network location despite their seemingly benign tweet content, whereas others may be prominent owing to their content patterns even if they are not centrally situated within the network.

Regarding practical identification on platforms, one method employed by Twitter and others is the implementation of intervention tags (such as "potentially misleading" labels) on material and monitoring subsequent engagement declines. A person who persists in disseminating content despite it being labelled or predominantly shares material that includes warnings is likely a habitual purveyor of misinformation. Papakyriakopoulos and Goodman examined the effects of Twitter's misinformation labels implemented during the 2020 U.S. election and concluded that certain habitual disseminators remained unaffected by these labels, persisting in the propagation of falsehoods despite the platform's

interventions [11]. This analysis can facilitate identification: individuals who regularly encounter and disregard such warnings may be flagged by automated systems for further scrutiny.

Progress is being made in the application of large language models (LLMs) and sophisticated AI for the analysis of misinformation. Although LLMs (such as GPT-based models) have predominantly been employed to identify counterfeit content or generate rationales for its inauthenticity [30], they may also serve to profile consumers. An LLM may be prompted with a user's recent posts and tasked with evaluating the probability that the user subscribes to or endorses conspiracy theories. This is a novel and exploratory concept, necessitating the resolution of problems regarding bias and the dependability of generative models. However, as LLMs are incorporated into moderation processes, their function may expand to evaluating user-level risk alongside content-level evaluation.

In summary, the array of methodologies for predicting potential disseminators of misinformation is varied. Less complex models with carefully selected features provide interpretability and have demonstrated robust performance in research contexts. More intricate models and graph-based methodologies seek to encapsulate the complete intricacy of user behaviour and social context, but at the expense of transparency. The most efficacious solutions in literature frequently integrate various methodologies, mirroring the complex nature of the issue, disseminating misinformation concurrently regarding the content being shared, the individuals sharing it, and the mechanisms of its propagation via the network.

4. FINDINGS FROM DETECTION AND IDENTIFICATION TECHNIQUES

The application of the above techniques in various studies has yielded a number of important findings about users who are likely to repost unreliable news. These findings deepen our understanding of the "misinformation spreader" profile and the dynamics of how false content propagates via user interactions:

4.1 A small minority drives the majority of misinformation spread

A notable feature consistently noticed is that the dissemination of misinformation is markedly uneven. A small percentage of users (about 10−3 to 10−2 of the total) constitutes a significant component of the overall dissemination of bogus news [1, 5]. During the 2016 election on Twitter, around 1% of users encountered over 80% of the false information, predominantly due to their engagement in following and disseminating it [5]. On Facebook, the majority of users did not disseminate any false information, however a minority of users posted numerous fabricated stories [1]. The power-law distribution indicates that interventions aimed at the most prolific disseminators could substantially diminish the overall spread of misinformation. This elucidates the rationale behind prediction initiatives concentrating on the identification of these superspreaders. Upon

identification, these users may undergo intensified fact-checking, debunking initiatives, or, in severe instances, suspension, to diminish their influence. Nonetheless, one caveat is that these users frequently possess significant influence (many followers) or are closely linked to like-minded individuals, thus, indiscriminate moderating may occasionally incite backlash or lead to migration to alternative platforms.

4.2 Superspreaders tend to have distinct profiles and behaviors

Top misinformation spreaders' characterisation reveals they are not random users but rather often have certain qualities. The most active Twitter superspreaders, according to DeVerna et al., were a mix of political pundits or activists with large followings, low-credibility "news" site accounts, affiliated personal accounts connected to those sites, and a variety of other influencers [16]. Primarily political in nature, either far-right or far-left, these accounts showed more poisonous and inflammatory language than average users spreading false information [16]. This fits the gut feeling that those who propagate false news successfully frequently employ polarising or aggressive language, which can drive involvement. These users also exhibit another behavioural marker: they publish extremely regularly, sometimes flooding their followers with a large amount of repetitious or thematically consistent material, for instance, dozens of tweets per day promoting the same false story. High-volume posting gives disinformation superspreaders control of the discussion and raises the possibility that their material will be viewed and shared by others.

Superspreaders might also have a tendency of jumping from one deceptive subject to the next: before elections they promote electoral fraud stories, then they switch to anti-vaccine or other conspiracy material, and so forth. This implies an opportunistic behaviour; such users are usually inclined to disseminate any false assertions that fit their objective or worldview instead of being deceived by one problem. For forecasting, this overall tendency is beneficial: Unless their behaviour changes significantly, a person who has been a large spreader in one domain is likely to do so in future domains of misinformation.

4.3 Demographic and psychosocial correlates: older age, political ideology, and certain personality traits

One constant result, as mentioned in the focal areas, is the age effect: older users on average communicate more false information [1]. Though Twitter's user base skews younger than Facebook's, this has been recorded on Facebook and is probably relevant to Twitter as well. The explanations could include generational variations in media trust or digital literacy, as well as cognitive changes. Political conservatives in the United States have been shown to share more false news than liberals [1], which some say is due to the specific distribution of misinformation during events like the 2016 election being mostly right-leaning material (therefore, conservatives had more of it to share) as well as possibly more exposure to partisan media. Misinformation is not limited to one ideology, hence one should remember that the relationship could change in other settings or nations.

Regarding personality, although big personality data is limited, the research from Shrestha et al. and comparable studies suggests significant variations [20, 21]. A spreader profile, for instance, might relate to greater extraversion (more active on social media, bigger network, thus more sharing) or higher neuroticism (more susceptible to anxiety or anger, which false information usually exploits), and maybe lower conscientiousness (less careful about checking before sharing). These theories fit seen behaviour such as hasty retweeting without truth checking. The "Need for Cognition" or analytical thinking ability is another idea: separate studies by Pennycook et al. (not specifically addressed in this review) revealed that people who performed worse on analytical thinking exercises were more likely to accept and disseminate false information, implying that cognitive reflection is inversely related to vulnerability. This fits with why older people (who on average performed worse on those tasks) could be more vulnerable [1].

4.4 Content and emotional drivers: false news spreads faster due to novelty and emotion

ASeveral studies show that false information frequently draws more user attention and encourages reshares than accurate information. Quantitatively, Vosoughi et al. demonstrated that false news articles spread "significantly farther, faster, deeper, and more broadly than the truth" [4]. They ascribed this to the novelty of fake news: individuals are more inclined to communicate information they find unexpected or alarming, both to inform others and to provoke responses. A bogus narrative can cascade through several user generations (retweets of retweets and so on) before verification catches up, according to this novelty effect. Moreover, as said, erroneous narratives often take use of emotional triggers. Research on the spread of false information throughout the COVID-19 epidemic revealed that tweets calling for moral outrage or terror (e.g., blaming a group for the pandemic, or fanning fear of vaccines) had exceptionally high resharing rates [12]. More broadly, Mousavi et al. showed that on social media, emotionally stimulating (good or negative) material tends to go viral more frequently than neutral information [13]. In the realm of false news, emotional arousal is usually negative (anger, indignation, terror), which drives a sense of urgency motivating people to fast share, maybe without stopping to check.

These content-driven results imply for user prediction that those who are regularly exposed to or interact with high-arousal material may be "primed" to share it. A user's personal timeline material can be examined for how much sensational or emotive content it has; a high frequency could suggest the user lives in an information environment ready for misinfo spreading. Furthermore, this emphasises why teaching people to "think twice" is difficult; the social incentive system (likes, retweets, the psychological gratification of seeing one's message spread) now favours emotionally charged sharing, which false information usually offers. Some actions, such as reminding users when they try to share articles they haven't opened or marking material, seek to reduce this automatic sharing. Work by

Papakyriakopoulos and Goodman in 2022 on Twitter's labelling revealed that when tweets were marked as misleading, their distribution (measured by retweets and likes) was reduced in comparison to analogous unlabelled tweets [11]. This suggests that some of the users ignore alerts and avoid disseminating labelled material, which is encouraging for mitigation plans. On the other hand, hardcore spreaders usually carried on regardless of labels, suggesting that for the most tenacious distributors of false information, more forceful measures could be necessary.

4.5 Efficacy of predictive models and feature importance

Studies on predictive modelling have also produced revelations about which signals most forecast the spread of false information. For instance, Su et al. show that cross-topic prediction is frequently better with user-centric characteristics than with content-centric ones [25]. Models that depended mostly on message content characteristics showed notable performance declines in tests on other topics (e.g., health misinformation) while those that included user behaviour and profile features kept more strong performance. This study indicates that a user's tendency to spread false news is rather constant and not only influenced by the specific subject or content of the news. Some users, therefore, are frequent spreaders regardless of the false news topic. This result underlines the need of user profiling, for example, using their past sharing history or consistent characteristics, as a key component in any long-term remedy.

In the work of Mu and Aletras, feature importance analysis revealed that several language characteristics, such as the use of profanity, which can suggest an aggressive communication style, and engagement statistics (number of likes the user gives/receives) were powerful predictors in their model [24]. The PAN 2020 entries likewise observed that no one feature type was adequate; lexical characteristics reflected topic preferences, sentiment traits reflected emotional style, and user metadata reflected influence/reach; their combination produced the optimal outcomes [20]. Other models provide an intriguing insight: not all spreaders of false information are very political or aggressive in their vocabulary; others seem "normal" in their tweets save for the fact that they sometimes transmit a viral false narrative. For these, network-based characteristics, such as the fact that they follow a lot of extreme accounts or frequently retweet a specific conspiratorial influencer, might be the giveaway. Different predictors therefore catch various subtypes of spreaders, which supports ensemble methods.

Raw performance-wise, state-of-the-art models in controlled tests have achieved approximately 0.75 - 0.80 AUC or F1 scores in spotting known spreaders [20, 24]. Though projections are not perfect, this shows a rather strong signal and room for development. Various factors contribute to the flaw: limited or biassed training data (e.g., labelled spreaders from one time period might not cover all kinds of spreaders), the constantly changing nature of misinformation strategies, and overlap with other behaviours (some genuine users might occasionally share misinformation unintentionally, therefore distorting the model). Furthermore, any predictive system would have to deal with false positives (mislabeling a careful user as a possible spreader might be unfair and damage their reputation) and false negatives (failing to catch a spreader in time).

5. SUMMARY

In summary, research conducted from 2020 onwards has markedly enhanced our comprehension of the mechanisms and motivations behind users' reposting of inaccurate news on X/Twitter. The results consistently indicate that both individual user characteristics and social network dynamics contribute to the dissemination of disinformation. Methods for identifying and predicting probable spreaders have advanced to utilise language indicators, user behavioural patterns, and network architecture, each providing integral components of the whole analysis. Through the integration of various methodologies, researchers have successfully identified high-risk users with notable accuracy and extracted insights regarding their characteristics. The discoveries, including the significant impact of superspreaders, the emotional appeal of incorrect content, and the importance of user-centric features for prediction, are essential for formulating ways to mitigate the dissemination of misinformation. Current and forthcoming efforts will enhance these models, tackle ethical implementation, and adjust to emerging types of misinformation; but, existing literature offers a solid basis for comprehending and predicting the human factors involved in the dissemination of false news on social media.

6. REFERENCES

[1] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science Advances*, vol. 5, no. 1, eaau4586, 2019.

[2] B. Guo *et al.*, "The future of false information detection on social media: New perspectives and trends," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–36, 2019.

[3] K. Shu *et al.*, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[5] N. Grinberg *et al.*, "Fake news on Twitter during the 2016 U.S. presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[6] K. Shu *et al.*, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatial Temporal Information for Studying Fake News on Social Media," *Big Data*, vol. 7, no. 1, pp. 68–79, 2019.

[7] M. N. Shah and A. Ganatra, "A systematic literature review and existing challenges toward fake news detection models," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–21, 2022.

[8] Y. Wang *et al.*, "Do Twitter users change their behavior after exposure to misinformation? An in-depth analysis," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–16, 2022.

[9] Y. Tian *et al.*, "Predicting rumor retweeting behavior of social media users in public emergencies," *IEEE Access*, vol. 8, pp. 87121–87132, 2020.

[10] M. Potthast *et al.*, "A Stylometric Inquiry into Hyperpartisan and Fake News," in *Proceedings of the 56th Annual Meeting of the ACL*, 2017.

[11] O. Papakyriakopoulos and E. Goodman, "The impact of Twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets," in *Proc. ACM Web Conf.*, 2022, pp. 2541–2551.

[12] K. Solovev and N. Pröllochs, "Moral emotions shape the virality of COVID-19 misinformation on social media," in *Proc. ACM Web Conf.*, 2022, pp. 3706–3717.

[13] M. Mousavi *et al.*, "Effective messaging on social media: What makes online content go viral?," in *Proc. ACM Web Conf.*, 2022, pp. 2957–2966.

[14] C. E. Robertson *et al.*, "Negativity drives online news consumption," *Nature Human Behaviour*, vol. 7, no. 5, pp. 812–822, 2023.

[15] S. B. Paletz *et al.*, "Emotional content and sharing on Facebook: A theory cage match," *Science Advances*, vol. 9, no. 39, eade9231, 2023.

[16] M. R. DeVerna *et al.*, "Identifying and characterizing superspreaders of lowcredibility content on Twitter," *arXiv preprint arXiv:2207.09524*, 2022.

[17] Q. Liao *et al.*, "An integrated multi-task model for fake news detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5154–5165, 2021.

[18] S. Y. Lin *et al.*, "Fake news detection model with hybrid features—news text, image, and social context," *Information Systems Frontiers*, 2025.

[19] S. Vilella *et al.*, "Measuring user engagement with low credibility media sources in a controversial online debate," *EPJ Data Science*, vol. 11, article 29, 2022.

[20] A. Shrestha, F. Spezzano, and A. Joy, "Detecting fake news spreaders in social networks via linguistic and personality features: Notebook for PAN at CLEF 2020," in *CEUR Workshop Proc.*, vol. 2696, 2020.

[21] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet prediction based on topic, emotion and personality," *Online Social Networks and Media*, vol. 25, p. 100165, 2021.

[22] X. Zhou *et al.*, ""This is fake! Shared it by mistake": Assessing the intent of fake news spreaders," in *Proc. ACM Web Conf.*, 2022, pp. 3685–3694.

[23] M. Mosleh and D. G. Rand, "Measuring exposure to misinformation from political elites on Twitter," *Nature Communications*, vol. 13, no. 1, pp. 1–9, 2022.

[24] Y. Mu and N. Aletras, “Identifying Twitter users who repost unreliable news sources with linguistic information,” *PeerJ Computer Science*, vol. 6, e325, 2020.

[25] X. Su *et al.*, “Mining User-aware Multi-relations for Fake News Detection in Large Scale Online Social Networks,” *arXiv preprint arXiv:2212.10778*, 2022.

[26] A. Gupta, H. Lamba, and P. Kumaraguru, “Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy,” in *Proceedings of the 22nd International Conference on World Wide Web (Companion)*, pp. 729–736, 2013.

[27] R. Singh and A. Sharma, “Predicting Image Credibility Analysis for Fake News Detection Using Deep Neural Networks,” *Multimedia Tools and Applications*, vol. 81, pp. 12345– 12368, 2022.

[28] Y. Wang *et al.*, “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*, 2018.

[29] P. Bazmi, M. Asadpour, and A. Shakery, “Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility,” *Information Processing & Management*, vol. 60, no. 1, p. 103146, 2023.

[30] J. Wang *et al.*, “LLM-Enhanced Multimodal Detection of Fake News,” *PLOS ONE*, vol. 19, no. 10, e0312240, 2024.

[31] Y. Wang *et al.*, “CCasGNN: Collaborative Cascade Prediction Based on Graph Neural Networks,” *arXiv preprint arXiv:2112.03644*, 2021.

[32] F. Rangel, P. Rosso, and M. Potthast, “Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter,” in *Working Notes of CLEF*, 2020.

[33] A. Giachanou *et al.*, “Leveraging Emotional Signals for Credibility Detection,” *Information Processing & Management*, vol. 57, no. 5, p. 102230, 2019.