

Article

MSDSI-FND: Multi-Stage Detection Model of Influential Users' Fake News in Online Social Networks

Hala Al-Mutair ^{1,2,*}  and Jawad Berri ^{1,3}

¹ Information Systems Department, College of Computer and Information Science, King Saud University, Riyadh 11543, Saudi Arabia; jawad.berri@udst.edu.qa

² Department of Information Systems, Imam Mohammad Ibn Saud Islamic University, Riyadh 11623, Saudi Arabia

³ College of Computing and Information Technology, University of Doha for Science and Technology, Doha 24449, Qatar

* Correspondence: 437204347@student.ksu.edu.sa

Abstract

The rapid spread of fake news across social media poses significant threats to politics, economics, and public health. During the COVID-19 pandemic, social media influencers played a decisive role in amplifying misinformation due to their large follower bases and perceived authority. This study proposes a Multi-Stage Detection System for Influencer Fake News (MSDSI-FND) to detect misinformation propagated by influential users on the X platform (formerly Twitter). A manually labeled dataset was constructed, comprising 68 root tweets (42 fake and 26 real) and over 40,000 engagements (26,700 replies and 14,000 retweets) collected between December 2019 and December 2022. The MSDSI-FND model employs a two-stage analytical framework integrating: (1) content-based linguistic and psycholinguistic analysis, (2) user profiles analysis, structural and propagation-based modeling of information cascades analysis. Several machine-learning classifiers were tested under single-stage, two-stage, and full multi-stage configurations. An ablation study demonstrated that performance improved progressively with each added analytical stage. The full MSDSI-FND model achieved the highest accuracy, F1-score, and AUC, confirming the effectiveness of hierarchical, stage-wise integration. The results highlight the superiority of the proposed multi-stage, influential user-aware framework over conventional hybrid or text-only models. By sequentially combining linguistic, behavioral, and structural cues, MSDSI-FND provides an interpretable and robust approach to identifying large-scale misinformation dissemination within influential user-driven social networks.

Keywords: fake news detection; influential users; machine learning; natural language processing; online social networks; social networks analysis



Academic Editor: Tehmina Amjad

Received: 16 October 2025

Revised: 17 November 2025

Accepted: 21 November 2025

Published: 26 November 2025

Citation: Al-Mutair, H.; Berri, J.

MSDSI-FND: Multi-Stage Detection Model of Influential Users' Fake News in Online Social Networks. *Computers* **2025**, *14*, 517. <https://doi.org/10.3390/computers14120517>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The widespread popularity of social media has transformed the world into a virtual global village. This evolution has encouraged individuals to freely express their thoughts and emotions. On the one hand, the right to freedom of expression has led to a significant increase in online content that, when properly utilized, can enhance human life, promote public awareness, support political campaigns, and serve other purposes. However, its abuse has generated antagonism in online social media, leading to misleading information in fake news, hate speech, propaganda, cyberbullying, etc. Online social networks (OSNs), such as X (Twitter), Facebook, and Weibo, have dramatically changed how our society produces, disseminates, and consumes information. This information directly affects users,

influencing them either positively or negatively and, in turn, triggers specific reactions and behaviors. Importantly, this information may come from an untrusted source, which can disseminate false information to OSNs. Consuming news via social media is frequently timely and less expensive than consuming traditional news media, such as newspapers or television. The effect of fake news and its propagation speed increase when it comes from influential users on online social networks. Misinformation has become a highly active area of research in the scientific community [1]. Misinformation campaigns may affect the opinions and thoughts of news consumers, thereby influencing individual reputations and societal behavior [2].

Although OSNs may offer an enjoyable experience to some, they are not free from deceptive forms of news such as fake news, hoaxes, and rumors. This fake news can harm human communities, particularly in political campaigns, health, sports, and other fields. Research indicates that fake news on X is typically retweeted and engaged with (i.e., in terms of likes, mentions, and replies) significantly more than the verified information [3]. The impact of fake news may increase if it is published by an influential user. Furthermore, influential users can intentionally leverage their popularity and social influence to disseminate fake news, thereby swaying followers' opinions. To illustrate the effect of fake news on OSN users when influential individuals publish it, consider the following tweet example of US President Donald J. Trump about the 2020 US presidential election, which was posted on 8 January 2021: "The 75,000,000 great American Patriots who voted for me, AMERICA FIRST, and MAKE AMERICA GREAT AGAIN, will have a GIANT VOICE long into the future. They will not be disrespected or mistreated in any way, shape, or form!!!" This tweet received thousands of retweets, likes, and replies, and was interpreted as a message to his supporters not to accept the results of the presidential election. As such, it may have prompted his supporters to target the inauguration ceremony, considering it a "safe target" since he would not be in attendance. Twitter judged that such tweets may encourage others to replicate violent acts and that this tweet, in particular, was likely to motivate and inspire individuals to repeat the criminal acts that took place at the US Capitol on 6 January 2021 [4].

Fake News Detection (FND) is an emerging area of research that has attracted tremendous interest, with researchers seeking to understand and identify its typical patterns and features. A community of scholars from various disciplines have proposed potential solutions to combat the spread of fake news and conspiracy theories on OSNs. The solutions were divided into two categories: automated and human-based. Human-based solutions are fact-checking websites that debunk fake news disseminated on OSNs. Examples include Snopes (<https://www.snopes.com>, accessed on 10 October 2025), FactCheck (<https://www.factcheck.org>, accessed on 10 October 2025), PolitiFact (<https://www.politifact.com>, accessed on 10 October 2025) and other websites that rely on human effort. It is not always easy for humans to verify new claims as rapidly as they are published in today's OSNs. Similarly, a set of fake news detection systems, including TweetCred, has been proposed [5]. Similarly, some leading social media platforms, including Facebook and X, have built their FND tools. X established a tool to check the trustworthiness of URLs embedded in tweets. Facebook adds warnings to posts that may have verified claims [6]. Therefore, the FND is a challenging task. Consequently, different computerized solutions have been proposed to address this issue. The most commonly used strategies primarily incorporate Artificial Intelligence (AI) techniques such as Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) models. FND methods can be divided into four categories depending on the factors that are focused on: the false knowledge carried, writing style, how the news propagates among users, and credibility of its source (user profiles) [7,8]. Despite these advancements, a noticeable

research gap still exists in understanding how misinformation disseminated by influential users propagates through social networks.

Most existing studies have concentrated on content-based detection approaches that rely solely on textual information or deep learning architecture such as BERT and BiLSTM, which classify the authenticity of individual posts. However, these models overlook critical dimensions of misinformation, such as user influence, audience engagement behavior, and propagation dynamics, which play an essential role in shaping the reach and credibility of fake news. To bridge this gap, the present study introduces a Multi-Stage Detection System of Influential Users' Fake News in Online Social Networks (MSDSI-FND) that integrates content, creator -influential user-, and propagation features to analyze both the linguistic and behavioral aspects of fake news diffusion. This framework specifically targets influencer-originated misinformation cascades on the X (Twitter) platform, providing a more holistic and realistic perspective on how fake news evolves and spreads in online environments. In this study, multi-stage' refers to a hierarchical detection framework in which each stage performs a specific analytical task: content evaluation stage, and user-influence profiling with propagation analysis stage. These stage-wise representations are subsequently fused within classical machine-learning classifiers. Although the final training phase operates as a unified classification step, the architectural and feature-engineering workflow preserves a hierarchical, stage-based organization, consistent with the multi-stage concept of layered information integration.

The primary contributions of this study are as follows:

- A novel English language dataset collected from X based on influential users' tweets related to COVID-19 and their followers' engagements and manually labeled as fake/real news tweets.
- Exploratory data analysis was applied to the created dataset using in-depth statistical analysis to understand and visualize the proposed dataset.
- (MSDSI-FND): A proposed multi-stage detection system for online fake news disseminated by influential users was developed and tested with promising performance for detecting fake news.

The remainder of this paper is organized as follows. Section 2 presents the background of fake news concepts, approaches, and techniques, followed by a review of the literature related to fake news detection. Section 3 presents the methodology used in the study, including an exploratory data analysis of the created dataset through in-depth statistical analyses and the proposed (MSDSI-FND) system, which is based on the collected dataset. Section 4 discusses the results of applying the proposed system. Finally, the paper concludes in Section 5, which offers concluding remarks and discusses pathways for future research.

2. Related Work

This section presents the most commonly used definitions of fake news, related concepts, and the techniques used to tackle and detect fake news in the literature.

2.1. Types of Fake News

The openness of social media allows users to communicate and exchange information quickly, making them vulnerable to misleading information and other suspicious online activities. Misleading information typically includes unverified information or news about upcoming events. Furthermore, they can affect people's attitudes, opinions, and behaviors in different ways. Misleading information contains other related terms, such as false/fake news, satire news [9], rumors [10], disinformation [11] and misinformation. This overlap of concepts is known as information disorder; disinformation, misinformation, and malinformation are the three primary types of information disorder. Misinformation

and disinformation relate to false or inaccurate information, but the critical difference is whether the information was created to be misled. Disinformation refers to intentional misinformation, whereas misinformation refers to accidental causes. Similarly, genuine information disseminated to cause damage is referred to as malinformation [12]. A Google Trends Analysis of the term reveals that it was heavily used during the 2016 US presidential election [13]. Additionally, there is no universal definition of fake news. Nevertheless, one of the most widely used definitions in academia is Stanford University, which defines fake news as “the news articles that are intentionally and verifiably false and could mislead readers” [14]. The broader definition introduced by Sharma et al. [15] stated that fake news refers to “a news article or message published and propagated through media, carrying false information regardless of the means and motives behind it”.

2.2. Fake News in OSNs

Recently, people have turned to social networking sites as a resource to stay informed about the changing world and the current news. Social media has become a popular platform for seeking information and consuming news. Individuals are encouraged to freely express their thoughts and emotions. It even went further than that, as some people decide, plan, and build their decisions and life matters based on the news and information the site markets publish, and present as facts that cannot be thoroughly questioned or debated. In recent years, fake news has become a significant focus, particularly in the study of online social media. This is reflected in the number of articles written on the role of social media in spreading misinformation and the prevalence of manipulated video clips. Following the 2016 US presidential election, the lack of quality control mechanisms on social media has received special attention. Studying, analyzing, and combating fake news is vital from a scientific perspective, not just an “informative” or “journalist” perspective, as some may expect. The research community is concerned with studying the types of fake news, and their associated spread and impact. For example, in 2016, a tweet claimed that President Obama had been injured by an explosion, resulting in a financial stock crisis that wiped out \$130 billion in stock value in minutes. Another example showing how false beliefs can jeopardize public health is a case in Iran during the COVID-19 pandemic, where an article stating that drinking methanol would cure COVID-19 resulted in approximately 500 deaths [16]. It is worth noting that some malicious accounts contribute to the dissemination of such news for propaganda.

In some cases, these accounts are not operated by humans; instead, so-called “social bots” are controlled by computer algorithms to produce content automatically and interact with humans (or other bot users) on social media [17,18]. Moreover, social media users tend to form groups of like-minded people who polarize their opinions, resulting in an echo chamber effect [19].

Another social theory that accounts for how individuals’ behaviors and beliefs are affected in OSNs is the “filter-bubble effect.” This refers to the intellectual isolation that occurs when social media websites use algorithms to personalize information for a user based on a prediction of what they want to see, which can reduce their exposure to opposing viewpoints [20]. Fake news spreads in OSNs because every user can be a source (creator) or spreader (by retweeting or reposting) of information. These actions increase the possibility that other users on each network will see their contributions. Furthermore, in OSNs, the “circle of acquaintances” is larger than that in offline life, which may reflect the reality that people tend to agree with the opinions of their friends, regardless of facts.

2.3. Fake News Detection Approaches and Techniques

Owing to the growth of OSNs in recent years, the distribution of fake news online has increased substantially. The information exchanged is numerous, rapid, and heterogeneous. Faking information in online communities can have a profound impact on society. Consequently, more researchers have focused on identifying false information and fake news. Various computational detection techniques have been studied and have yielded promising results. Several state-of-the-art FND approaches have been discussed in this section. Figure 1 illustrates these approaches.

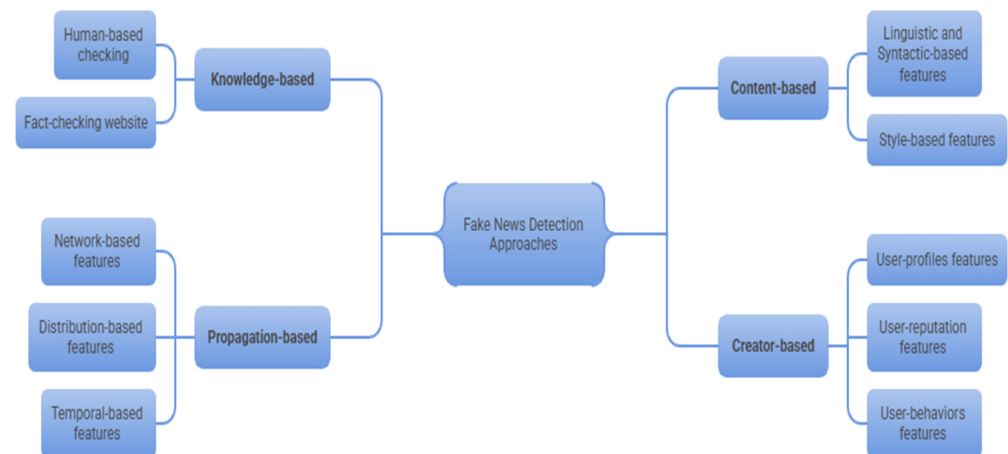


Figure 1. Fake News Detection Approaches.

2.3.1. Knowledge-Based Fake News Detection

Knowledge-based FND is based on a fact-checking process that assesses the authenticity of news by comparing the extracted knowledge from the to-be-verified news content with known facts. The fact-checking process can be performed either manually or automatically. Manual fact-checking uses one or more references: domain experts and fact-checking websites (e.g., PolitiFact, Snopes, The Washington Post) or a crowd of regular individuals who function as fact-checkers (e.g., Amazon Mechanical Turk). Table 1 summarizes some popular fact-checking websites that are trending because of the massive volume of posts on OSNs. It relies on Information Retrieval (IR), NLP, ML techniques, and network/graph theory [21].

Table 1. Fact-checking websites.

Website	Theme	Analyzed Content	Classification Labels
FactCheck.org	American politics	News, interviews, debates, TV shows, and speeches	True, No evidence, False
PolitiFact	American politics	Statements	True, mostly true, Half true, mostly false, False, Pants on fire

Table 1. Cont.

Website	Theme	Analyzed Content	Classification Labels
Snopes	Politics and other general topics	News articles and videos	True, mostly true, Mixture, mostly false, False, Unproven, Outdated, Miscaptioned, correct attribution, misattributed, scam, legend
The Washington Post Fact Checker	American politics	Statements and claims	One Pinocchio, Two Pinocchio, Three Pinocchio, Four Pinocchio.
Misbar	Claims on social media (specifically in the Middle East, North Africa, and the US)	Statements and claims	Fake, Misleading, True, Myth, Selective, Suspicious, Commotion, and Satire

2.3.2. Content-Based Fake News Detection

This approach seeks to analyze news content based on an evaluation of its intention (i.e., whether it was created to mislead the public). Figure 2 presents the content feature taxonomy. One of the content-based feature categories used in this approach comprises linguistic and syntactic-based features. These apply to natural language structure and semantics and can be classified into three levels: word-level features such as bag-of-words, n-gram, term frequency (TF) [22], and term frequency-inverted text frequency (TF-IDF) [23]; sentence-level features such as parts of speech (POS); and content-level features, which refer to the basic information of the meta news content (e.g., location, educational background, and habits) [24,25].

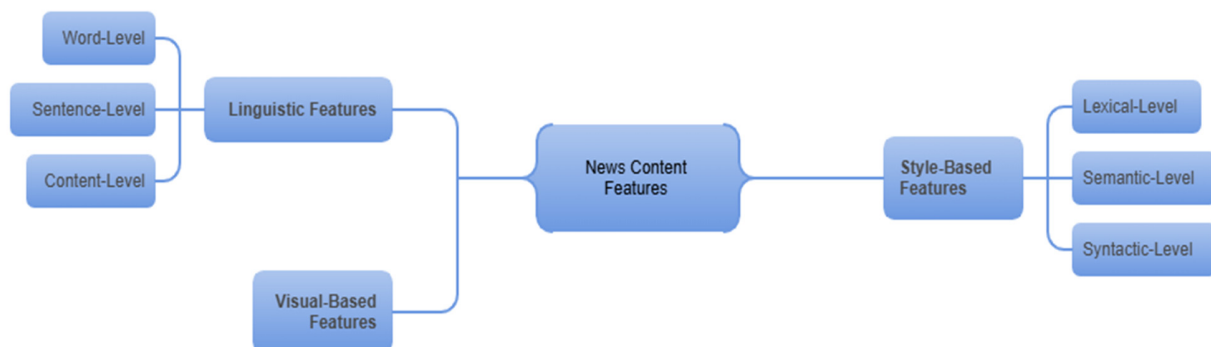


Figure 2. News content features.

Style-based features are another category of content-based features that reveal the distinct characteristics of the writing styles used by fake news authors. Style-based features are grounded in self-defined (non-latent) ML features that can represent fake news and differentiate it from the truth (e.g., word-level statistics based on TF-IDF, n-grams, or Linguistic Inquiry and Word Count LIWC features) [26]. The last category comprises visual-based features related to images or videos in news content. With a content-based approach, fundamental social and forensic psychology theories have not played a significant role. By highlighting potential fake news trends and fostering explainable ML models for FND, such theories can significantly enhance FND.

2.3.3. Creator/Source-Based Fake News Detection

This approach captures the specific characteristics of suspicious user accounts or bots (i.e., non-human) accounts to determine the credibility of the fake news source. The approach is divided into three categories: user profiling features (account name, geolocation information, and verified user), features of user reputation (number of followers and credibility score), and features of user behavior for both deceptive users and regular/valid users [27–29]. Credibility is often defined as “offering reasonable grounds for being believed” [30].

2.3.4. Propagation-Based Fake News Detection

Propagation-based social context-based features (FNDs) exploit social context information to detect fake news (e.g., how fake news posts propagate on social networks, who spreads fake news, and how spreaders connect). Some studies have further divided propagation-based features into three categories: network-based, distribution-based, and temporal-based. Network-based features focus on an online user community and discuss their similarities or differences from various viewpoints, such as their preferences, location, or level of education. Distribution-based features can help extract distinct dissemination patterns of online news [31].

In addition, time-based features were applied to describe the behavior of posting online news initiators in a time series. Examples of time-based features include the time interval between the two posts, posting frequency, reply rate, and comment rate for a specific account [32]. It is also possible to extract contextual characteristics by considering relevant information related to social media posts or fake news. The most frequently used context features are the analysis of users, sources of rumors or news, structures for disseminating social media information, and other users’ reactions to news [31–33].

In conclusion, the previous subsections (Sections 2.3.1–2.3.4) briefly discussed FND approaches from four perspectives: the false knowledge they contain, their writing style, the creators/users who post them, and their propagation networks. The content-based approach primarily focuses on the textual and visual content of the news during the detection process. Table 2 summarizes the differences between the aforementioned FND approaches. Furthermore, unlike traditional media outlets, OSNs are widespread, exacerbating the problem because they are not subject to censorship. Although it is crucial to identify trustworthy and untrustworthy users in OSNs, this is outside the scope of the proposed study. Instead, this study investigates fake news published by influential users on OSNs and examines how it influences their followers. The credibility score, which refers to “the quality of trustworthiness,” can strongly indicate whether a user is trustworthy, whether they have shared fake news, and whether their account is malicious or normal. Automatic detection of fake news is vital.

Many existing methods have contributed to identifying fake news tweets in X. Therefore, there is a need for more powerful techniques to study and analyze such information from different viewpoints. In addition to examining the content of fake news, exploring the context and dissemination network of this information is valuable (e.g., studying user engagement, creator/spreader profile characteristics, and patterns of network diffusion and propagation). Based on the extracted features, an ML model was trained to classify news as real or fake. The performance of each detection approach is measured in terms of its extracted set of features. Furthermore, DL is a form of ML that can learn unsupervised learning from unstructured datasets. Multiple hidden layers are used to progressively extract higher-level features from the input. The two main deep learning architectures for FND applications are convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [34].

Table 2. FND approaches comparison.

FND Approach	The Goal	The Context	Features Set
Knowledge-based	To assess the authenticity of the news	News content	Textual and visual features set
Content-Based	Analyze news content based on an evaluation of its intention	News content	Textual and visual features set
Creator/Source-Based	Captures the specific characteristics of suspicious user accounts or bots	User profile	Users' profile data: implicit and explicit
Propagation-Based	Exploits social context information	User social engagements	User posts, retweets, likes, and temporal features

Moreover, FND has been studied from diverse computing perspectives, including ML and NLP. NLP techniques have recently emerged as some of the most successful approaches for FND. They are primarily concerned with helping computers to understand and manipulate the natural, unprocessed language that humans use for communication. NLP enables machines to perform various tasks related to natural language, ranging from parsing and POS tagging to sentiment analysis, stance analysis, machine translation, and dialog system [35]. Nonneural network techniques have been widely used in various NLP tasks. These studies encompassed a range of diverse features and methods. Some studies have primarily used discriminatory features, whereas others have used stylistic features to identify fake news on X such as the BOW model, n-grams, and POS tags. Despite their simplicity and age, these solutions have been successfully used in FND. Other studies have used word and character embeddings.

Most existing studies focus on news stories' text content (e.g., news headlines and body text), but few studies have investigated image and video content. Early studies in this field, undertaken by [23], extracted textual features from news content to build classifiers to identify fake news. Examples of these features include emoticons, symbols, sentiments, and pronouns. By contrast, BOW and n-grams are the most commonly used methods for representing news texts. Another study [36] proposed a supervised ML classification model and word-based n-gram analysis to automatically classify Twitter posts into credible and non-credible. They employed five different classifiers, and the best performance (84.9% accuracy) was achieved using a combination of unigrams and bigrams, LSVM as the classifier, and TF-IDF as the feature extraction technique. Similarly, the work in [37] proposed a benchmarking pipeline for FND based on linguistic features and ML classifiers on UNBiased (UNB) dataset, which reached up to 95% accuracy.

It is essential to recognize that traditional ML classifiers (e.g., SVM, DT, and RF) depend on feature engineering, which is typically time-consuming and labor-intensive. Therefore, researchers have begun using DL techniques for FND to overcome the difficulties of using handcrafted features. Several studies have reported that DL techniques outperform ML techniques in detecting fake news on OSNs [32]. Recent research [38] indicates a distinct shift towards transformer-based models (BERT, GPT, RoBERTa), hybrid neural networks (e.g., BiLSTM with attention), and ensemble techniques that integrate various architectures to enhance performance.

State-of-the-art models now leverage attention mechanisms, ensemble learning, and large language models (LLMs) to achieve high performance, with some reporting F1-scores

and accuracies exceeding 95% on benchmark datasets such as the work presented in [39]. Authors in [40] extracted features based on sentiment analysis of news articles and emotion analysis of users' comments related to the news in Fakeddit dataset, and then fed them to a proposed bidirectional long short-term memory model (BiLSTM) and obtained a high detection accuracy of 96.77%.

A recent study was introduced by [41], who built a hybrid model combining Convolutional Neural Networks and LSTM, enriched with FastText embeddings. Later, they tested the model on three different datasets: WELFake, FakeNewsNet, and FakeNewsPrediction, and surpassed other techniques in classification performance across all datasets, registering accuracy and F1-scores of 0.99, 0.97, and 0.99, respectively. [42] developed a multitask framework to predict the emotion and legitimacy of a text using different deep learning models in single-task (STL) and multitask settings for a more comprehensive comparison and achieved high accuracy. In [43], the authors combined news content with user behavior using a proposed DL approach that incorporated BERT-CNN-BiGRU-ATT and BERT base-BiGRU-CNN-ATT for fake news detection on the FakeNewsNet dataset and achieved a high accuracy of approximately 91.37%.

The authors in [44] developed an exceptional transformer architecture, which consists of two parts: the encoder part that learns useful representations from fake news data and the decoder part that predicts future behavior, and proposed a novel deep neural framework for FND that is divided into three parts: the news content, social context, and detection. Wang et al. [45] fine-tuned BERT for COVID-19 fake news detection by adding BiLSTM and CNN layers; their best model (BERT-based) achieved state-of-the-art performance on that specific dataset. Further, Azizah et al. [46] compared several transformer models: BERT, ALBERT and RoBERTa and found that ALBERT slightly outperformed the others (accuracy ~87.6%) on their Indonesian fake news dataset, underscoring the value of transformer fine-tuning across languages.

Another study by Raza et al. [47] compared BERT-style encoder models and larger large-language-models (LLMs) for fake news detection, demonstrating that while LLMs have promise, fine-tuned encoder models still tend to outperform them on classification tasks in this domain. However, these studies largely remain content-centric, focusing on the text of news items and/or accompanying visuals, and seldom extend to influencer-centric propagation patterns or cascade/network features, which is the focus of the present work. In summary, Table 3 presents an overview of all related studies discussed in this manuscript. To sum up, the field of fake news detection is changing quickly. The next generation of detection systems promise to use cutting-edge AI methods along with ideas from social sciences and network analysis to make digital information ecosystems more reliable and trustworthy.

Table 3. Summary of Related Works on Fake News Detection (FND).

Ref.	Technique	Features	Dataset(s)	Accuracy
[36]	LSVM + TF-IDF	Unigrams, bigrams	Twitter posts	84.9%
[37]	ML benchmarking pipeline	Linguistic features	UNBiased (UNB)	95%
[40]	Bi-LSTM	Sentiment and emotion from news + comments	Fakeddit	96.77%
[41]	CNN + LSTM + FastText	Textual + semantic	WELFake, FakeNewsNet, FakeNewsPrediction	0.99, 0.97, 0.99 respectively

Table 3. Cont.

Ref.	Technique	Features	Dataset(s)	Accuracy
[43]	BERT _{base} -CNN-BiGRU-ATT	Content + User behavior	FakeNewsNet	92%
	BERT _{base} -BiGRU-CNN-ATT			87%
	BERT _{base} -CNN-BiLSTM			91%
[44]	Transformer encoder-decoder	Content, social context, detection	NELA-GT-19 and Fakeddit	74%
[45]	Fine-tuned BERT + CNN + Bi-LSTM	COVID-19 tweets	COVID-19 dataset	80%
[46]	BERT, ALBERT, RoBERTa	Text features	Indonesian dataset	87.6%

3. Methodology

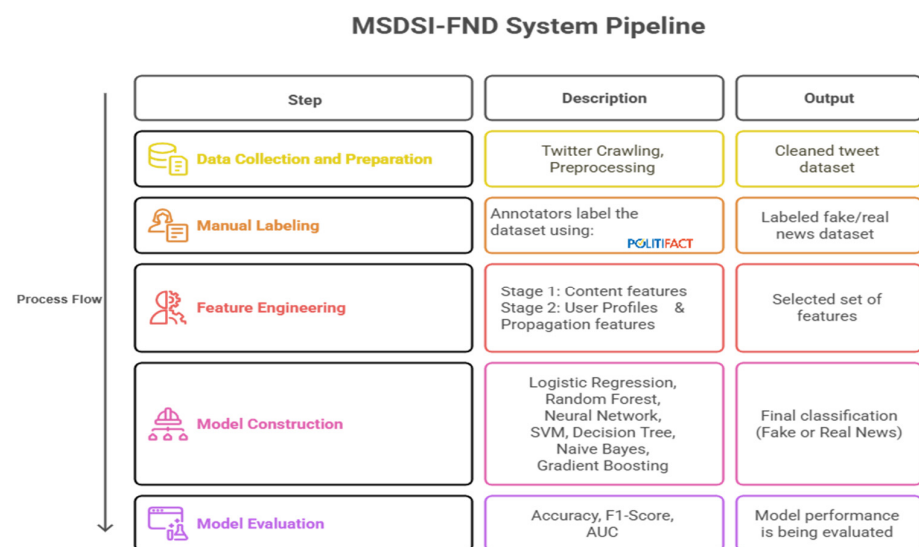
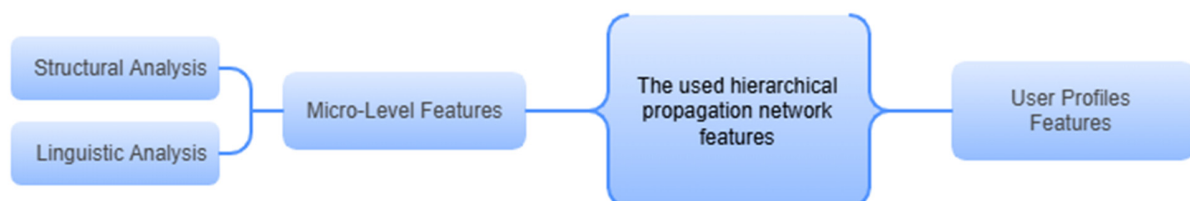
The originality of this study is twofold. The first lies in its focus on fake news generated and disseminated by a distinct category of users—namely, influential users. These individuals possess thousands or even millions of followers who engage frequently with their content through likes, retweets, mentions, and replies. Such engagement amplifies the visibility and perceived credibility of their posts. Consequently, fake news originating from influencers tends to spread more rapidly and persistently than that from ordinary users. Detecting misinformation produced by influential users is therefore a critical step in strengthening fake news detection (FND) mechanisms, as it allows the analysis of both content and behavioral dimensions linked to these high-impact sources.

The second aspect of this study's contribution lies in the development of a multi-stage detection framework: Multi-stage Detection System for Influential Users Fake News (MSDSI-FND) that integrates key analytical dimensions: the content of fake news posts and the profiles of influential users, with the associated propagation networks that reveal diffusion structures and user-sharing behaviors, along with the social context captured through interactions, sentiments, and engagement patterns. This integrated framework introducing a multi-stage analytical architecture that reflects the real-world evolution of misinformation, from its creation to its amplification and eventual diffusion across the network. Figure 3 shows the overall methodology of this study; its different phases and tasks are described in the subsequent sections. Various types of features as represented in Table 4 and Figure 4 were extracted. Simultaneously, a comparative analysis is applied between fake news and real news published by influential users to investigate whether significant differences may enhance the detection process. MSDSI-FND system is built as outlined in the following steps

1. Construct a labeled dataset from X specific to this study, as it will be described in detail later.
2. Analyze and model the content-based features of fake news tweets shared by influential users by examining their tweets and followers' replies.
3. Analyze and model influential user profiles and social context features related to fake news by exploring users' sharing behaviors (followers) along with news tweets cascades.
4. Develop and evaluate a detection system that leverages extracted features to identify fake news shared on OSNs by influential users.

Table 4. The selected feature types.

Feature Type	Feature Description
Content-Based	Used to extract the textual features of influential users' fake/real news posts.
Hierarchical Propagation	-Explicit user profile features include the basic user information from X API (e.g., account name, registration time, verified account or not, and number of followers). -Implicit user profile features can only be obtained with different psycholinguistic patterns tools, such as LIWC. -The distribution pattern of online news and the interaction between online users, as the features presented in [14] and in [48]

**Figure 3.** Proposed MSDSI-FND Pipeline.**Figure 4.** The used hierarchical propagation network features.

3.1. Data Collection and Preparation

The first stage in developing the proposed model involved constructing a reliable and representative dataset. The data were collected from the X platform and consisted of fake and real news posts shared by influential users concerning COVID-19 and its vaccines.

3.1.1. Data Collection

The collection period spanned from December 2019 to December 2022, covering the key phases of the pandemic when misinformation was particularly widespread. The dataset was collected using the X API, based on a predefined set of keywords and seed hashtags related to COVID-19, vaccination, and public health discourse. These keywords were organized in a custom-built dictionary designed specifically for this study. Alongside tweet content, the profile attributes of the posting users such as follower count, number of likes, and verification status were also retrieved.

Additionally, engagement data including replies, retweets, and likes were collected to capture the social propagation and interaction dynamics associated with each tweet. The focus on influential users was deliberate and based on clear methodological rationale. Influencers were selected according to three main criteria:

1. Verified status on X, ensuring that the accounts belonged to authentic public figures rather than bots or anonymous users.
2. Public-figure identity, such as politicians, journalists, scientists, or media personalities, whose posts naturally receive wide visibility and impact public discourse.
3. Follower count exceeding 10,000, ensuring that each selected account had substantial audience reach and engagement potential.

This selection strategy aligns with previous research emphasizing that misinformation originating from influential users tends to spread faster and reach wider audiences than content shared by ordinary users. By focusing on verified, high-reach accounts, the dataset captures the real-world influence dynamics that drive viral misinformation cascades, thereby providing a more realistic and meaningful foundation for the proposed detection model.

3.1.2. Data Preprocessing

Once the data is collected, it undergoes a series of preprocessing steps to clean and structure it for subsequent analysis. This includes:

- Text Cleaning: Removing URLs, mentions, hashtags, special characters, and punctuation from the tweet text.
- Tokenization: Splitting the text into individual words or tokens.
- Stop Word Removal: Eliminating common words (e.g., “the,” “a,” “is”) that do not contribute significantly to the meaning.
- Stemming/Lemmatization: Reducing words to their root form to normalize the vocabulary.
- Lowercasing: Converting all text to lowercase to ensure consistency.

3.2. Manual Labeling

To ensure the accuracy and reliability of labeling, the dataset was manually annotated by two independent human annotators, both with computer science backgrounds and trained in online misinformation assessment. The annotators reviewed each tweet and its associated context before assigning a label of “fake” or “true.” The annotation followed a two-step verification process:

- First, the annotators checked whether the tweet had been flagged or labeled as misleading by X itself. If so, it was categorized as fake news as shown in Figure 5.
- If no platform-based flag was found, the annotators cross-verified the content using PolitiFact.com; a well-established fact-checking organization that evaluates claims based on evidence from multiple reputable sources as shown in Figure 6.

In ambiguous cases, the annotators discussed their reasoning to reach a consensus, ensuring consistency and minimizing subjectivity. This approach helped maintain labeling reliability across the dataset and ensured that each tweet’s classification was grounded in verifiable, credible information.



Figure 5. Twitter classifies influential user tweets as having misleading information.



Figure 6. PolitiFact labels Kevin McCarthy’s statement about San Francisco’s COVID relief funds as False.

3.3. Feature Engineering

In this study, the feature engineering process was conducted in two main phases. The first phase involved feature extraction, where linguistic, user-based, and propagation-based attributes were derived from the dataset. The second phase focused on feature selection, aimed at identifying the most discriminative and statistically significant features that contribute effectively to fake news detection. An independent samples T-test was applied to evaluate the statistical differences between the two classes fake and real news. This test was used to determine which features showed significant mean differences between the two categories, retaining only those with a p -value < 0.05 as significant predictors. This filtering step reduced noise, improved computational efficiency, and enhanced model interpretability by focusing on features with proven discriminative power. Based on the t -test results, the most significant features were selected for each analytical stage. These refined feature sets were subsequently integrated and used to construct the MSDSI-FND model.

3.3.1. Stage 1: Content Features Extraction and Selection

This subsection examines two categories of content-based features used in the analysis: LIWC linguistic features and Exploratory Data Analysis (EDA) features, as detailed below.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a methodology employed to investigate datasets, derive valuable and actionable insights, discover correlations among explanatory variables, identify erroneous data points, and select potential models. It uses descriptive statistics and visualization techniques to enhance data comprehension and interpretation [49]. We employed a univariate graphical technique to analyze individual variables sequentially for categorical data. In contrast, a multivariate graphical method was used to assess two or more variables concurrently and to investigate their interrelationships. In the second case, correlation analysis was employed, which is a technique capable of quantifying the total correlation among two or more numerical variables. The resulting dataset comprises 68 root tweets posted by influential users 42 (61.8%) labeled as fake and 26 (38.2%) labeled as true along with their complete engagement cascades. These posts generated over 40,000 total interactions, including approximately 26,700 replies and 14,000 retweets as shown in Table 5. Figure 7 shows the class distribution of fake and real news root tweets for the same dataset. Such engagement depth provides a rich behavioral and structural foundation for analyzing the temporal, linguistic, and social propagation patterns of misinformation in online social networks. The word and character counts for the replies in both classes were calculated from the dataset as shown in Figure 8.

Table 5. The collected dataset's tweets distribution.

Category	Fake	Real	Total
Root tweets	42	26	68
Engagement Replies	15,362	11,340	26,702
Engagement Retweets	3733	10,327	14,060
Total Engagements	19,095	21,667	40,762

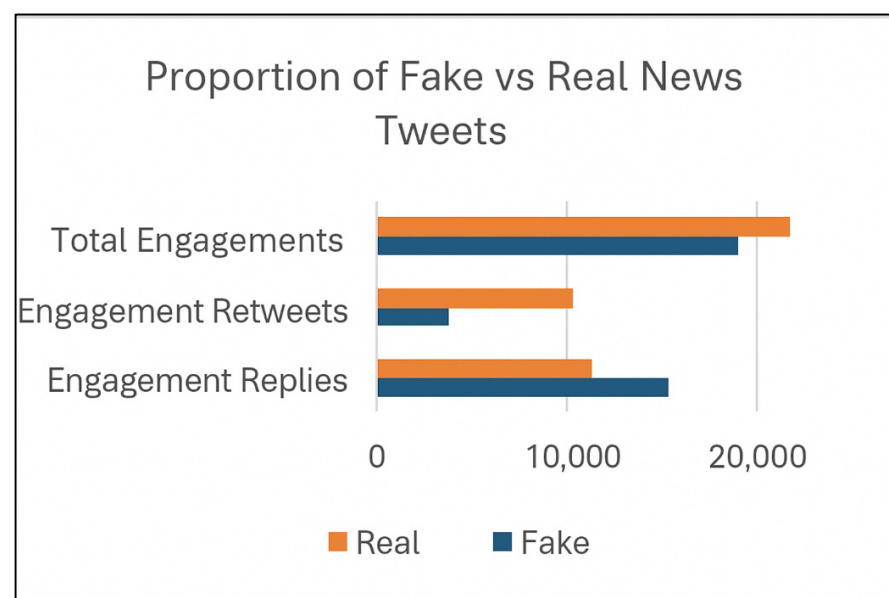


Figure 7. Proportion of Fake vs. Real News Tweets.

The data show that replies to fake news tweets contain a higher average number of words and characters than real news tweets. The average word and character count per reply resembled that of fake and real news. The length of a reply may not serve as a reliable indicator of its relationship to fake or true news. The distinction between fake news and real news tweets was negligible. Fake news often contains more verbose content, incorporating

elaboration or emphasis to enhance its convincing nature and address counter-arguments, rather than simply disseminating information. By contrast, true news is associated with more words and characters in retweets than in replies. This suggests that individuals may be more inclined to disseminate accurate news than to participate in discussions.

The time distributions of fake news and real news retweets and replies are investigated and measured over time, as shown in Figures 9 and 10. Figure 9 shows how fake retweets were shared over time. The pattern is highly inconsistent, with clear spikes at specific times, particularly around significant events such as the COVID-19 outbreak and vaccine debates. These sharp bursts show that fake news spreads quickly and powerfully soon after it is posted, but then it fades just as quickly. In other words, fake content receives attention immediately, often because it is emotional or provocative, but it fails to hold people's interest for a long time. Figure 10 shows how long it took for people to respond to real and fake tweets. The fake tweets (orange bars) show quick, short-lived spikes in activity. Many people respond right away, often with disbelief or strong feelings, and then the conversation fades away. The true tweets (blue bars), on the other hand, receive more steady and consistent replies over time, indicating that the conversations are more extended and more meaningful. In general, these numbers indicate that fake news spreads quickly and has a significant impact, whereas real news spreads more slowly and maintains people's interest. This difference highlights the importance of examining both content and behavioral propagation patterns when identifying false information, particularly when influential users are involved.

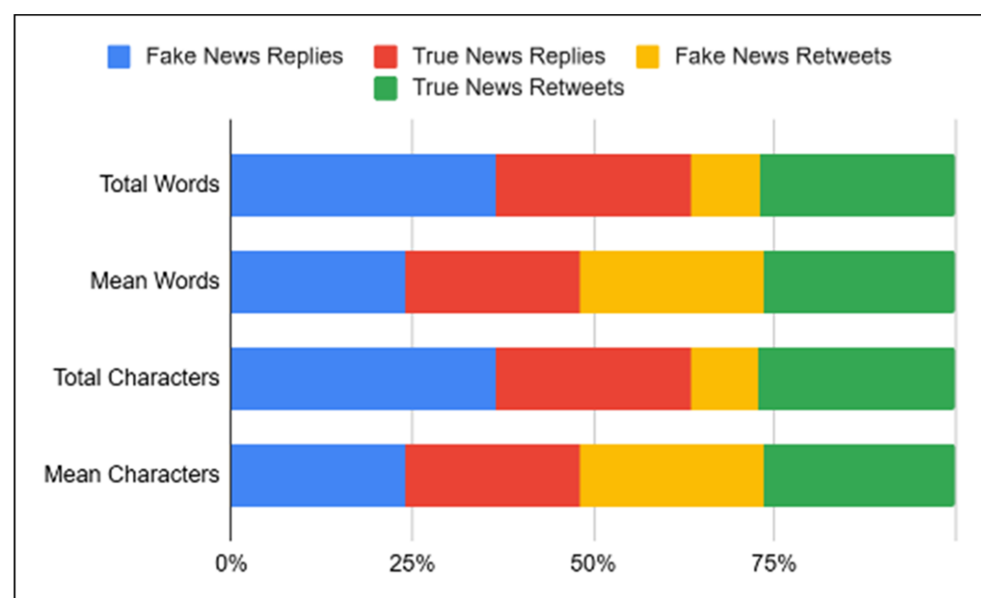


Figure 8. Word counts and character counts for fake/real news replies/retweets.

The dataset also explores various linguistic features, including sentiment analysis, lexical analysis diversity, readability scores, and the use of punctuation and special characters. The sentiment of tweets (e.g., positive, negative, neutral) was measured to examine whether there was a difference between the sentiment of fake and real news tweets. However, for replies and retweets, the sentiment scores for both real and fake news were almost indistinguishable on average. This suggests that sentiment alone may not be a strong differentiator between the two categories in the dataset. Lexical diversity is another measure that is used to assess variations in vocabulary within a text. This dataset shows a minimal difference in lexical diversity between fake and real categories, with true news retweets exhibiting slightly higher lexical diversity and lower variability than fake retweets. The Flesch-Kincaid Readability score and Gunning Fog Index were used to determine the

differentiator between the two categories in the dataset. The Gunning Fog Index estimates the number of years of formal education needed to understand a text; higher scores indicate more complex text. After applying these two measures to the dataset for both parts (replies and retweets), we observed that replies in both the fake and real categories had similar readability levels. This suggests that the complexity of the language used in the replies does not differ significantly. Fake news retweets have a lower mean score than real news retweets, indicating that fake news retweets are slightly easier to read. Furthermore, we examined second-level replies, retweets, and likes for the main replies and r tweet originating from root tweets. Correlations between numerical variables were used to check for relationships. Finally, as mentioned earlier, feature selection was conducted using *t*-tests to retain only the most relevant and effective set of features that contribute meaningfully to the detection process.

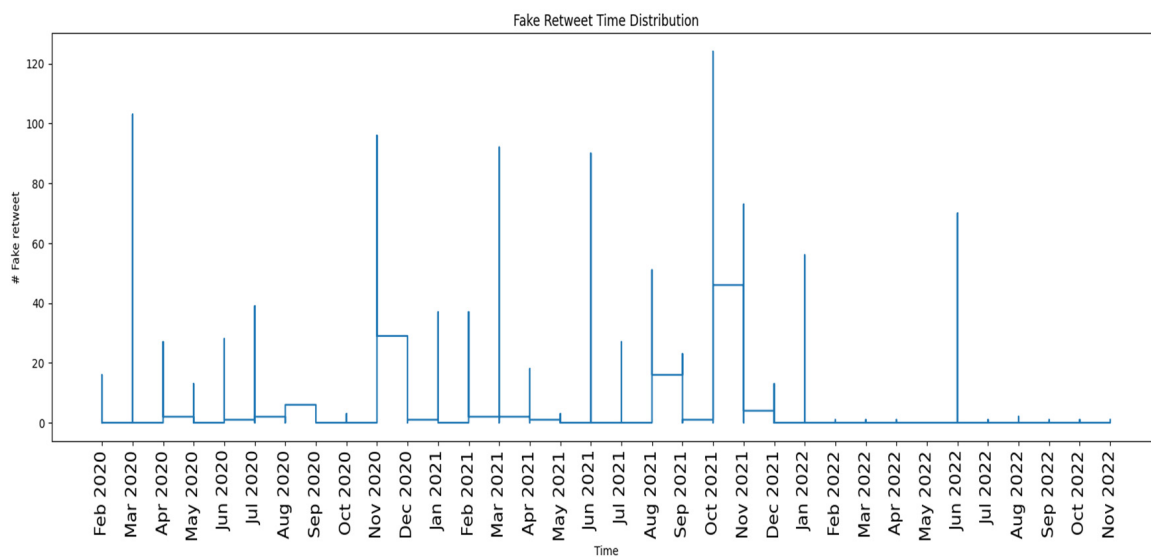


Figure 9. Time distribution of fake-news retweets. The *y*-axis (# Fake retweets) represents the number of retweets recorded at each point in time, showing how frequently fake news was reshared across the timeline.

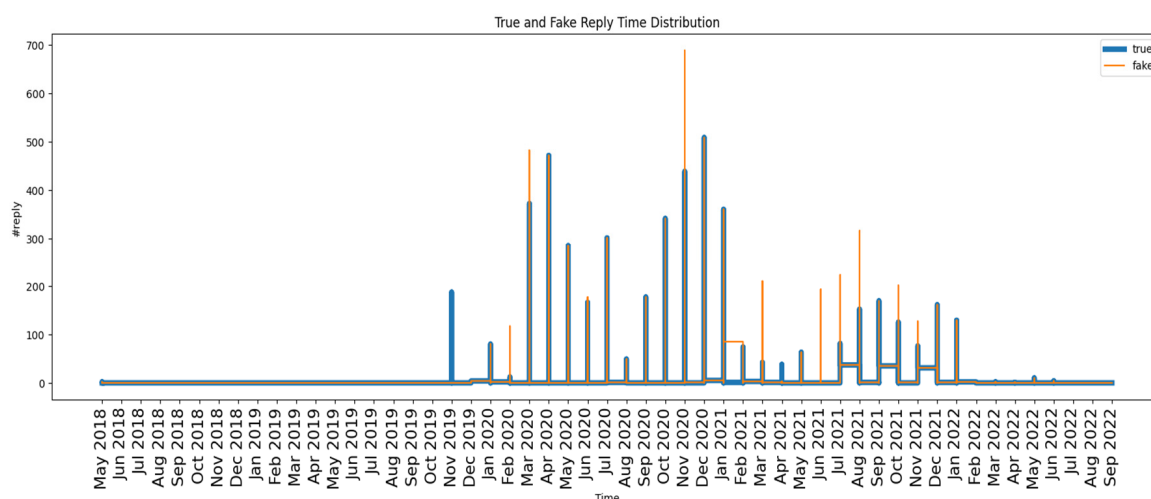


Figure 10. Time distribution of replies for true and fake news. The *y*-axis (# reply) represents the number of replies recorded at each time point, illustrating how audience engagement through replies varies over the observed timeline for both true and fake news tweets.

Linguistic Inquiry and Word Count: LIWC

Another measure used in the dataset was LIWC-22 [50]. When paired with additional features, such as metadata and network analysis, the LIWC can be helpful in identifying fake news. The strength of LIWC lies in its ability to identify linguistic and psychological patterns in texts, which are often distinct in fake news compared with real news. The LIWC-22 analysis highlighted apparent differences in the language style and emotional tone between real and fake news tweets. Fake news tweets scored slightly higher on the Analytic dimension (≈ 49) but lower on Authenticity (≈ 34) compared with real ones (≈ 44 and ≈ 41 , respectively). This pattern suggests that fake news often adopts a structured, formal tone to sound credible, yet the language feels less genuine or personally grounded. When it comes to emotions, real news tweets included a wider range of emotional words (Affect $\approx 7.2\%$) than fake ones ($\approx 5.4\%$). Both positive and negative emotion words appeared more often in real tweets, indicating that authentic sources tend to express empathy or emotional awareness when reporting.

In contrast, fake tweets contained fewer explicit emotion cues, perhaps to maintain a neutral or persuasive facade. Real tweets also used more personal pronouns such as I, we, and you ($\approx 8.9\%$ vs. 6.8%), reflecting a more direct and socially engaging tone. In addition, they featured more cognitive-process words ($\approx 10.9\%$ vs. 8.0%), including expressions of reasoning and uncertainty like because, think, or perhaps. This suggests that real news is more analytical and nuanced, while fake news tends to simplify or overstate information. Regarding motivation and style, real tweets contained slightly higher references to affiliation, achievement, and power, which fits with the communication style of institutional or professional accounts. Interestingly, fake tweets relied more on visual cues—they used more emojis ($\approx 6.36\%$), but fewer exclamation marks and internet slang, possibly to make the message look casual or emotionally appealing. These results show that real news tweets are richer, more expressive, and more socially connected, whereas fake news tweets appear more detached, less authentic, and more visually stylized. Finally, as mentioned earlier, feature selection was conducted using *t*-tests to retain only the most relevant and effective set of features that contribute meaningfully to the detection process.

3.3.2. Stage 2: User Profiles and Propagation Features Extraction and Selection

This study adapts and builds upon the approach outlined in [14] to explore social context-based features (replies to cascades and user profiles features), making necessary modifications to align with the characteristics of our dataset. The proposed model will build a micro-level network (replies-networks) to analyze influential users' followers sharing behaviors related to fake news, for example, whether they support or deny the news content and their sentiments. Later, feature selection was conducted using *t*-tests to retain only the most relevant and effective set of features that contribute meaningfully to the detection process as shown in Table 6. Structural features show how tweets (and their replies) spread through the network by showing the shape and depth of information propagation and how people act when they are part of engagement cascades. User-based features describe the social and influence traits of the users in each cascade, especially the differences between the source (root: influential user) and participant nodes. Linguistic features look at how the emotions and feelings of replies change over time within each cascade. Incorporating these features provides a deeper understanding of how users interact, how deep the cascade is, and how the sentiment changes from one conversation to the next.

Table 6. The user profiles and propagation features.

Feature Type	Features List
Structural	S10: Tree depth
	S11: Number of nodes
	S12: Maximum out-degree
	S13: Number of cascades with replies
	S14: Fraction of cascades with replies
User Profile	U1: nodes' average friends
	U2: source friends
	U3: nodes' average followers
	U4: source followers
	U5: nodes' average favorites
	U6: source favorites
	U7: nodes average listed.
	U8: source listed.
	U9: nodes' average statuses
	U10: source statuses
Linguistic	L1: Sentiment ratio of all replies
	L2: Average sentiment of all replies
	L3: Average sentiment of first-level replies
	L4: Average sentiment of replies in deepest cascade
	L5: Average sentiment of first-level replies in deepest cascade

3.4. Model Construction

We leveraged the extracted and selected set of features from the previous steps to build our proposed MSDSI-FND model. The proposed FND model combines content-based features (Section 3.3.1) with the social-context features (hierarchical propagation network) features previously described in Section 3.3.2. Seven baseline ML supervised learning classification algorithms have been used to examine the engagement of influential users, as their tweets may contain fake or real news.: logistic regression LR, random forest RF, decision tree DT, support vector machine SVM, naïve Bayes NB, gradient boosting GB, and neural networks NN. Each classification model was trained in a supervised environment using binary-labeled tweets. All experiments were performed using a combination of feature sets extracted from the dataset.

3.5. Model Evaluation

Different performance metrics were used to evaluate the classifier's performance. *Accuracy*, *Precision*, *Recall*, *F1-score*, and *AUC* (Area Under the Curve) were the chosen metrics for this dataset. Where *TP* (true positive) is the number of correctly classified real news tweets, *TN* (true negative) is the number of correctly classified fake news tweets, *FN* (false negative) is the number of real news tweets incorrectly classified as fake news, and *FP* (false positive) is the number of fake news tweets incorrectly classified as real news.

Precision measures how many of the tweets predicted as tweets contain real news (positive class) are actually real, and it has the following equation:

$$Precision = \frac{TP}{TP + FP}, \quad (1)$$

where *Recall* measures how many of the actual real news tweets were correctly identified by the model.

$$Recall = \frac{TP}{TP + FN}. \quad (2)$$

Accuracy is the most widely used metric for evaluating a classifier and has the following equation:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}. \quad (3)$$

The *F1 score* is defined as the weighted average of the *FN* and *FP*, as shown in the following equation:

$$F1\ Score = 2 \times \frac{Precision + Recall}{Precision \times Recall}. \quad (4)$$

Initially, the dataset was split based on the train-validation-test split process. The training set consisted of 70% tweets, and the remaining 30% was used equally for validation and testing (15% for both). In addition, to reduce the risk of overfitting, data leakage, and to ensure that the models genuinely generalize to unseen data, we applied user-level stratification when splitting the dataset. In this setup, all tweets and engagement data from a specific influential user appear only in one set either training or testing but never in both. If the same user's posts appeared in both sets, the model could easily "memorize" these patterns, leading to inflated accuracy. By separating users, we make the evaluation more realistic and confirm that the models are actually learning to detect fake versus real news across different users, not within the same author's style. We also used stratified sampling to maintain balanced representation of fake and real content in both splits. Together, these precautions significantly reduce the risk of overfitting.

Furthermore, Table 7 summarized the seven classifiers used in our experiments along with their tuned or selected parameters. All models were trained using the same training set and tuned using the validation set. After identifying the best configuration for each algorithm, final performance metrics (*Accuracy*, *F1*, and *AUC*) were calculated on the independent test set to ensure unbiased evaluation. This setup provides a fair, consistent, and reproducible basis for comparing classifiers and supports the validity of the results presented in this paper.

The MSDSI-FND system performance of applying the combined set of features: the user profiles with propagation features (HT), EDA, and LIWC (HT_EDA_LIWC) results are shown in Table 8. We can observe that the performance gives the best results among all classifiers which demonstrate that fake news can be effectively recognized through the synergy of content and context. The results showed that combining linguistic and psychological content features with EDA features, along with social context and network architecture features (HT) demonstrate strong performance. Moreover, we can see that RF was among the best-performing models ($F1 = 93.1\%$, $AUC = 98.5\%$), as it captured the complex relationships between linguistic features (such as emotional tone, positivity, and negativity) and network features (such as the number of responses and tree depth). These observations indicate that fake news is characterized by a shallow, rapid diffusion structure driven by a few influential accounts, while real news shows a deeper, more interactive diffusion. It is also noteworthy that the neural network (NN) achieved the highest overall performance ($F1 = 96.6\%$, $Accuracy = 96.7\%$, $AUC = 99.4\%$) which reflects its ability to learn the complex and intertwined relationships between language and social behavior. It

seems the network recognized the “fingerprint” of fake news, which combines emotional motivation, influencer style, and collective diffusion behavior.

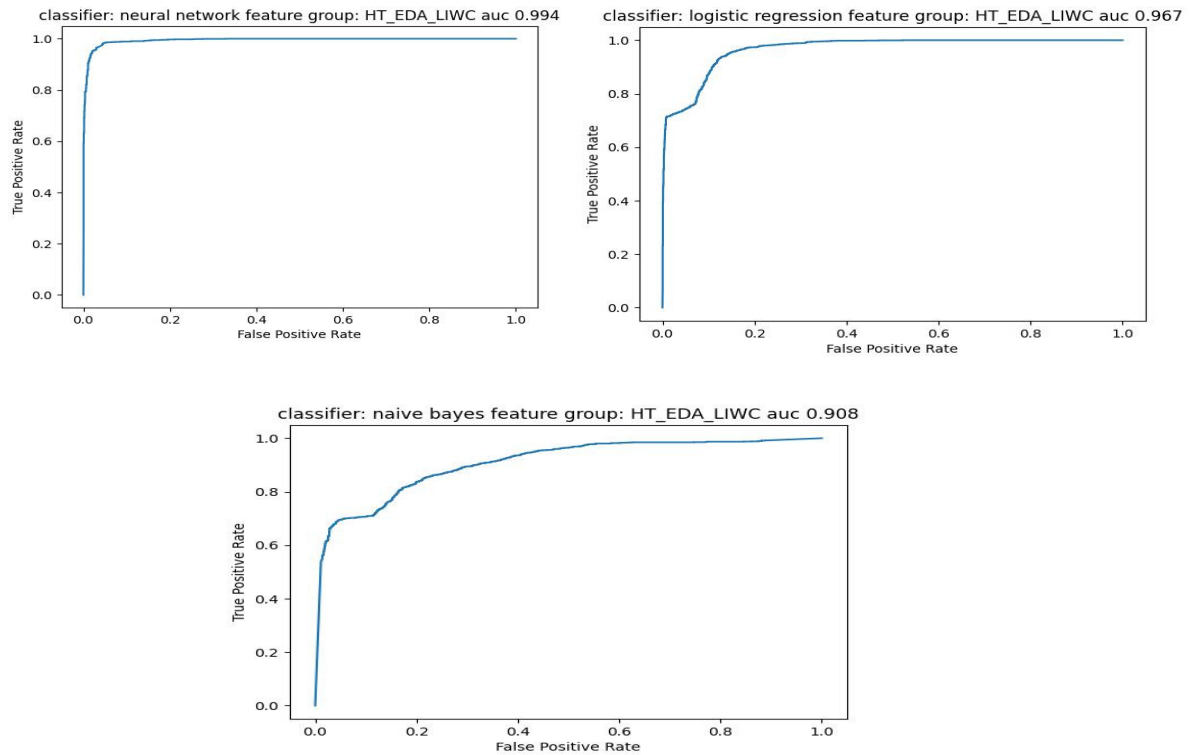
Table 7. The seven selected ML Algorithms’ parameters tuning.

Classifier	Key Parameters	Justification
Logistic Regression	solver = “liblinear”	Suitable for binary classification.
Naive Bayes	default parameters	Used as a simple probabilistic baseline model.
Random Forest	max_depth = 2, random_state = 0	Depth limited to prevent overfitting; random state ensures reproducibility.
Decision Tree	random_state = 0	Default settings used; random state for consistent results.
Gradient Boosting	n_estimators = 10, learning_rate = 1.0, max_depth = 1, random_state = 0	Parameters tuned on validation data for a balance between bias and variance.
Neural Network (MLP)	solver = ‘lbfgs’, alpha = 1×10^{-5} , hidden_layer_sizes = (20, 2), random_state = 0	Small hidden layers and regularization are chosen for stable convergence on limited data.
SVM	kernel = ‘linear’, probability = True, random_state = 0	Linear kernel preferred for interpretability; probability enabled for AUC computation.

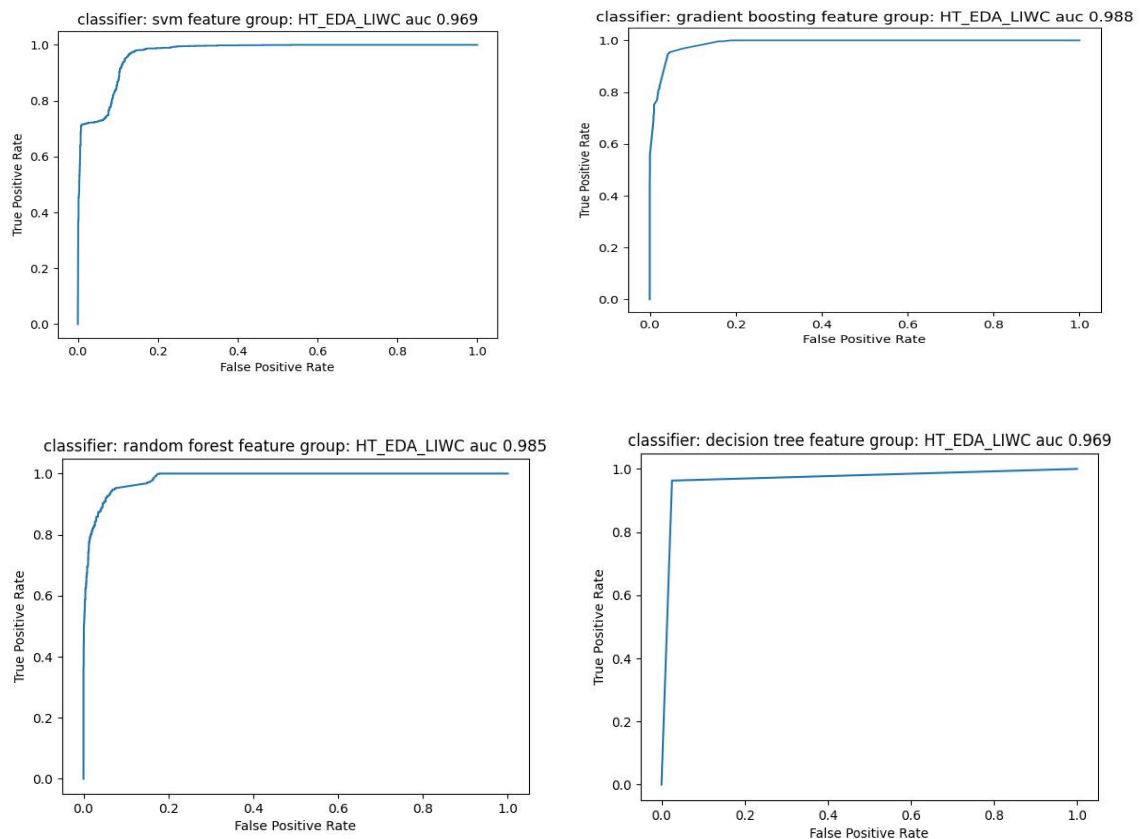
Table 8. MSDSI-FND Performance results.

No.	Algorithm	Features	Metrics		
			F1 Score	Accuracy	AUC
1	LR	HT_EDA_LIWC	87.3	87.7	97.6
2	SVM	HT_EDA_LIWC	87.7	88	96.9
3	DT	HT_EDA_LIWC	62	50.9	96.9
4	RF	HT_EDA_LIWC	93.1	93.2	98.5
5	NB	HT_EDA_LIWC	55.4	58.3	90.8
6	GB	HT_EDA_LIWC	95.2	95.3	98.8
7	NN	HT_EDA_LIWC	96.6	96.7	99.4

Furthermore, Figure 11 represents ROC curves (Receiver Operating Characteristic curves) for all seven classifiers applied to the selected feature set (HT_EDA_LIWC) of MSDSI-FND model. Each plot displays the True Positive Rate (TPR) against the False Positive Rate (FPR) for a given classifier. The closer the ROC curve is to the top-left corner, the better the model’s performance. The AUC value quantifies this performance numerically, where 1.0 represents perfect classification and 0.5 indicates random guessing. The results demonstrate that all classifiers achieve high discrimination power, confirming that the integrated multi-stage features (content + user + structural) provide a rich and complementary signal for distinguishing fake from real news. The NN and GB models yield the best trade-off between sensitivity and specificity, validating the framework’s robustness and its hierarchical, multi-stage feature design.



(a)



(b)

Figure 11. (a) ROC Curves of NN, LR, and NB Classifiers using the combined HT_EDA_LIWC Feature Set. (b) ROC Curves of SVM, GB, RF, and DT Classifiers using the combined HT_EDA_LIWC Feature Set.

4. Results and Discussion

The experimental results demonstrate the effectiveness of the proposed MSDSI-FND framework in detecting fake news propagated by influential users. As illustrated in the ROC curves and summarized in the performance metrics, all classifiers achieved high accuracy and strong discriminative ability, confirming the robustness of the multi-stage feature design. Among the tested models, Neural Network, Gradient Boosting, and Random Forest yielded the highest overall performance, achieving AUC values above 0.98 and F1-scores exceeding 93%. This indicates integrating heterogeneous features: content-based linguistic indicators, user-behavioral metrics, and hierarchical propagation structures substantially improve the model's capacity to identify misinformation signals that are otherwise missed by text-only classifiers. Furthermore, consistent performance across multiple classifiers highlights the generalizability of the feature representations rather than dependence on a specific learning algorithm. The HT_EDA_LIWC configuration, representing the complete multi-stage feature integration, consistently outperformed all partial combinations, reinforcing the effectiveness of the proposed framework. Building on these findings, the next subsection presents an ablation study designed to quantify the contribution of each analytical stage and empirically validate the necessity of the multi-stage design in enhancing fake-news detection performance.

Ablation Study on Multi-Stage Feature Contributions

To evaluate the contribution of each analytical stage within the MSDSI-FND framework, an ablation study was conducted by systematically removing one or more feature groups from the complete feature combination (HT_EDA_LIWC). This analysis aims to quantify the individual impact of each stage content, user, and structural on overall model performance and to empirically validate the rationale for a multi-stage design. In the first experiment, classifiers were trained using only single-stage features such as EDA (content-based textual cues), LIWC (psycholinguistic and linguistic indicators), or HT (hierarchical propagation and user-network features). In subsequent experiments, these feature sets were incrementally combined: first pairing HT + EDA, then integrating LIWC to form the complete HT_EDA_LIWC configuration. This progressive integration reflects the hierarchical logic of the MSDSI-FND framework, where each stage contributes complementary information extracted at different analytical levels. Table 9 compares the performance of the selected classifiers when trained separately on EDA-based and LIWC-based content features. The EDA features capture surface-level textual and stylistic attributes (e.g., sentiment ratios, lexical diversity, readability), while the LIWC features reflect deeper psycholinguistic and emotional indicators. The results highlight how each feature set contributes to detecting fake news through linguistic patterns: EDA features emphasize structural and stylistic complexity, whereas LIWC features provide richer insight into psychological and affective cues that distinguish deceptive from authentic content.

We observed that models trained with LIWC features consistently outperformed those trained with EDA features across all metrics. This indicates that psycholinguistic cues captured by LIWC provide far greater discriminative power for detecting fake news than simple textual or engagement statistics alone. The average accuracy across LIWC models $\approx$ is 86%, compared to only 52% for EDA models. For most classifiers, the AUC values for LIWC models were over 90%, indicating that they performed very well in classification. The DT, GB, and RF models all performed best, with accuracy over 90% and AUCs over 95%. These results demonstrate that LIWC features substantially enhance model performance across all classification algorithms. The high AUC values indicate that the models not only classify correctly but also maintain a clear distinction between fake and real news tweets across all thresholds.

Table 9. Comparative Performance of Classifiers Using EDA and LIWC (Stage 1—Content Analysis).

No.	Algorithm	Features	Metrics		
			F1-Score	Accuracy	AUC
1	LR	EDA features	40.7	57.9	58.4
		LIWC	77.7	78.4	91.4
2	SVM	EDA features	36.5	57.5	52.1
		LIWC	78.9	79.4	91.7
3	DT	EDA features	24.9	28.5	77.4
		LIWC	90.3	90.6	90.2
4	RF	EDA features	36.5	57.5	57.4
		LIWC	86	86.7	95.9
5	NB	EDA features	35.7	45.1	57.6
		LIWC	77	79.2	82.8
6	GB	EDA features	48.1	56.8	57.4
		LIWC	88	88.3	96.3
7	NN	EDA features	38.1	57.7	58.1
		LIWC	78.9	79.4	91.7

The achieved accuracy for LIWC based features was higher than for EDA features, which provides insight into how fake news texts had lower authenticity, greater emotional polarity, and lower cognitive complexity. Real news, on the other hand, had a more balanced linguistic and cognitive profile. Therefore, it is necessary to include linguistic and psychological features to obtain reliable, understandable results when detecting fake news. The model also shows how influential users strategically craft messages to make people more involved by examining linguistic and psychological patterns. They often use emotionally charged or assertive language to spread fake news. Models trained on basic statistical or engagement-driven EDA features showed lower predictive performance than those trained on LIWC features. This apparent performance gap shows that information, whether linguistic or psychological, is better at distinguishing events than basic metrics alone. These findings become even more important when viewed through the lens of influential users, who play a key role in creating and spreading fake news. Influencers often change how they tell stories to make them more relatable and emotional because they have many followers and people trust them.

LIWC's psycholinguistic patterns show that tweets that spread false news are usually less authentic, less logical, and more selective in how they express emotion. People often use this kind of language to get people's attention, make them feel something, or split their opinions. Real news tweets, on the other hand, show more analytical and cognitive processing and use a balanced tone. This aligns with factual reporting and evidence-based communication, showing that fake and real news tweets occupy distinct linguistic clusters in terms of psycholinguistic features. This difference suggests that language patterns, without prevalence metrics such as retweets or follower counts, can signal early signs of false information. Fake news detection systems could therefore use these linguistic cues to identify potentially false or misleading content before it goes viral. In the second phase, the models were trained on micro-level social context features, including structural metrics (e.g., tree depth, number of replies) and user profile attributes (e.g., followers, favorites, statuses) as shown in Table 10.

Table 10. Performance of Classifiers Using HT: (Stage 2—User profiles and propagation features).

No.	Algorithm	Features	Metrics		
			F1 Score	Accuracy	AUC
1	LR	Original DS + HT	62	69.8	81.4
2	SVM	Original DS + HT	62	69.8	81.4
3	DT	Original DS + HT	62	69.8	81.4
4	RF	Original DS + HT	68.3	69.1	81.4
5	NB	Original DS + HT	57.3	59.8	80.3
6	GB	Original DS + HT	62.0	69.8	81.4
7	NN	Original DS + HT	62.0	69.8	81.4

Even without linguistic cues, these models identified important patterns in user behavior and network structure. Fake news cascades had trees that were wide but not very deep, a high maximum out-degree (meaning that only a few influential users spread the news) and intense emotional reactions in the first-level replies. Real news cascades, on the other hand, had more in-depth discussions and greater participant engagement.

Furthermore, the low AUC values (~0.8) showed that social-contextual features are helpful but not enough on their own. This is because different influencers can use the exact diffusion mechanisms to spread both real and fake content. The last step is comparing the performance for the feature sets, the (HT + EDA) and (HT + EDA + LIWC) categories, as shown in Table 11.

Table 11. Comparative Performance of Classifiers Using (Stages 1 and 2).

No.	Algorithm	Features	Metrics		
			F1 Score	Accuracy	AUC
1	LR	HT_EDA	62.9	69.9	81
		HT_EDA_LIWC	87.3	87.7	96.7
2	SVM	HT_EDA	61.9	69.7	80.9
		HT_EDA_LIWC	87.7	88	96.9
3	DT	HT_EDA	48.3	50.9	54.6
		HT_EDA_LIWC	62	50.9	96.9
4	RF	HT_EDA	68.4	69.2	81.5
		HT_EDA_LIWC	93.1	93.2	98.5
5	NB	HT_EDA	43.4	49.7	79.5
		HT_EDA_LIWC	55.4	58.3	90.8
6	GB	HT_EDA	62	69.8	81.4
		HT_EDA_LIWC	95.2	95.3	98.8
7	NN	HT_EDA	64.9	67.7	79.3
		HT_EDA_LIWC	96.6	96.7	99.4

This combination improved the performance of all the algorithms. The NN and GB models performed almost flawlessly in classification (AUCs of 99.4% and 98.8%, respectively). The average F1-score was over 87%, the accuracy was over 90%, and the AUC values were between 96% and 99%. Combining content and context helped the models learn about different kinds of deception. For example, they learned not only what linguistic indications identify fake news, but also how that kind of news acts socially after it is published. The two dimensions worked together to help the models identify manipulative tone or exaggerated

certainty in text (content level) and influence-driven amplification patterns and emotionally driven followers' reactions (context level). The combination of these features yields a strong, understandable model that can detect fake news. By combining LIWC's psycholinguistic insights with HT_EDA's contextual metrics, the system achieves near-optimal detection performance while maintaining interpretability. These findings contribute to the broader understanding that misinformation is a socio-linguistic phenomenon its identification requires capturing both the language of deception and the dynamics of influence that sustain its spread. As conclusion, Table 12 shows the comparative performance of MSDSI-FND Model across different feature sets.

Table 12. Comparative Performance of MSDSI-FND Model across different feature sets.

Feature-Set	Feature Type	Avg. Accuracy	Avg. AUC	Key Contribution
I	LIWC_EDA (content only)	~86%	~95%	Captures linguistic deception and psychological tone
II	Social context (HT)	~69%	~81%	Models' user influence, engagement, and propagation dynamics
III	HT_EDA_LIWC (combined)	>90%	>96%	Learns integrated linguistic + behavioral patterns of misinformation

5. Conclusions

In the context of OSNs, not all tweets have the same impact level. Tweets posted by influencers tend to spread widely and quickly, generating significant audience interactions that often extend beyond their immediate followers. Given that influencers are regarded as credible sources by numerous followers, even a single incident of misinformation can quickly mislead a substantial portion of the audience, making it challenging to clarify the issue because of its rapid dissemination and widespread impact. It may exhibit greater complexity and diversity, encompassing a combination of robust support, correction, aggression, and the dissemination of tweets by other influencers. All of these provide analytical indications that must not be overlooked. Therefore, studying the audience's interaction with them provides an accurate picture of the risks or benefits of each tweet issued by an influencer. This study demonstrated that combining text analysis and context analysis enables the model to comprehensively capture the structural patterns of news dissemination, along with a precise understanding of the nature of the social interaction surrounding a tweet in terms of sentiment, corrections, and polarization. The role of this integrated analysis is particularly prominent in the case of influential users' tweets given their extraordinary power and reach. Fake news issued by influencers can spread more quickly and reach larger audiences, making early detection and remediation crucial. Furthermore, the nature of interactions around an influencer's tweet is often more complex and diverse, giving micro-level features a pivotal role in enhancing the model's accuracy. The model presented in this paper not only focuses on text analysis but also monitors the entire social context surrounding an influencer's tweet, making it more robust and realistic than traditional models. It is suitable for immediate use in content moderation systems or even for developing tools to assist fact-checking teams on major platforms. While this study offers important findings on detecting fake news shared by influential users, it is not without limitations. The dataset mainly focuses on COVID-19 discussions during a specific period, which means the results may not fully apply to other topics or time frames. In addition, the work relies on text and structural data from Twitter only, leaving out images

and videos that often shape how misinformation spreads online. The analysis also does not consider users' long-term behavior or their activity across different platforms, both of which could provide deeper insights into how fake news evolves. Lastly, because the study depends on publicly available and manually labeled data, there is always a chance of bias in the labeling or sampling process. Future studies could address these points by including multimodal content, extending the time span of analysis, and exploring behavioral and reputational aspects to build a more complete picture of how misinformation spreads on social media.

Accordingly, this study recommends the importance of developing artificial intelligence models that integrate structural and micro-features, with a priority given to analyzing influencer posts, to effectively contribute to reducing the spread of fake news and protecting the digital community's awareness from misinformation. This research underscores the importance of continually refining the fake news detection model to keep pace with rapid changes in the social media environment and evolving digital disinformation methods. To achieve this, future research will investigate expanding the scope of analysis to include image and video content and study the evolution of diffusion patterns over time using dynamic network analysis techniques, along with leveraging advanced deep learning algorithms. The accuracy of the model can also be enhanced by analyzing users' digital reputations, incorporating geographical and temporal characteristics, and utilizing external signals from news-verification platforms. It is also interesting to study the reliability of link sources and track changes in the behavior of influencers themselves as well as to develop visual analysis tools that support the interpretation of the results. Implementing such suggestions will ensure the development of a more comprehensive, flexible, and accurate model to address the challenges posed by fake news on OSNs.

Author Contributions: Conceptualization, H.A.-M.; supervision, J.B.; data collection and analysis, H.A.-M.; methodology, H.A.-M.; formal analysis, H.A.-M.; original draft writing H.A.-M. review and editing, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, King Saud University, through the initiative of DSR Graduate Students Research Support (GSR).

Data Availability Statement: The dataset used in this research contains data obtained through Twitter's API and cannot be publicly shared in full due to platform restrictions. However, the processed features and analytical scripts are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, L.; Morstatter, F.; Carley, K.M.; Liu, H. Misinformation in social media: Definition, Manipulation, and Detection. *SIGKDD Explor. Newsl.* **2019**, *21*, 80–90. [CrossRef]
2. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 15. [CrossRef]
3. Vosoughi, S.; Roy, D.; Aral, S. The Spread of True and False News Online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
4. Twitter Permanent Suspension of @realDonaldTrump. Available online: https://blog.x.com/en_us/topics/company/2020/suspension (accessed on 27 May 2021).
5. Gupta, A.; Kumaraguru, P.; Castillo, C.; Meier, P. TweetCred: Real-Time Credibility Assessment of Content on Twitter. In *Social Informatics*; Aiello, L.M., McFarland, D., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Germany, 2014; Volume 8851, pp. 228–243. ISBN 978-3-319-13733-9.
6. Here's How We're Using AI to Help Detect Misinformation. 2020. Available online: <https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/> (accessed on 22 May 2021).
7. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 109. [CrossRef]
8. Bondielli, A.; Marcelloni, F. A Survey on Fake News and Rumour Detection Techniques. *Inf. Sci.* **2019**, *497*, 38–55. [CrossRef]

9. Rubin, V.L.; Chen, Y.; Conroy, N.K. Deception Detection for News: Three Types of Fakes. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–4. [CrossRef]
10. Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; Procter, R. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
11. Kshetri, N.; Voas, J. The Economics of ‘Fake News’. *IT Prof.* **2017**, *19*, 8–12. [CrossRef]
12. Shu, K.; Wang, S.; Lee, D.; Liu, H. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*; Shu, K., Wang, S., Lee, D., Liu, H., Eds.; Lecture Notes in Social Networks; Springer International Publishing: Cham, Germany, 2020; ISBN 978-3-030-42698-9.
13. Google Trends. “Fake News”. Available online: <https://trends.google.com/trends/explore?date=2013-12-06%202021-03-06&geo=US&q=fake%20news> (accessed on 10 October 2025).
14. Shu, K.; Mahudeswaran, D.; Wang, S.; Liu, H. Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. *Proc. Int. AAAI Conf. Web Soc. Media* **2020**, *14*, 626–637. [CrossRef]
15. Sharma, K.; Qian, F.; Jiang, H.; Ruchansky, N.; Zhang, M.; Liu, Y. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–42. [CrossRef]
16. Delirrad, M.; Mohammadi, A.B. New Methanol Poisoning Outbreaks in Iran Following COVID-19 Pandemic. *Alcohol Alcohol.* **2020**, *55*, 347–348. [CrossRef]
17. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The Rise of Social Bots. *Commun. ACM* **2016**, *59*, 96–104. [CrossRef]
18. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.; Flammini, A.; Menczer, F. The Spread of Low-Credibility Content by Social Bots. *Nat. Commun.* **2018**, *9*, 4787. [CrossRef]
19. Jamieson, K.H.; Cappella, J.N. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishmen*; Oxford University Press: Oxford, UK, 2008.
20. Pariser, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*; Penguin: London, UK, 2011.
21. Ciampaglia, G.L.; Shiralkar, P.; Rocha, L.M.; Bollen, J.; Menczer, F.; Flammini, A. Computational Fact Checking from Knowledge Networks. *PLoS ONE* **2015**, *10*, e0128193. [CrossRef]
22. Luhn, H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.* **1957**, *1*, 309–317. [CrossRef]
23. Jones, K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
24. Castel, S.; Almeida, T.; Elghafari, A.; Santos, A.; Pham, K.; Nakamura, E.; Freire, J. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of the 2019 World Wide Web Conference*; ACM Digital Library: New York, NY, USA, 2019; pp. 975–980. [CrossRef]
25. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic Detection of Fake News. *arXiv* **2017**, arXiv:1708.07104. [CrossRef]
26. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. *arXiv* **2017**, arXiv:1702.05638. [CrossRef]
27. Castillo, C.; Mendoza, M.; Poblete, B. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web—WWW ’11*, Hyderabad, India, 28 March–1 April 2011; ACM Press: Hyderabad, India, 2011; p. 675.
28. Abbasi, M.-A.; Liu, H. Measuring User Credibility in Social Media. In *Social Computing, Behavioral-Cultural Modeling and Prediction*; Greenberg, A.M., Kennedy, W.G., Bos, N.D., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7812, pp. 441–448. ISBN 978-3-642-37209-4.
29. Zhang, X.; Ghorbani, A.A. An Overview of Online Fake News: Characterization, Detection, and Discussion. *Inf. Process. Manag.* **2020**, *57*, 102025. [CrossRef]
30. Viviani, M.; Pasi, G. Credibility in Social Media: Opinions, News, and Health Information—a Survey: Credibility in Social Media. *WIREs Data Min. Knowl. Discov.* **2017**, *7*, e1209. [CrossRef]
31. Liu, Y.; Xu, S. Detecting Rumors Through Modeling Information Propagation Networks in a Social Media Environment. *IEEE Trans. Comput. Soc. Syst.* **2016**, *3*, 46–62. [CrossRef]
32. Islam, M.R.; Liu, S.; Wang, X.; Xu, G. Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives. *Soc. Netw. Anal. Min.* **2020**, *10*, 82. [CrossRef] [PubMed]
33. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.F.; Cha, M. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, New York, NY, USA, 9–15 July 2016; AAAI Press: New York, NY, USA, 2016; p. 9.
34. Al-Sarem, M.; Boulila, W.; Al-Harby, M.; Qadir, J.; Alsaeedi, A. Deep Learning-Based Rumor Detection on Microblogging Platforms: A Systematic Review. *IEEE Access* **2019**, *7*, 152788–152812. [CrossRef]
35. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *arXiv* **2018**, arXiv:1708.02709. [CrossRef]

36. Hassan, N.; Gomaa, W.; Khoriba, G.; Haggag, M. Credibility Detection in Twitter Using Word N-Gram Analysis and Supervised Machine Learning Techniques. *Int. J. Intell. Eng. Syst.* **2020**, *13*, 291–300. [[CrossRef](#)]
37. Gravanis, G.; Vakali, A.; Diamantaras, K.; Karadais, P. Behind the Cues: A Benchmarking Study for Fake News Detection. *Expert Syst. Appl.* **2019**, *128*, 201–213. [[CrossRef](#)]
38. Alghamdi, J.; Lin, Y.; Luo, S. Towards COVID-19 Fake News Detection Using Transformer-Based Models. *Knowl.-Based Syst.* **2023**, *274*, 110642. [[CrossRef](#)]
39. Padalko, H.; Chomko, V.; Chumachenko, D. A Novel Approach to Fake News Classification Using LSTM-Based Deep Learning Models. *Front. Big Data* **2024**, *6*, 1320800. [[CrossRef](#)]
40. Hamed, S.K.; Ab Aziz, M.J.; Yaakub, M.R. Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. *Sensors* **2023**, *23*, 1748. [[CrossRef](#)]
41. Hashmi, E.; Yayilgan, S.Y.; Yamin, M.M.; Ali, S.; Abomhara, M. Abomhara Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI. *IEEE Access* **2024**, *12*, 44462–44480. [[CrossRef](#)]
42. Choudhry, A.; Khatri, I.; Jain, M.; Vishwakarma, D.K. An Emotion-Aware Multitask Approach to Fake News and Rumor Detection Using Transfer Learning. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 588–599. [[CrossRef](#)]
43. Alghamdi, J.; Lin, Y.; Luo, S. Does Context Matter? Effective Deep Learning Approaches to Curb Fake News Dissemination on Social Media. *Appl. Sci.* **2023**, *13*, 3345. [[CrossRef](#)]
44. Raza, S.; Ding, C. Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach. *Int. J. Data Sci. Anal.* **2022**, *13*, 335–362. [[CrossRef](#)] [[PubMed](#)]
45. Wang, Y.; Zhang, Y.; Li, X.; Yu, X. COVID-19 Fake News Detection Using Bidirectional Encoder Representations from Transformers Based Models. *arXiv* **2021**, arXiv:2109.14816. [[CrossRef](#)]
46. Azizah, S.F.N.; Cahyono, H.D.; Sihwi, S.W.; Widiarto, W. Performance Analysis of Transformer Based Models (BERT, ALBERT, and RoBERTa) in Fake News Detection. In Proceedings of the 2023 6th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 10 November 2023; IEEE: New York, NY, USA, 2024; pp. 425–430.
47. Raza, S.; Paulen-Patterson, D.; Ding, C. Fake News Detection: Comparative Evaluation of BERT-like Models and Large Language Models with Generative AI-Annotated Data. *Knowl. Inf. Syst.* **2025**, *67*, 3267–3292. [[CrossRef](#)]
48. Shu, K.; Zhou, X.; Wang, S.; Zafarani, R.; Liu, H. 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, Vancouver, BC, Canada, 27–30 August 2019; ACM: Vancouver, BC, Canada, 2019; pp. 436–439.
49. Sahoo, K.; Samal, A.K.; Pramanik, J.; Pani, S.K. Pani. Exploratory Data Analysis Using Python. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 4727–4735. [[CrossRef](#)]
50. Tausczik, Y.R.; Pennebaker, J.W. Pennebaker The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* **2009**, *29*, 24–54. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.