How early can we detect? Detecting Misinformation on Social Media Using User Profiling and Network Characteristics

Shreya Ghosh[1][0000-0002-6970-8889] ✉ and Prasenjit Mitra[1,2]

[1] College of IST, Pennsylvania State University, USA
[2] L3S Research Center, Leibniz University, Hannover, Germany
{shreya,pmitra}@psu.edu

Abstract. The rise of social media has amplified the need for automated detection of misinformation. Current methods face limitations in early detection because crucial information that they rely on is unavailable during the initial phases of information dissemination. This paper presents an innovative model for the early detection of misinformation on social media through the classification of information propagation paths and using linguistic patterns. We have developed and incorporated a causal user attribute inference model to label users as potential misinformation propagators or believers. Our model is designed for early detection of false information and includes two auxiliary tasks: predicting the extent of misinformation dissemination and clustering similar nodes (or users) based on their attributes. We demonstrate that our proposed model can identify fake news on real-world datasets with 86.5% accuracy within 30 minutes of its initial distribution and before it reaches 50 retweets, outperforming existing state-of-the-art benchmarks.

Keywords: Misinformation · social network · discourse analysis.

1 Introduction

Misinformation on social media platforms poses significant challenges to society, as it can influence public opinion [1], exacerbate polarization [18], and even jeopardize public health [20,5]. To mitigate the effect of misinformation, it is crucial to identify false information as early as possible, followed by the implementation of targeted and efficient countermeasures [23]. Detecting false information on social media is inherently challenging due to several factors, especially when focusing on the early detection of false information. Firstly, fake news is deliberately crafted to deceive readers, making it difficult to identify solely based on its content. Secondly, social media data is vast, multi-modal, predominantly user-generated, occasionally anonymous, and often noisy, which complicates the detection process. Thirdly, social media platforms facilitate the cheap and rapid dissemination of news, causing information, whether true or false, to spread quickly and extensively through complex networks. This rapid propagation adds to the challenge of identifying and containing fake news at an early stage. Current methods primarily focus on linguistic patterns [26,7] and external knowledge bases [15] to identify misinformation, which is insufficient for capturing the complex interactions and user behaviours that drive its spread. Furthermore, existing approaches are often limited by their reliance on single-task learning, which can hinder generalizability and robustness across different domains and stages of misinformation dissemination.

To address these limitations, our proposed framework (**CIPHER**[3]) combines advanced NLP techniques to understand the linguistic differences between false and true information, a graph convolutional network for capturing user interactions and propagation characteristics, and attention mechanisms to capture both linguistic patterns and network features that characterize the spread of misinformation. By employing a multi-task learning approach, our model simultaneously tackles the primary task of misinformation detection and auxiliary tasks of predicting propagation depth and clustering users based on their reactions to false information. This integrated approach enables a more comprehensive understanding of the misinformation spread on social media platforms, facilitating early detection and effective countermeasures.Our contributions can be summarized as follows:

- We propose a novel multi-task learning framework that captures both linguistic patterns and network features to effectively detect misinformation, predict its propagation depth, and cluster users based on their reactions to false information.
- We introduce a user causal inference model to identify user's contribution to false information propagation or prevention, and a dynamic attention mechanism that weighs the importance of tokens in the text according to their significance in the misinformation dissemination process, providing a more refined understanding of the linguistic patterns involved in the spread of misinformation.
- We demonstrate the efficacy and robustness of our framework through extensive experiments on multiple datasets (AntiVax, FakeNewsNet, Pheme, and Constraint), comparing its performance against baseline models and showcasing its generalizability across different domains and stages of misinformation spread.

The specific research questions, we are interested to investigate are: *RQ: How can we integrate the user causal model and retweet, follow, and mention network features into a unified framework for the early detection of misinformation on social media platforms? Additionally, how does multi-task learning help in improving the efficacy of early misinformation detection?* (See section 3 for problem definition of multi-task objectives)

CIPHER advances the state-of-the-art in misinformation detection by proposing a novel multi-task learning framework that not only enables early detection of

[3] CIPHER: **C**atching **I**nternet **P**ropaganda: A **H**olistic **E**arly False Information **R**ecognition

misinformation, but also provides valuable insights into the propagation process and user behaviors, laying the foundation for more effective interventions and countermeasures.

2 Related Works

Several studies have explored the use of linguistic features, such as n-grams, syntactic and semantic patterns, and sentiment analysis, in conjunction with machine learning algorithms for misinformation detection [17,25]. These approaches often rely on feature extraction and selection techniques to identify discriminative patterns in text data. However, they struggle to capture the complex interactions between users and the dynamic nature of misinformation spread on social media platforms. With the advent of deep learning, various NLP models, including recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformers, have been employed for misinformation detection [14,10]. These models can learn high-level semantic representations from large-scale text data, improving the detection performance. However, they often overlook the importance of network features and user behaviors, which are crucial for understanding the dissemination process of misinformation. Ruchansky et al. [14] proposed a hybrid model named CSI (Capture, Score, and Integrate) for detecting fake news on social media. This study shows that combining textual, publisher, and user interaction features can effectively detect fake news. Ezeakunne et al. [4] focused on detecting misinformation by analyzing user behavior on social media. They proposed a deep learning model that leverages user behavior patterns, including retweeting, liking, and replying to tweets, to predict the credibility of the information. Ma, et al., [10] developed an attention-based recurrent neural network (RNN) model to detect fake news on social media. They demonstrated that the attention-based RNN model outperformed other state-of-the-art methods in early misinformation detection. However, these methods have limitations. The CSI model's [14] performance is limited by the quality and availability of publisher credibility data and user engagement features, while the BEHIND model [4] relies on user behavior patterns, which may be susceptible to manipulation by malicious actors or bots. The attention-based RNN model [10] struggles to detect misinformation when textual features alone are insufficient or ambiguous. Yang, et al., [22] employed linguistic cues and user features for early detection of rumors, but the approach may be limited by language-specific characteristics and evolving user behaviors. Volkova et al. [19] focused on identifying truthful versus deceptive news headlines using linguistic analysis, but their model struggles with misleading headlines that are factually accurate. Rashkin, et al., [13] focused on identifying truthful versus deceptive news headlines using linguistic analysis, but the model might struggle with misleading headlines that are factually accurate. Monti, et al., [11] proposed a geometric deep learning approach to detect misinformation, but the model's performance may be limited by the structural complexity and scale of real-world social networks. Liu, et al., [8] proposed a novel deep neural network combining crowd response features and user reactions to effectively detect misinformation early. Liang, et al., [21] proposed a model incorporating stance information from users to improve fake news detection, but the model's performance may be limited by the availability and quality of user-generated stance information.

3 Problem Definition

Given a social network S represented by a graph $G(V, E)$, where V is the set of nodes (users) and E is the set of edges (connections), and a set of microblog posts T, let $t \in T$ have a timestamp t_t and be classified as either misinformation (M) or fact/true/genuine information (F). CIPHER addresses the following problems: **Early False Information Identification (EFII)**: For each $t \in T$, determine a classification task $C(t) \in \{M, F\}$, minimizing the time taken for classification, t_{cl}, such that $t_{cl} \leq \theta$, where θ is the maximum allowable time for early detection. The performance is measured by F1 score within the range $[\varphi, 1]$ (i.e., how much minimum time is required to reach at least φ F1]. **User Classification (UC)**: For each user $u \in V$, assign a label $L(u) \in$ {M_spreader, M_preventer, M_initiator, M_skeptic} based on their role in spreading or mitigating misinformation. **Predicting Depth of False Information Reach (PDFIR):** For each misinformation tweet $m \in M$, estimate the depth of reach $d(m, t_p)$ within the social network $G(V, E)$ at a future time point t_p, where $t_p = t_t + \mathit{\Delta} t$, and $\mathit{\Delta} t$ is the prediction horizon. Next, we cluster users in V into 4 groups based on their UC labels and characteristics. The multi-task learning problem addresses the objectives (EFII, UC, PDFIR) considering the linguistic pattern of the posts, temporal dynamics of social networks and information propagation.

4 CIPHER: Methodology

4.1 Network Construction

In this section, we introduce a three-layered graph that combines the Retweet, Mention, and Follow networks to better understand user interactions and information dissemination patterns on social media platforms. This integrated graph aims to improve early misinformation detection by leveraging the diverse interactions across the three networks.

Let $G = (V, E)$ denote a directed, weighted multilayer network, where V is the set of nodes (users) and E is the set of edges (user interactions) across the three layers. The three-layered graph is represented by $G_3L = (V_3L, E_3L)$, where (i) V_3L: the union of V_{RT}, V_M, and V_F, representing the set of users involved in retweets, mentions, or follows. (ii) E_3L: the set of edges (u, v, k) across the three layers, with $k \in \{RT, M, F\}$, where (u, v, RT) denotes a retweet interaction, (u, v, M) represents a mention interaction, and (u, v, F) signifies a follow interaction. (iii) $w(u, v, k)$: the weight of edge (u, v, k), indicating the interaction strength between users u and v in layer k. The three-layered graph captures diverse interaction types, providing a more accurate and holistic understanding of user behavior and information flow on social media platforms. We show that by combining the Retweet, Mention, and Follow networks, the integrated graph can help identify potential misinformation spreaders and influential users across different types of interactions, enhancing early detection capabilities.

Next, to incorporate the credibility of users in the three-layered graph, we propose assigning edge weights based on a personalized PageRank trust model [2]. The intuition behind this approach is that if *a user B primarily shares unreliable content, user A, who interacts with user B, is also likely to share unreliable content.* This trust model helps us capture the importance of channels through which misinformation or true information spreads. Here, the personalized PageRank trust model computes a trust score for each user based on their credibility. Let $T(u)$ denote the trust score of user u. To compute T(u), we use a personalized PageRank algorithm [9] with a preference vector that prioritizes users who are known to be credible, as determined by fact-checking organizations or other reliable sources. Using the trust scores $T(u)$ and $T(v)$ for users u and v, we update the edge weights in the three-layered graph as follows: $w(u, v, RT) = T(u) * T(v) * N_{RT}(u, v)$, where $N_{RT}(u, v)$ denotes the number of times user u retweets content from user v. $w(u, v, M) = T(u) * T(v) * N_M(u, v)$, where $N_M(u, v)$ represents the number of times user u mentions user v. $w(u, v, F) = T(u) * T(v)$, as user u either follows or does not follow user v, and we consider the trust scores of both users to assign the weight. This weight assignment strategy accounts for the credibility of both users involved in the interaction, making the graph more informative for early misinformation detection.

Transmitter and Receiver Characteristics in Misinformation Propagation Next, we analyze the roles and characteristics of transmitters and receivers in the context of misinformation propagation on social media platforms. *Transmitter:* Transmitters are individuals who propagate information on social media. We select the following characteristics that can influence the spread of misinformation by transmitters: (1) *Reaction time:* The speed at which a user forwards or shares received information upon encountering it. (2) *Perseverance:* Persistence in spreading information despite difficulties or delays in convincing others. Users may spread information at different time scales, ranging from single forwards to long-term efforts (super-spreaders). (3) *Authority level:* The number of followers or the user's relevance to a specific domain, such as healthcare, can impact their influence on misinformation spread. (4) *Sensitivity:* Users may exhibit different levels of sensitivity when encountering misinformation, including (i) believe-and-forward, (ii) being neutral, or (iii) not believing and persuading others to act the same.

Receiver: Receivers are individuals who consume and potentially propagate information further. CIPHER considers the factors influence the likelihood of receivers spreading misinformation: (1) *Attitude:* Receivers may immediately change their state (e.g., adopt a belief), require some time before being convinced (e.g., by seeking additional information), or be completely insensitive to the information (i.e., no change of state). (2) $Number of Messages$: The

Dataset	Post
AntiVax	(Transmitter)“AntiVaxWarrior” posts a tweet claiming, “The MMR vaccine causes autism! Don’t let them poison your kids! #StopTheVax #Autism” (Receiver) “ConcernedParent” who is unsure about vaccines retweets ”@AntiVaxWarrior” ’s tweet and adds, “Is this true? I am worried about vaccinating my child now”. #VaccineSafety”
FakeNewsNet	(Transmitter) “@FinanceGuru” shares an article, “Cryptocurrency scams are on the rise. Make sure to research and verify before investing. #Crypto #InvestWisely” Resistance and Misinformation: (Receiver 1) “@CryptoKing” retweets and adds, “This is just fear-mongering by mainstream media! Cryptos are the future!” #Cryptocurrency #FinancialFreedom” (Receiver 2) “@MoneyMatters” retweets and comments, “The banks are spreading lies to protect their outdated financial system! #CryptoRevolution #BankingLies”

Table 1: Illustration of receiver and transmitter

likelihood of misinformation propagation can be influenced by the frequency of messages from the same sender or multiple, different users. (3) $Sourceauthority$: The popularity (e.g., number of followers) or recognized expertise of the transmitter can be a critical factor in the propagation success of misinformation. Receivers may be more likely to trust and spread information from authoritative sources. By considering the characteristics of transmitters and receivers in the context of misinformation propagation, we develop more effective early detection algorithms and strategies for mitigating the spread of misinformation on social media platforms. We provide two examples in the context of misinformation propagation in Table 1. In the example, the transmitter is responsible for disseminating misinformation, while the receiver, depending on their attitude and susceptibility, may contribute to the further spread of the information.

4.2 User Causal Model

We propose a *Causal User Attribute Inference (CUAI)* model that employs a Graph Attention Network (GAT) to infer the causal relationships between user attributes and their propensity to spread misinformation. Let $G = (V, E)$ represent the social network graph. Each user $v \in V$ has an associated attribute vector $A(v)$, consisting of features such as reaction time, perseverance, authority level, sensitivity, and source authority. The CUAI model consists of the following components:

1. Attribute Embedding Layer: We use a linear transformation layer to convert user attributes, A(v), into continuous feature vectors $h_0(v) \in R^d$, where d is the embedding dimension:

$$h_0(v) = W_0 * A(v) + b_0, \tag{1}$$

where W_0 and b_0 are the learnable weight matrix and bias vector, respectively.
2. GAT-based Causal Inference: We employ a multi-head Graph Attention

Network (GAT) model to learn the causal relationships between user attributes and their propensity to spread misinformation:

$$h_l + 1(v) = ||_{k=1}^{K} Attention_k(h_l(v), h_l(u) : u \in N(v)), \tag{2}$$

where || denotes concatenation, K is the number of attention heads, N(v) represents the neighbors of user v, and $Attention_k()$ is the k-th attention mechanism. The attention mechanism computes the importance of neighboring nodes' features based on the input features:

$$\alpha_k(u, v) = softmax_u(LeakyReLU(W_k^T [h_l(u)||h_l(v)])) \tag{3}$$

where W_k is the learnable weight matrix for the k-th attention head, and [||] denotes the concatenation of two vectors.

3. Causal Effect Estimation: We use the learned user embeddings $h_L(v)$, where L is the number of GAT layers, to estimate the causal effects of user attributes on misinformation propagation. By employing GATs for causal inference, the CUAI model provides a powerful and flexible approach to identifying the key causal factors driving the spread of misinformation in social networks. This knowledge is used to design more effective intervention strategies and reduce the impact of false information on society. Next, we use the outcome of the framework to estimate the causal effects of user attributes on misinformation propagation. For each user v, we define two potential outcomes: $Y_v(1)$ and $Y_v(0)$ based on if the user attribute was set to a specific value (1) or not set (0), respectively. The causal effect for user v is then defined as the difference between the two potential outcomes:

$$\tau(v) = E[Y_v(1) - Y_v(0)], \tag{4}$$

where $E[]$ denotes the expectation. To estimate $\tau(v)$, we use the learned user embeddings $h_L(v)$ and fit two separate regressions:

$$Y_v(1) = g_1(h_L(v); \theta_1), Y_v(0) = g_0(h_L(v); \theta_0), \tag{5}$$

where $g_1(;\theta_1)$ and $g_0(;\theta_0)$ are regression functions with parameters θ_1 and θ_0, respectively. We can then estimate the causal effect as the difference between the predicted outcomes:

$$\tau(v) \approx g_1(h_L(v); \theta_1) - g_0(h_L(v); \theta_0). \tag{6}$$

By estimating the causal effects, we identify the most influential user attributes that can be targeted for interventions to reduce misinformation propagation.

4. Misinformation Propensity Prediction: To predict whether a given user v is likely to propagate misinformation, we train a supervised classifier using the learned user embeddings $h_L(v)$ as features. Let $f(;\theta)$ be the classifier function with parameters θ, and let $y(v) \in 0, 1$ denote the ground truth label for user v, where 1 represents a user who propagates misinformation and 0 represents a user who does not. We can define the classification loss as:

$$L(\theta) = \sum_{v \in V} L_{cls}(y(v), f(h_L(v); \theta)), \tag{7}$$

where the cross-entropy loss is denoted by $L_{cls}(,)$. During training, the classifier minimizes the loss function with respect to the parameters $\theta : \theta* = argmin_{\theta} L(\theta)$. Once the classifier is trained, we predict the misinformation propensity for a given user v by computing the probability of the user propagating misinformation:

$$P(y(v)) = (1|h_L(v)) = f(h_L(v)); \theta*). \tag{8}$$

By predicting misinformation propensity using the learned user embeddings, the CUAI model is used to identify users who are more likely to spread misinformation, enabling targeted interventions and mitigating the impact of false information on social networks.

4.3 Temporal Characteristics

Observation (AntiVax Dataset) A tweet claiming that the MMR vaccine is linked to autism initially receives retweets and likes from users who agree with the statement. As the tweet spreads, users who express their surprise at such a claim, asking for evidence or research supporting the claim. Further down the line, users begin to question the claim's validity and ask for reliable sources, engaging in conversations to debunk the misinformation. To incorporate the observed patterns into a model that considers the dynamic nature of information dissemination, we propose a novel method that utilizes a dynamic attention value for each post. Let P be a post and $t(P)$ be the time when the post was made. Let E be the event corresponding to the initial tweet and $t(E)$ be the time when the event started. We define the time interval for a post as: $\delta t(P) = t(P) - t(E)$. Next, let G be the graph representing the social network, where $V(G)$ is the set of nodes (users) and $E(G)$ is the set of edges (interactions). Let $N(P)$ be the set of nodes (users) that have already interacted with post P. Then, we define the interaction ratio $R(P)$ as: $R(P) = |N(P)|/|V(G)|$. Now, let $F(G)$ be the follower-followee network of the users in G, and $L1(P)$ be the set of nodes reachable from P using a BFS search algorithm. We define the *BFS ratio* L(P) as: $L(P) = |L1(P)|/|V(F(G))|$. The dynamic attention value $A(P)$ for a post P can be calculated as a weighted sum of the time interval, interaction ratio, and BFS ratio:

$$A(P) = \alpha * \delta t(P) + \beta * R(P) + \gamma * L(P) \tag{9}$$

whereα, β, γ are weights that can be tuned based on the importance of each factor in the dissemination process. Finally, let $S(P)$ represent the linguistic pattern score for the post P. The overall score for a post P, considering both dynamic attention and linguistic patterns, can be calculated as:

$$Score(P) = \gamma * A(P) + (1 - \gamma) * S(P) \tag{10}$$

where γ is a weight that balances the influence of dynamic attention and linguistic patterns in the model. By incorporating this dynamic attention value and linguistic pattern into our model, we can more effectively capture the patterns observed in the information dissemination process and improve the early detection of misinformation.

For a given news story propagating on social media, we first construct its propagation path by identifying the users who engaged in propagating the news. **User profiling:** We convert user profiles into fixed-length sequences. Let U_i be the fixed-length sequence representing the user profile of user i. For each user i, we create a propagation path P_i that consists of the user profile sequence U_i and the interactions in which user i participated. In layer 1, we apply a Gated Recurrent Unit (GRU) layer to learn the vector representation V_i for each propagation path P_i: $V_i = GRU(P_i)$. In layer 2, we deploy Graph Convolutional Network (GCN) layer to learn the transformed propagation path T_i for each user profile vector V_i: $T_i = GCN(V_i)$. We concatenate transformed propagation paths by combining the transformed propagation paths T_i into a single vector C that represents the overall transformed propagation path: $C = Concat(T_1, T_2, ..., T_n)$. Finally, we deploy a multi-layer feedforward neural network to predict the maximum depth D for the corresponding propagation path: $D = FNN(C)$. By incorporating user profiling and propagation paths into the model, we can more effectively capture the propagation patterns observed in the information dissemination process and improve the early detection of misinformation. This approach allows us to account for the impact of individual users on the overall propagation of news stories and to better understand the dynamics of misinformation spread. The classifier uses a binary cross-entropy loss $L_{misinfo}$ as the loss for the primary task of detecting false information. L_{depth} is the loss for the auxiliary task of predicting the depth of false information propagation. This is a mean squared error (MSE) loss. $L_{cluster}$ is the loss for the auxiliary task of clustering users based on their reactions to false information. This is a categorical cross-entropy loss, with clusters represented as one-hot encoded vectors. Task-specific weighting factors, γ_1, γ_2 and γ_3, control the relative importance of the tasks in the joint loss function. The combined loss function L_total can be defined as:

$$L_{total} = \gamma_1 * L_{misinfo} + \gamma_2 * L_{depth} + \gamma_3 * L_{cluster} \tag{11}$$

We also introduce adaptive weighting factors that dynamically adjust the importance of each task during training. We use the inverse training progress as a weight factor: $\gamma_i(t) = \alpha_i/(1 + \beta_i * tr)$. Here, tr is the current training step, α_i and β_i are positive hyperparameters for each task i, and γ(t) is the weighting factor for task i at step t. This formulation ensures that the weighting factors decrease as training progresses, allowing the model to focus on the most relevant tasks at each stage of training.

4.4 Linguistic Pattern Analysis

CIPHER deploys the following three components using linguistic pattern analysis: (A) A novel Semantic Similarity Analysis (SSA) approach using a pre-trained RoBERTa model, multi-layer attention mechanism, and contrastive learning to enhance early misinformation detection. We employ the Hugging Face Transformers and spaCy libraries in Python. Firstly, we tokenize the text, convert it to lowercase, and remove special characters, URLs, and user mentions using spaCy and regular expressions. Next, we initialize the pre-trained RoBERTa model (DistilRoBERTa[4]), denoted as R. Next, we use a multi-layer attention mechanism A, consisting of L self-attention layers, each followed by a feed-forward network and layer normalization. Let the output of the last transformer layer in RoBERTa be denoted as H_i for each tweet t_i. The attention mechanism computes a weighted representation $A(H_i) = \sum_{l=1}^{L} W_l F_l(H_i)$, where F_l denotes the l-th self-attention layer followed by the feed-forward network and layer normalization, and W_l are learnable weights. Then, we construct the training dataset of paired examples, with each pair consisting of a tweet t_i and a reference source r_j. For each tweet, generate positive pairs (t_i, r_j^+) with verified information sources sharing semantic similarity and negative pairs (t_i, r_j^-) with unrelated or contrasting sources. We fine-tune the RoBERTa model with the multi-layer attention mechanism on the paired dataset using contrastive learning. The objective is to learn semantic embeddings $\varphi(t_i)$ and $\varphi(r_j)$ that minimize the distance between positive pairs and maximize the distance between negative pairs:

$$\mathcal{L}contrastive = \sum_{i=1}^{N} \left[d(\varphi(t_i), \varphi(r_j^+)) - \alpha + \max_{r_j^-} d(\varphi(t_i), \varphi(r_j^-)) \right]_+, \tag{12}$$

where $d(\cdot, \cdot)$ denotes a distance metric (e.g., cosine distance), α is a margin parameter, $[\cdot]_+$ represents the hinge function, and N is the number of tweets in the dataset.

Argument Mining and Logical Fallacy Detection (AMLF): The module is designed to identify argumentative structures and logical fallacies in textual data. Given a dataset D containing text samples t_i, our objective is to extract argument components, such as claims C_i, premises P_i, and conclusions Q_i. Let F_{ext} denote an extraction function, parameterized by a pre-trained RoBERTa, which is fine-tuned for argument component extraction. The extraction process can be defined as follows: $(C_i, P_i, Q_i) = F_{ext}(t_i)$, where t_i is a text sample from the dataset D. For each extracted argument component, we aim to identify argumentative relations R_{ij} between them, such as support, attack, or neutral. Let F_{rel} denote a relation identification function, parameterized by a pre-trained NLP model fine-tuned on an argument relation dataset. The relation identification process can be defined as: $R_{ij} = F_{rel}(C_i, P_j)$. where C_i and P_j are argument components extracted from the text samples. Next, to detect logical fallacies, we define three fallacy patterns $\mathcal{F} = f_1, f_2, f_3$, such as *ad hominem, straw man, or false cause.* We aim to recognize and classify these patterns in argumentative structures. Some examples of tweets containing logical fallacies from the Anti-Vax dataset are provided in Table 2.

Sentiment analysis: Given a dataset D containing text samples t_i, our objective is to classify the sentiment expressed in each text as positive, negative, or neutral. To enhance early misinformation detection, we extract specific features from sentiment analysis, such as: **Sentiment Polarity Score:** Calculate a sentiment polarity score P_i for each text sample, indicating the degree of positivity or negativity expressed in the text. **Subjectivity Score:** U_i is the level

[4] `https://huggingface.co/distilroberta-base`

Fallacy type	Post
Ad Hominem	“You can’t trust Dr. XXX’s opinion on vaccines; he’s just a puppet for Big Pharma! #VaccineTruth”
Straw Man	“Pro-vaxxers want us to believe that vaccines are 100% safe and have no side effects, but my child developed a fever after getting vaccinated. #VaccineInjury”
False Cause	“I saw a news report that a child was diagnosed with autism just days after being vaccinated. Clearly, vaccines are the cause of autism. #VaccineHarm”

Table 2: Illustration of fallacy patterns (AntiVax)

of personal opinion, emotion, or judgment in the text i. **Stance Confidence Score:** C_i is the model’s certainty in the detected stance towards the target in the text i. Some examples of tweets from the AntiVax dataset, illustrating the differences in sentiment and stance between true and fake information are presented in Table 3. For true information, the sentiment is positive or neutral, and the stance may favor a particular viewpoint or report on factual events. Conversely, false information often exhibits negative sentiment and may adopt a stance against specific topics, entities, or claims.

Type of news	Post
True information (Real news)	“The World Health Organization has declared COVID-19 a global pandemic. Countries worldwide are implementing preventive measures to curb the spread of the virus.” Sentiment: Neutral. Stance: Reporting
False information	“Shocking news! COVID-19 is a hoax created by the government to control the population! They’re using the pandemic to enforce strict surveillance on citizens! #COVIDHoax” Sentiment: Negative. Stance: Against

Table 3: Sentiment difference in Fake vs true news (AntiVax)

5 Experimental Evaluations

[5] **Dataset Description** We evaluated our proposed framework, CIPHER, using four real-life datasets as described in Table 4. In this section, we discuss the experimental analysis and evaluation of CIPHER, focusing on three main evaluation tasks: (1) how early misinformation can be detected with $\geq$ 85% accuracy, (2) predicting the depth of false information propagation, and (3) clustering users based on their characteristics. We highlight interesting findings

[5] The experiment details, codebase and additional information is available HERE

Dataset	Details
PHEME [3]	Rumors and their veracity in social media. It contains approximately 330 rumor threads collected from Twitter, with each thread having an average of 100 tweets. The dataset covers nine different events, including terrorist attacks, shootings, and natural disasters.
AntiVax [6]	Anti-vaccination movement and contains over 1.8 million tweets collected between 2019 and 2021. The dataset includes tweets, retweets, mentions, and replies, along with associated metadata.
CONSTRAINT [12]	Created for the CONSTRAINT AAAI-21 Shared Task on detecting misinformation during the COVID-19 pandemic. It comprises over 17,000 English tweets, annotated as either real or fake, with an equal distribution between the two classes
FakeNewsNet [16]	A comprehensive dataset containing both real and fake news articles. It consists of data from two popular fact-checking websites, PolitiFact and GossipCop over 23,000 news articles, with metadata such as social engagements, user information, and propagation patterns.

Table 4: Four real-life datasets used for CIPHER's performance evaluation

extracted from the PHEME, AntiVax, FakeNewsNet, and Constraint datasets. Table 5 represents the comparison of misinformation detection in terms of F1-score within 24hour, 12hour and 30mins. Apart from PHEME dataset, CIPHER outperforms the baselines[6] for all experimental settings by a significant margin. Table 6 presents CIPHER's performance on predicting infected nodes (users who initiate, transmit and believe misinformation, at time-step t+1) and predicting depth of misinformation reach.

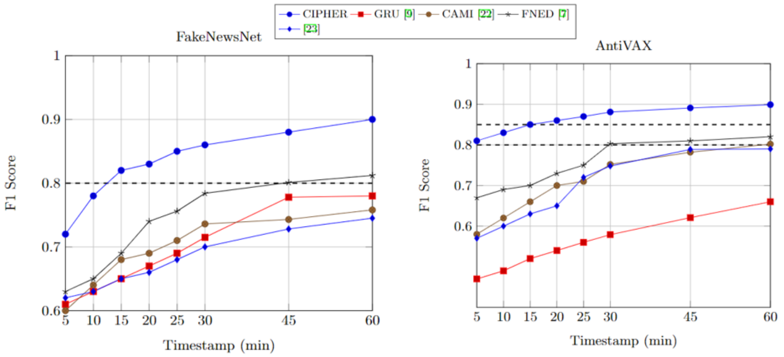

Fig. 1: Comparison of the minimum time required for the identification of misinformation with ≥ 0.80 F1

[6] Baselines have been selected as state-of-the-art models for misinformation detection and early misinformation detection on social media

Dataset / Time	**F1 Score**				
	CIPHER	CAMI [23]	FNED [8]	GRU [10]	[24]
AntiVax / 24h	**0.9408**	0.861	0.892	0.814	0.820
AntiVax / 12h	**0.901**	0.812	0.842	0.683	0.790
AntiVax / 30m	**0.881**	0.752	0.803	0.579	0.748
CONSTRAINT / 24h	**0.931**	0.843	0.866	0.802	0.801
CONSTRAINT / 12h	**0.895**	0.808	0.832	0.661	0.772
CONSTRAINT / 30m	**0.870**	0.736	0.791	0.518	0.721
FakeNewsNet / 24h	**0.92**	0.848	0.840	0.849	0.810
FakeNewsNet / 12h	**0.872**	0.791	0.831	0.790	0.760
FakeNewsNet / 30m	**0.856**	0.736	0.784	0.715	0.691
PHEME / 24h	**0.905**	0.852	0.848	0.867	0.816
PHEME / 12h	**0.856**	0.768	0.819	0.828	0.772
PHEME / 30m	0.811	0.701	0.723	**0.820**	0.607

Table 5: F1 score for detecting misinformation (post): comparison with baseline models and proposed framework (CIPHER). Best score is in bold font.

Timesep	**AntiVax**		**FakeNewsNet**		**PHEME**		**Constraint**	
	Depth	Node	Depth	Node	Depth	Node	Depth	Node
T/3	0.801	0.78	0.76	0.75	0.72	0.703	0.812	0.802
T/2	0.85	0.80	0.798	0.784	0.768	0.718	0.868	0.854
2T/3	0.910	0.845	0.823	0.810	0.827	0.781	0.920	0.907

Table 6: CIPHER's performance (F1) on predicting maximum depth of false information propagation in network and predict infected nodes at time-step t+1

We use a specific example from the AntiVax dataset to illustrate the workings of the proposed multi-task learning framework. Tweet A: "This new vaccine causes severe side effects! It paralyzed my friend's arm! #AntiVax" Retweeted by user B with a comment: "I've heard similar stories. Are vaccines really safe? #QuestioningVaccines" User C, a healthcare professional, replies to user B: "Vaccines are safe and rigorously tested. Side effects are rare and usually mild. Here's a link to the CDC's vaccine safety information. #VaccinesSaveLives". Now, CIPHER's lingusitic model extract patterns from Tweet A, user B's comment, and user C's reply using a pre-trained model RoBERTa. And the attention layer weighs the importance of tokens in each text (e.g., "paralyzed", "side effects", "safe", and "CDC"). The 3-layered graph represents the relationships between users A, B, and C, capturing retweets, mentions, and follow relationships. The GCN layer is deployed to the 3-layered graph, capturing structural features and user interactions in the network. Then the context vector from the attention layer (focusing on crucial tokens) with the output from the GCN layer, creating a unified representation that captures both linguistic patterns and network features are deployed together followed by the multi-task learning (e.g., identifying Tweet A as misinformation) and auxiliary tasks of predicting the depth of false information propagation (e.g., estimating how many layers of users the misinformation will reach) and clustering users based on their reactions to false information (e.g., grouping users A and B as vaccine skeptics and user C as a healthcare professional). It has been observed through ablation study that CIPHER's performance has been improved by 8% and 6% by adding User causal model and temporal characteristics with the linguistic pattern respectively. Alongside, the multi-task learning framework has enhanced the overall accuracy by $6\% - 9\%$ in the misinformation detection task.

Our experiments demonstrate that our proposed approach (CIPHER) can effectively detect misinformation early, achieving over 85% accuracy in most cases. By combining the Semantic Similarity Analysis (SSA), Argument Mining and Logical Fallacy Detection (AMLF), and Sentiment Analysis modules, we were able to identify key linguistic features that contribute to the early detection of misinformation. We observed that the PHEME and AntiVax datasets contained noticeable differences in sentiment and stance between true and false information, as well as the presence of logical fallacies in the latter. In the FakeNewsNet dataset, we found that the false stories often contained sensational language and misleading claims, which could be identified using our SSA and AMLF modules. Similarly, the Constraint dataset exhibited distinct patterns in terms of argument structure and fallacies, which contributed to the early detection of false claims. Fig. 1 shows that CIPHER is able to identify false information within 3-12 minutes with more than 0.80 F1 at the earliest which is better than state-of-the-art models. Our approach also allows for the prediction of the depth of false information propagation, helping to estimate the potential reach and impact of misinformation (See Table 6). By analyzing the content and context of the misinformation, as well as the characteristics of the users involved in its dissemination, we were able to model the spread of misinformation within social media networks.

6 Conclusion

We presented a comprehensive approach to early misinformation detection on social media platforms by leveraging user profiling, linguistic analysis and network analysis modules. Our proposed methodology aims to identify and flag potential misinformation in its early stages of propagation to mitigate its spread and impact on society. CIPHER demonstrated promising results in detecting misinformation, highlighting the importance of incorporating various linguistic features and network characteristics to build an effective detection system.

Acknowledgements

This research was funded by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003.

Ethical consideration

Several ethical considerations were taken into account. Data was obtained from publicly available sources. We took measures to ensure the privacy and anonymity of the individuals whose data was used. We recognize that social media data can be biased in many ways, and we took measures to mitigate these biases. We ensured that our research is conducted in a responsible and ethical manner.

References

1. H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.
2. Y. Asim, A. K. Malik, B. Raza, and A. R. Shahid. A trust model for analysis of trust, influence and their relationship in social network communities. *Telematics and Informatics*, 36:94–116, 2019.
3. L. Derczynski and K. Bontcheva. Pheme: Veracity in digital social networks. In *UMAP workshops*, 2014.
4. U. Ezeakunne, S. M. Ho, and X. Liu. Sentiment and retweet analysis of user response for early fake news detection. In *The International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS'20)*, pages 1–10, 2020.
5. S. Ghosh, P. Mitra, and B. L. Hausman. Evade: Exploring vaccine dissenting discourse on twitter. In *epiDAMIK 5.0: The 5th International workshop on Epidemiology meets Data Mining and Knowledge discovery at KDD 2022*, 2022.
6. K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30, 2022.
7. S. Jiang and C. Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.
8. Y. Liu and Y.-F. B. Wu. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33, 2020.
9. P. A. Lofgren, S. Banerjee, A. Goel, and C. Seshadhri. Fast-ppr: Scaling personalized pagerank estimation for large graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1445, 2014.
10. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.
11. F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.
12. P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 21–29. Springer, 2021.

13. H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.
14. N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.
15. N. Seddari, A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi, and A. Bouras. A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, 10:62097–62109, 2022.
16. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.
17. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
18. J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.
19. S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 647–653, 2017.
20. Y. Wang, M. McKee, A. Torbica, and D. Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552, 2019.
21. L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, pages 637–645, 2018.
22. Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*, 2018.
23. F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, et al. A convolutional approach for misinformation identification. In *IJCAI*, pages 3901–3907, 2017.
24. Z. Yue, H. Zeng, Z. Kou, L. Shang, and D. Wang. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433, 2022.
25. H. Zhang, S. Qian, Q. Fang, and C. Xu. Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Transactions on Multimedia*, 23:4441–4454, 2020.
26. C. Zhou, K. Li, and Y. Lu. Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness. *Information Processing & Management*, 58(6):102679, 2021.