Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Clock against Chaos: Dynamic Assessment and Temporal Intervention in Reducing Misinformation Propagation

Shreya Ghosh[1], **Prasenjit Mitra**[2,3], **Preslav Nakov**[4]
[1] Indian Institute of Technology (IIT) Bhubaneswar, India
[2] College of IST, The Pennsylvania State University, USA
[3] L3S Research Center, Leibniz University Hannover, Germany
[4] Mohamed bin Zayed University of Artifcial Intelligence, UAE
shreya.cst@gmail.com, pmitra@psu.edu, preslav.nakov@mbzuai.ac.ae

Abstract

As social networks become the primary sources of information, the rise of misinformation poses a significant threat to the information ecosystem. Here, we address this challenge by proposing a dynamic system for real-time evaluation and assignment of misinformation scores to tweets, which can support the ongoing efforts to counteract the impact of misinformation public health, public opinion, and society. We use a unique combination of Temporal Graph Network (TGN) and Recurrent Neural Networks (RNNs) to capture both structural and temporal characteristics of misinformation propagation. We further use active learning to refine the understanding of misinformation, and a dual model system to ensure the accurate grading of tweets. Our system also incorporates a temporal embargo strategy based on belief scores, allowing for comprehensive assessment of information over time. We further outline a retraining strategy to keep the model current and robust in the dynamic misinformation landscape. The evaluation results across five social media misinformation datasets show promising accuracy in identifying false information and reducing propagation by a significant margin.

Introduction

The emergence of social networks as a primary source of information has brought with it the rise of misinformation, a challenge that threatens the very fabric of our information ecosystem (Meel and Vishwakarma 2020). Rapid, unchecked propagation of unverified information on these platforms can have far-reaching societal impact, including influencing public opinion (Allcott and Gentzkow 2017; Ghosh and Mitra 2023c), exacerbating polarization (Tucker et al. 2018), and even jeopardising public health (Wang et al. 2019; Ghosh and Mitra 2023b). The unprecedented transition from traditional news media, where the roles of the journalists and the consumers are well-defined, to the democratic landscape of social media, where news is crowd-sourced and anyone can function as a reporter, has posed significant challenges. As a result, there is an increasing trend of traditional news outlets to disseminate unverified information in an attempt to stay ahead of the curve (Fung et al. 2022). A pertinent example of this is the Russia–Ukraine conflict, where numerous reports are hastily published, leading to confusion and misinterpretation of the actual events on the ground (Park et al. 2022; Ghosh and Mitra 2023a).

At the same time, news consumers, i.e., the general public, find it difficult to discern fact from fiction given the speed at which the news spreads, and the varying degrees of reliability, e.g., unverified claims and conspiracy theories circulated widely on social media platforms, leading to confusion and hesitancy about vaccine safety and efficacy.

The real-time identification and mitigation of such misinformation are increasingly pressing research challenges. First and foremost, the deliberate crafting of fake news with the intent to deceive readers inherently complicates its identification based purely on content analysis (Vosoughi, Mohsenvand, and Roy 2017). This issue is further compounded by the characteristics of social media data, which is voluminous, multimodal, predominantly user-generated, occasionally anonymized, and frequently plagued by noise, thereby exacerbating the complexity of the detection process. In addition, the inherent structure of social media platforms enables a low-cost and rapid diffusion of news content. This rapid propagation capability allows for the swift and widespread dissemination of information, regardless of its veracity, through intricate network structures. This augments the difficulty of identifying and halting the proliferation of fake news in its early stages. The urgency and the scale of the task pose considerable challenges to the development of effective countermeasures as well.

Limitations and Challenges

Existing methods (See ‡ Related Work) to detect and to grade misinformation, however, are met with significant challenges and limitations: (I) *Static Evaluation*: Existing models predominantly use static scoring systems, which prove inadequate due to their failure to adapt to the dynamic nature of social networks. This lack of adaptability often results in delayed detection of misinformation. (II) *Temporal Oversights*: Current approaches typically do not account for temporal changes in the information context or the continuous influx of new data, leading to suboptimal detection and grading accuracy. (III) *Immediate Classification:* It may not always be possible or accurate, particularly in the face of uncertain or incomplete information. (IV) *Scalability Concerns*: The enormous scale of the data on social media and the high user activity levels pose significant computational

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

efficiency and scalability challenges to existing methods.

Motivating Examples. In our misinformation grading system, a tweet is assigned a score ranging from 1 to 5. A score of 1 indicates an entirely accurate tweet, 2 is for a largely accurate tweet that contains minor inaccuracies, 3 denotes a tweet with significant inaccuracy, yet still holding some elements of truth, 4 is a mostly false tweets with minor truthful elements, and 5 represents a completely false tweet. The factuality analysis of a tweet can culminate in one of three ways: it can be confirmed as true (scores 1 or 2), proven as false (scores 4 or 5), or remain unresolved (most likely score 3). Given a single event or topic, multiple variations of misinformation can circulate simultaneously, each holding varying degrees of truth. The resolution of one tweet can often lead to the automatic resolution of other related tweets. For instance, in the context of the Russia–Ukraine conflict, consider a situation where false information is circulating in social media about the origins of a particular military escalation: (1) *The escalation was triggered by a minor border incident involving local forces.* (Score of 3, as it is partially true that the conflict started from a border incident; however, this ignores the larger geopolitical context.) (2) *The escalation was entirely instigated by Ukraine.* (Score of 5, as this completely ignores Russia's role.) (3) *The escalation was completely unprovoked and unexpected.* (Score of 5, as it is false given that there were clear signs of escalating tensions.) (4) *The escalation was the result of long-standing tensions between Russia and Ukraine, magnified by a specific border incident.* (Score of 1, assuming this is a confirmed truth.) Now, once (4) is confirmed as true, it invalidates (2) and (3) entirely and provides additional context for (1). Thus, we demonstrated two important issues: (a) the need for continuous and real-time evaluation of the information shared on social media, assigning a misinformation score that could help users gauge the veracity of a tweet instantly; (b) the need for an adaptive model to account for the dynamic nature of social media content, which becomes vital in situations where the context and the validity of information may change over time, as seen in the evolving narratives during conflicts or health crises in COVID-19 (Loomba et al. 2021).

Contributions

We address these issues by proposing a dynamic, comprehensive system to evaluate and to assign misinformation scores to tweets in real time. We make several contributions:

- *Integration of TGNs and RNNs:* We combine Temporal Graph Networks (TGNs) and Recurrent Neural Networks (RNNs) to capture both structural and temporal characteristics of misinformation propagation. RNNs model temporal aspects, considering the content and the associated timestamped events, while TGNs capture both structural and temporal aspects by maintaining evolving node-level states based on interactions and timestamps. TGNs use temporal message passing and memory modules to capture the spread of misinformation over time and to identify key propagators in the network. This novel combination enables quick adaptation to changes in the information context, thereby enhancing the timeliness and the accuracy of misinformation detection.
- *Active Learning:* Our system uses active learning techniques to iteratively refine its understanding of misinformation, soliciting feedback on the most uncertain or challenging cases, leading to improved accuracy over time.
- *Dual Model System:* We propose a unique dual-model system that comprises a strict model with a higher threshold for identifying misinformation and a normal model, following a more lenient approach. By combining the outputs of these models, we ensure a more nuanced and accurate grading, improving the system's reliability.

We propose the concept of a "temporal embargo", which refers to the practice of withholding judgment on the veracity of a piece of information for a certain time period. This delay allows for a more comprehensive assessment, as additional context or information might emerge over time that can influence the final decision. The belief score is a measure of the model's confidence in the veracity of a piece of information. In this work, we propose a novel dynamic assessment model that, rather than issuing an immediate verdict, implements a temporal embargo based on the belief score. If the belief score falls within a certain range, indicating uncertainty, the model will withhold the judgment for different time intervals, periodically reassessing the information until the model's prediction stabilizes or new information arises. The challenge lies in determining the appropriate embargo periods and the belief score thresholds, as well as ensuring that the model can adapt and scale to the high volume and real-time nature of social media content. We investigate the effectiveness of this temporal embargo strategy in reducing misinformation propagation across various social media misinformation datasets. Our system achieves promising accuracy in identifying false information on five datasets.

Problem Statement

The objectives of our system are as follows:

(R1) *To develop a system that continuously evaluates and assigns a misinformation score to tweets in a social network, on a scale from 1 (completely true) to 5 (completely false).* Formally, given a set $T = \{t_1, t_2, \ldots, t_n\}$ of tweets, and a function $F : T \rightarrow [1, 5]$, the problem is to design a system S such that, for each tweet $t_i \in T$, computes $F(t_i)$ in real-time, where $F(t_i)$ represents the misinformation score of t_i.

(R2) *To develop an effective and adaptive model for misinformation detection that takes into account the dynamic nature of social media content and incorporates a temporal embargo strategy based on belief (confidence of the model's prediction) scores of the misinformation classification models.* Formally, given a set $T = \{t_1, t_2, \ldots, t_n\}$ of tweets, a function $F : T \rightarrow [1, 5]$, a dynamic nature of social media content represented by a time-varying function $G(t)$, the problem is to design an adaptive model M that, for each tweet $t_i \in T$, computes $F(t_i)$ at time $G(t)$, and incorporates a temporal embargo strategy based on belief scores $B : M \rightarrow [0, 1]$ of the misinformation classification models. $B(M)$ represents the belief score of model M, such that $B(M) = 0$ implies no belief in the model's classification,

and $B(M) = 1$ implies complete belief in the model's classification.

(R3) *To study the impact of retraining in an incremental training step, looking into aspects such as periodicity, change detection in the network, and linguistic score change.* Given a model M, a set $T = \{t_1, t_2, \ldots, t_n\}$ of tweets, a function $F : T \to [1, 5]$, a time-varying function $G(t)$ representing the dynamic nature of social media content, and a change detection function $\Delta : (T, G) \to \{0, 1\}$, the problem is to define a retraining strategy $R : M \times \Delta \to M'$ such that M' is an improved model after retraining. The retraining strategy must consider factors such as the periodicity P of retraining, represented by a function $P : M \to \mathbb{N}$, and the linguistic score change $L : T \times M \to \mathbb{R}$, where $L(t_i, M)$ is the change in the misinformation score for tweet t_i under model M.

The problem statements outlined here address aspects that have not been adequately tackled by existing approaches. While previously developed systems attempt to classify misinformation, our work (R1) to assign a continually updating misinformation score to social media posts in real time is an advancement over previous work. This dynamic scoring system can promptly identify the spread of harmful misinformation before it gains traction, a capability not fully realized in current models. Next, incorporating a temporal embargo strategy (R2) based on the belief scores of the misinformation classification models, allows for a better understanding of the information's veracity over time. This adaptability is essential to account for changes in the content, the context, and the tactics adopted by misinformation spreaders. While the importance of model training and retraining is recognized, the outlined strategy (R3) for defining a systematic and context-responsive retraining approach is a substantial contribution. This strategy takes into account factors such as periodicity, network change detection, and linguistic score changes, thus ensuring that the model remains current and robust in the dynamic misinformation landscape. This adaptive approach to retraining is not adequately covered in current methodologies.

Proposed Framework

Our proposed framework for misinformation detection combines several novel elements: a dynamic grading system that adapts to new information or to context changes, incorporation of temporal features to capture the evolution of misinformation over time, use of active learning techniques to refine the model's understanding of misinformation and its spread, and adoption of a dual grading system combining a strict and a normal model.

Dynamic Grading System (DGS)

The Dynamic Grading System (DGS) aims to provide a more responsive and adaptable measure of misinformation, accounting for the fluctuating nature of information spread in social media. The grading system is based on the concept of misinformation scores, which are continuously updated as new information or user feedback comes in. Each piece of content (e.g., a tweet) is assigned a misinformation score $b \in [1, 5]$, with 1 denoting absolute truth and 5 denoting absolute falsehood.

Incremental Learning. We adopt an incremental learning strategy to update misinformation scores as new data comes in. Here, new data does not only mean a new tweet, but a variety of contextual elements. This includes the profiles of users who retweet or like the tweet, the time at which the tweet was propagated, the network of users who interacted with the tweet, and the sentiment and the content (agree or disagree towards the topic of the discourse) of any responses. Let X_t denote the set of features associated with a tweet at time t, and b_t denote its misinformation score at time t. The model learns a function $f' : X_t, C_t \to b_t$ that maps features to misinformation scores. Here, C_t represents the contextual features at time t. When new data X_{t+1} comes in at time $t + 1$, the model updates its function f to $f' : X_{t+1}, C_{t+1} \to b_{t+1}$, effectively learning from the new data. Formally, the model updates its function as follows:

$$f' = f + \alpha \cdot (b_{t+1} - f'(X_{t+1}, C_{t+1})) \cdot \nabla f'(X_{t+1}, C_{t+1}) \quad (1)$$

where α is the learning rate, $\nabla f'(X_{t+1}, C_{t+1})$ is the gradient of f' with respect to (X_{t+1}, C_{t+1}). The term $(b_{t+1} - f'(X_{t+1}, C_{t+1}))$ represents the error in the model's prediction at time $t+1$, considering both the tweet and its context.

Real-Time Feedback. In addition to incremental learning, we incorporate real-time user feedback into our dynamic grading system. We introduce a feedback function $g : R \to [0, 1]$, where R represents user reactions,e.g., likes, shares, comments. This function maps user reactions to misinformation score adjustments, which are then used to update the misinformation score of a tweet. Formally, the updated misinformation score b_{t+1} at time $t + 1$ is computed as follows:

$$b_{t+1} = b_t + \beta \cdot g(R_{t+1}) \quad (2)$$

where β is a parameter controlling the impact of user feedback on misinformation scores, and R_{t+1} is the user reactions at time $t + 1$.

For each tweet, our system can query a set of trusted fact-checking websites, e.g., Snopes, PolitiFact, FullFact etc.[1] with the key claims present in the text. These sites then return a fact-checking score $f_c \in [1, 5]$ based on their assessment of the claim, with 1 indicating that the claim is entirely true, and 5 meaning that the claim is entirely false. Next, to integrate the fact-checking score with the existing misinformation score b_t, we introduce a fact-checking weight parameter $\gamma \in [0, 1]$. The updated misinformation score b_{t+1} at time $t + 1$ is then computed as follows:

$$b_{t+1} = (1 - \gamma) \cdot (b_t + \beta \cdot g(R_{t+1})) + \gamma \cdot f_c \quad (3)$$

The fact-checking weight parameter γ determines the importance of the fact-checking score relative to the updated misinformation score. To make the system more responsive to real-time events, we make the fact-checking weight γ adaptive. In situations where the fact-checking score and the misinformation score are significantly different, we increase γ to put more emphasis on the fact-checking score. Conversely, when the scores are relatively similar, we decrease γ to put more weight on the updated misinformation score.

[1] In this work, we use FullFact.

Temporal and Structural Features

Capturing the evolution of misinformation over time and its propagation through the structure of a social network is a critical aspect of our framework. We propose the use of Recurrent Neural Networks (RNNs) and Temporal Graph Networks (TGNs) to address these challenges.

Temporal Features with Recurrent Neural Networks (RNNs). In the context of misinformation grading, we consider each piece of content along with its associated time-stamped events (likes, shares, comments, etc.) as a sequence of inputs. Let $X = x_1, x_2, \ldots, x_T$ denote the sequence of features associated with a piece of content from time 1 to T, and $b = b_1, b_2, \ldots, b_T$ be the corresponding sequence of misinformation scores. An RNN updates its hidden state h_t at each time step t based on the current input x_t and the previous hidden state h_{t-1}:

$$h_t = \sigma(W_h h_{t-1} + U_h x_t + b_h) \tag{4}$$

where W_h and U_h are weight matrices, b_h is a bias vector, and σ is an activation function. The misinformation score at time t is then computed as

$$b_t = \sigma(W_o h_t + b_o) \tag{5}$$

where W_o is a weight matrix and b_o is a bias vector.

Temporal Graph Networks (TGNs). TGNs (Rossi et al. 2020) capture both the structural and the temporal aspects of a graph, making them an ideal tool for our task. TGNs work by maintaining an evolving node-level state that is updated every time an interaction involving that node occurs. The TGN updates the state of a node based on its previous state, the timestamp of the interaction, and the states of the interacting nodes. We propose to deploy a TGN to capture the temporal evolution of each node's neighbourhood, which is often overlooked in traditional GCN-based methods.

TGNs model the social network as a graph $G(V, E, T)$, where V represents the set of nodes (users), E is the set of edges (relationships and interactions), and T is the timestamp associated with each edge (interaction). Each node $v \in V$ maintains an evolving state $s(v, t)$ that is updated every time an interaction (or update in the relationship, like a user follows/ unfollows another one) involving node v occurring at time t. The updated state $s(v, t')$ (where $t' > t$) is computed as

$$s(v, t') = f(s(v, t), s(u, t), t' - t, x(v, u, t')) \tag{6}$$

where f is a function that combines the previous state of the node $s(v, t)$, the state of the interacting node $s(u, t)$, the time difference between the current and the previous interaction $t' - t$, and the features $x(v, u, t')$ associated with the interaction.

Temporal Message Passing: The temporal aspect of TGNs is crucial here. Misinformation does not propagate instantaneously, but rather spreads over time in a social network. TGNs, through temporal message passing, can model this aspect by learning from the sequence of interactions. For instance, a retweet or a reply that supports a rumor at a later stage may be more impactful in the spread of the rumor than an earlier interaction.

Memory Modules: The memory modules in TGNs can capture the state of a user (node) over time. This is helpful as user behavior might change as the rumor spreads: they might initially believe and propagate the rumor, but could later debunk it when provided with new information.

By continuously updating user states and their interactions, TGNs can provide a dynamic grading of the propagation of misinformation. For instance, a rumor might initially have a high misinformation score because many users are propagating it. However, as more users debunk the rumor, its misinformation score can decrease over time. TGNs can also help identify key propagators of misinformation in a social network. These are the nodes that play a significant role in the spread of misinformation. Our proposed model has significant advantages. For instance, a user tweeted a doctored image falsely showing a major Ukrainian city being under siege. This tweet might initially propagate rapidly among certain user groups, leading to a high misinformation score. As the image gets fact-checked and debunked by reliable sources, our framework will update the misinformation score, reflecting the new truth value of the claim. This is important because it ensures that users who encounter the tweet later, perhaps through search or late shares, would have an accurate understanding of its credibility. Also, a tweet claiming that vaccines cause autism initially spreads rapidly and receives a high misinformation score due to its wide propagation by anti-vaccination advocates. However, using the proposed framework, as more credible users, health experts, and fact-checking organizations debunk this claim, the misinformation score would dynamically decrease. This is crucial to reflect the changing nature of the information landscape and to prevent outdated misinformation scores from misleading social network users.

Transformers for Temporal and Contextual Attention. For a given sequence of time-stamped social media events $E = \{e_1, e_2, ..., e_n\}$ associated with a piece of content (tweet), we feed these events into RoBERTa. Each event e_i is represented as a vector combining its features $x(e_i)$ and the time difference $\Delta t(e_i)$ from the previous event. For instance, in the context of the FakeNewsNet dataset, each retweet, reply, or like event associated with the tweet T_1 can be an event in E. The features $x(e_i)$ for an event e_i include the text of the tweet, the user profile information, and the number of retweets/likes/replies at the time of the event. The time difference $\Delta t(e_i)$ is the time elapsed since the previous event. RoBERTa processes this sequence and produces a sequence of output vectors $O = \{o_1, o_2, ..., o_n\}$, where each o_i is a weighted combination of all input events:

$$o_i = Attention(Q_i, K, V), \tag{7}$$

where $Q_i = W_q * e_i$ is the query associated with event e_i. $K = W_k * E$ and $V = W_v * E$ are the keys and the values computed from all events, W_q, W_k, W_v are learned weight matrices, and Attention is the scaled dot-product attention function.

By combining TGNs and RoBERTa, we effectively capture the structural and the temporal aspects of misinformation propagation, thus providing a more comprehensive basis for the dynamic grading of misinformation.

Ensemble

Our ensemble combines three models:

(1) Content Analysis Model (CAM). To analyze the semantic content of the posts, we use RoBERTa, which we fine-tune for our task of dynamic misinformation grading. To capture stylistic features prevalent in misinformation, such as the usage of capital letters, exclamation marks, and other non-standard grammatical structures, we propose implementing a stylometric analysis. In this study, our stylometric analysis focuses on the extraction of stylistic features that have been shown in previous research to be indicative of misinformation (Przybyla 2020). These features include: *(i) Lexical features*: measures of lexical richness, such as type-token ratio, hapax legomena, and average word length. Lexical features also include the use of specific types of words, like the use of emotionally charged words, biased language, or sensational terms. *(ii) Syntactic features*: This refers to the arrangement of words in a sentence and grammatical structures. Unusual or non-standard syntactic structures could potentially be indicative of misinformation. *(c) Punctuation*: The use of certain punctuation marks, such as exclamation points or capital letters, has been shown to be indicative of misinformation. For example, excessive use of exclamation marks or all capital letters can indicate a sensationalistic tone that is common in misinformation. *(d) Complexity*: This includes sentence length, sentence complexity, and the use of passive versus active voice. The output of the Content Analysis Model (CAM) is a probability that a given post is misinformation, denoted as $P_{CAM(misinformation)}$. It combines the semantic understanding from RoBERTa and the stylistic insights from the stylometric analysis to provide a comprehensive assessment of the post content.

(2) User Behavior Model (UBM). This model focuses on user behavior features. This includes user posting frequency, the ratio of original posts to reposts, followers-to-following ratio, and account age. These features are fed into a gradient boosting model (XGBoost), which can handle high-dimensional heterogeneous feature spaces and provides feature importance.

(3) Propagation Pattern Model (PPM). This model analyzes the propagation patterns of the posts in the social network. The Temporal Graph Networks (TGNs) combined with the Transformer model (as described in the previous section) are used here to capture the temporal and the structural aspects of the propagation of misinformation. The outputs from each of these models are probabilities that a given post is misinformation. We combine these probabilities using a decision-level fusion strategy. Specifically, we use weighted averaging where the weights are learned from the validation data:

$P(misinformation) = w_1 * P_{CAM(misinformation)} + w_2 * P_{UBM(misinformation)} + w_3 * P_{PPM(misinformation)}$ where $P_{CAM(misinformation)}, P_{UBM(misinformation)}$ and $P_{PPM(misinformation)}$ are the probabilities output by the Content Analysis Model, the User Behavior Model, and the Propagation Pattern Model, respectively, and w_1, w_2, w_3 are the weights learned from the validation data (and they sum up to 1).

Active Learning

The goal of Active Learning is to construct a model that fits the data accurately, while reducing the amount of labeled data needed to construct the model. In the context of our framework, we incorporate active learning (See Algorithm 1) to iteratively refine our misinformation grading model, particularly during the incremental training steps. Given a pool of unlabeled data U, and a smaller set of labeled data L, the goal of active learning is to select data points from U that, when labeled and added to L, are most beneficial for improving the model's performance. Let F be our misinformation grading model, parametrized by θ, and let us denote the misinformation score at time t as $B_t(x;\theta)$, for a content x. We define the selection function $S(U, F) \rightarrow x$, which selects a data point x from the pool U based on the current model F. This function is typically designed to select data points for which the model is most uncertain or which are expected to provide the most information if labeled. A common strategy for S is to select the data point x for which the model's prediction is closest to the threshold of 0.5 (i.e., the model is most uncertain). Formally, this can be defined as follows:

$$S(U, F) = argmin_{x \in U} |B_t(x;\theta) - 0.5| \quad (8)$$

We propose an event-driven approach, where re-training is triggered based on either change detection in the network, or a significant change in the linguistic score. Let us denote as ΔN_t the change in the network at time t, and as $\Delta L_t(x)$ the change in the linguistic score of content x at time t. We can then define a re-training condition as follows: If $\Delta N_t > \tau_N$ or $max_{x \in U} \Delta L_{t(x)} > \tau_L$, re-train the model, where τ_N and τ_L are pre-defined thresholds. Periodicity can be further incorporated as a fallback option, to ensure that the model is re-trained at regular time intervals in case the above condition is not met for a prolonged period of time. We denote as T the periodicity of re-training (e.g., once every 7 days).

Algorithm 1: Active Learning for Misinformation Grading

Require: U: unlabeled data pool, L: labeled data, F: misinformation grading model, τ_N: network change threshold, τ_L: linguistic score change threshold, T: re-training periodicity

```
Initialize F with initial parameters θ using L
while stopping condition not met do
    x ← arg min_{x'∈U} |B_t(x';θ) − 0.5|  Select a data point
    Obtain a label for x and add it to L
    ΔN_t ← change in network at time t
    ΔL_t(x) ← change in linguistic score of x at time t
    if ΔN_t > τ_N or max_{x'∈U} ΔL_t(x') > τ_L or current time − last re-training time > T then
        Re-train F with L
    end if
end while
```

Dual Grading System

Given a social media post x, the misinformation grading models, the Strict Model (S) and the Normal Model (N), output misinformation scores $S(x;\theta_S)$ and $N(x;\theta_N)$ respectively, where θ_S and θ_N are the model parameters.

The models are trained to output a misinformation score in the range $[1, 5]$, where 5 indicates that the post is definitely misinformation, and 1 that it definitely is not. The strict model has a higher threshold τ_S for classifying a post as misinformation. Conversely, the Normal Model, with its lower threshold τ_N, is more lenient. Accordingly, we define two distinct loss functions to train these models:

(A) Strict Model Loss. The Strict Model is designed to minimize the number of false negatives (FN), where an actual misinformation post is not flagged as misinformation. The loss function for this model is as follows:

$$L_S(y,\hat{y};\theta_S) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\beta(y_i-\hat{y_i})^2 & \text{if } y_i \geq 3 \text{ and } \hat{y_i} < y_i\\ (y_i-\hat{y_i})^2 & \text{otherwise}\end{cases} \tag{9}$$

In this equation, y and $\hat{y}$ represent the actual and the predicted misinformation scores, respectively, n denotes the number of samples, and $\beta > 1$ (along with $\gamma > 1$, which is explained below) is a weight that can be adjusted based on the specific needs of the problem at hand.

(B) Normal Model Loss. This model is designed to minimize the number of false positives (FP), where an actual non-misinformation post is inaccurately flagged as misinformation. The loss for this model is as follows:

$$L_N(y,\hat{y};\theta_N) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\gamma(y_i-\hat{y_i})^2 & \text{if } y_i \leq 3 \\ & \text{and } \hat{y_i} > y_i\\ (y_i-\hat{y_i})^2 & \text{otherwise}\end{cases} \tag{10}$$

In both loss functions, we use a threshold of three in order to differentiate between misinformation and non-misinformation posts. We tune the models separately using their respective loss functions to output a misinformation score in $[1, 5]$. This dual-model approach to misinformation detection ensures a robust yet flexible system capable of addressing varying degrees of misinformation severity, thereby enhancing the integrity of the information circulating in social networks.

To measure the misinformation score effectively by deploying a temporal embargo strategy, we are interested to capture the model's degree of certainty (or belief) about its own predictions. In the context of neural networks, this can be interpreted as the entropy of the predictive distribution (Gal and Ghahramani 2016). The entropy of a probability distribution provides a measure of the uncertainty associated with the distribution. We used the softmax model output to quantify its predictive uncertainty. A lower entropy corresponds to a higher confidence level, as the model predictions are more concentrated on certain classes. This confidence can be computed as one minus the normalized entropy. Let $P_S(x;\theta_S)$ and $P_N(x;\theta_N)$ represent the softmax outputs (predictive distributions) of the strict and of the normal models, respectively, for content x. Then, the entropy of these distributions is given by

$$E_S(x) = -\sum_i P_S(x;\theta_S)_i \log P_S(x;\theta_S)_i \tag{11}$$

$$E_N(x) = -\sum_i P_N(x;\theta_N)_i \log P_N(x;\theta_N)_i \tag{12}$$

Here, the summations are over all classes i. To convert these entropy measures into confidence levels, we first normalize them to the range [0, 1] by dividing by $\log K$, where K is the number of classes. Then, we subtract the normalized entropy from 1:

$$C_S(x) = 1 - \frac{E_S(x)}{\log K} \tag{13}$$

$$C_N(x) = 1 - \frac{E_N(x)}{\log K} \tag{14}$$

In the dual grading system, a post is classified as misinformation if both models agree that it is misinformation after a certain time span t of observation. Mathematically, this can be expressed as follows:

$$M(x;t) = \begin{cases}1, & \text{if } S(x;\theta_S) \geq \tau_S \text{ and}\\ & N(x;\theta_N) \geq \tau_N \text{ after time } t,\\ 0, & \text{otherwise,}\end{cases} \tag{15}$$

where $M(x;t)$ is the final decision on whether the post x is misinformation after observing it for time t.

In addition, a belief score threshold is introduced to further refine the grading system. If the belief score is greater than or equal to 0.80, the post is labeled as FALSE. If it falls in between 0.30 and 0.80, the post is embargoed for different time intervals and periodically checked if the model changes its prediction. This can be expressed as follows:

$$\text{Label}(x) = \begin{cases}\text{FALSE}, & \text{if } M(x;t) \geq 0.80,\\ \text{EMBARGO}, & \text{if } 0.30 \leq M(x;t) < 0.80,\\ \text{TRUE}, & \text{otherwise.}\end{cases} \tag{16}$$

where $\text{Label}(x)$ is the final label assigned to the post x.

Experimental Analysis

The primary objectives of this work are three-fold: (*i*) to continuously evaluate and assign a misinformation score to tweets, (*ii*) to develop an effective and adaptive model for misinformation detection that incorporates a temporal embargo strategy, and (*iii*) to define an effective retraining strategy for incremental learning.[2]

Data

Datasets. For our evaluation, we used a subset of the tweets in the five datasets shown in Table 1: AntiVax (anti-vaccine), CONSTRAINT (COVID-19-related fake news), FakeNewsNet, PHEME (rumor detection), and RU War (Russia–Ukraine War). First, we carried out data annotation to assign a misinformation score to tweets from each dataset. We

[2]Codebase: https://github.com/shreyaghosh-2016/Misinformation-detection

Dataset	Details
PHEME (Derczynski and Bontcheva 2014)	330 rumor threads collected from Twitter, each having an average of 100 tweets.
AntiVax (Hayawi et al. 2022)	Over 1.8 million tweets from the anti-vaccination movement collected between 2019 and 2021.
CONSTRAINT (Patwa et al. 2021)	17,000 English tweets (COVID-19), annotated as either real or fake, with an equal distribution.
FakeNewsNet (Shu et al. 2020)	Data from two fact-checking websites, PolitiFact and GossipCop, over 23,000 news articles.
RU War (Chen and Ferrara 2022)	Tweets from Feb 22,2022 through Jan 8, 2023 collected using hashtags related to RU war.

Table 1: The five real-world datasets we used for the performance evaluation of our framework.

randomly selected 1,000 tweets that had more than 50 interactions including re-tweets, quote-tweets, likes, and comments. Note that all tweets are timestamped.

Annotators. We used Amazon Mechanical Turk (MTurk) to perform the annotation. The MTurk workers, also known as *Turkers*, were presented with the tweets and were asked to evaluate the veracity of each tweet on a scale from 1 (meaning completely true) to 5 (meaning completely false). Each tweet was evaluated by at least three different Turkers to ensure a robust annotation process. The Turkers were given clear and detailed guidelines in order to help them distinguish between different levels of misinformation. We aggregated the scores by the different Turkers by averaging.

Inter-Annotator Agreement. The annotations were largely consistent, with a very high inter-annotator agreement of 0.83 in terms of Krippendorff's Alpha.

Average Misinformation Scores. The average misinformation scores for the tweets in the PHEME, the AntiVax, the FakeNewsNet, the CONSTRAINT, and the RU War datasets were 3.1, 3.8, 2.9, 3.2, and 3.2, respectively, suggesting a varying degree of misinformation across the different datasets. We noted that the AntiVax dataset had a much higher average misinformation score of 3.8, while the scores for the other datasets were quite close, ranging in [2.9; 3.1].

Experiments and Evaluation

Comparison to Human Annotations Using QWK. First, we used the Quadratic Weighted Kappa (QWK) as an evaluation measure to quantify the performance of our misinformation detection model. QWK considers the possibility of agreement occurring by chance, which makes it a more reliable measure than simple percent agreement calculation. QWK also takes into account the order of the categories and gives more weight to disagreements when the categories are further apart. QWK is a robust statistical measure, specifically designed to evaluate the agreement between two raters, each of which classifies N items into C mutually exclusive categories. In our scenario, the two raters were the human annotators (averaged scores, used as a gold standard) and the scores from our misinformation detection model, both assigning a misinformation score on a scale from 1 to 5. Figure 1 shows that our model achieves a QWK score of 0.85 on the PHEME dataset, 0.89 on the AntiVax dataset, 0.83 on the FakeNewsNet dataset, 0.77 on the RU War dataset, and 0.81 on the CONSTRAINT dataset. These scores indicate a substantial agreement between the model's predictions and the human annotations (gold standard), reflecting the robustness and the effectiveness of our proposed model in continuous misinformation scoring. Our model has consistently produced QWK scores above 0.77, thus demonstrating its ability to adapt to different types of data and different types of misinformation.

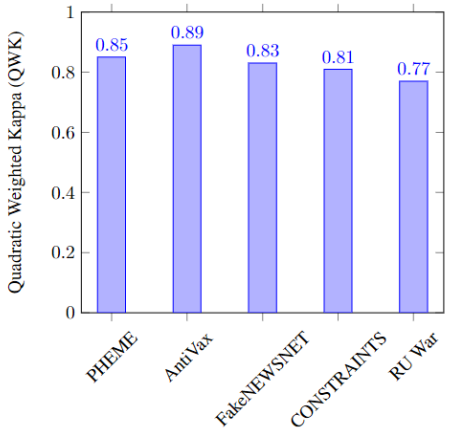

Figure 1: QWK score for our model across five datasets.

Comparison to Previous Models over Different Timestamps. Table 2 compares our system's performance to five state-of-the-art misinformation detection models: CAMI (Yu et al. 2017), FNED (Liu and Wu 2020), GRU (Ma et al. 2016), and (Yue et al. 2022). In order to understand the model's real-time performance and its ability to adapt to rapidly evolving information landscapes, we compare over three different time frames: 24 hours, 12 hours, and 30 minutes. We can see that our system consistently outperforms the other systems in terms of F1 score across the five datasets and also across the three timeframes. The most sizable improvements are observed for AntiVax and RU War.

Per-class Performance over Time. Figure 2 shows the per-class misclassification ratio for our framework on the FakeNewsNet dataset for tweets at different timestamps. We

Dataset / Time	F1 Score				
	Our framework	CAMI (Yu et al. 2017)	FNED (Liu and Wu 2020)	GRU (Ma et al. 2016)	(Yue et al. 2022)
AntiVax / 24h	**0.941**	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	**0.920**	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
RU War / 24h	**0.942**	0.758	0.728	0.810	0.711
RU War / 12h	**0.863**	0.622	0.548	0.692	0.506
RU War / 30m	**0.620**	0.573	0.501	0.606	0.481

Table 2: Comparing our framework to various models across five datasets and three timestamps (F1 score). For fair comparison, here the *temporal embargo* strategy is turned off for our method.

can see for Class 1 (entirely true) a consistent decrease in the misclassification ratio over time from 0.21 at 15 minutes to 0.15 after a day. For Class 2 (mostly true tweets with minor inaccuracies), the decrease is from 0.19 to 0.14. Class 3 (partially true tweets with significant inaccuracies) maintains a steady misclassification ratio of 0.25 during the initial 45 minutes, and then gradually decreases to 0.18. The most sizable decrease is observed for Classes 4 and 5, which represent false tweets with minor elements of truth and entirely false tweets, respectively. Class 4 begins with a high misclassification ratio of 0.53 at 15 minutes, and decreases to 0.11. For Class 5, the misclassification ratio drops from 0.58 to 0.15. Overall, the results indicate a steady decrease in misclassification for all classes over time, particularly for false tweets.

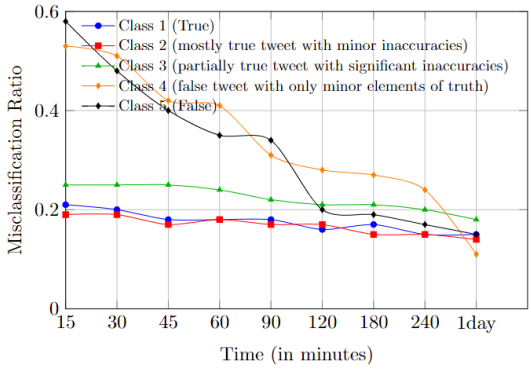

Figure 2: Per-class misclassification ratio for our framework on FakeNewsNet for tweets at different timestamps.

Note that the FakeNewsNet dataset includes examples collected from two fact-checking organizations: PolitiFact and GossipCop. We looked into the performance for each subset and we found some notable differences. As PolitiFact largely deals with political claims while GossipCop focuses on celebrity gossips, the nature of misinformation is inherently different. Political misinformation is more strategically crafted and nuanced, making it more challenging to classify accurately within a short period of time. In contrast, celebrity gossips are less sophisticated, enabling quicker and more accurate classification early on. Since the PolitiFact subset contains more time-sensitive and rapidly evolving information, the accuracy of classification improves significantly over time as more context becomes available. On the other hand, GossipCop data is relatively static, and thus the classification accuracy does not improve as much over time.

Impact of the Temporal Embargo and the Dual Loss. To evaluate the impact of the temporal embargo strategy, we compared our model's initial classification (before the embargo) and the final classification (after the embargo). The results are shown in Table 3, where we also compare the impact of the loss used: strict, normal, and dual. We can see sizable improvements when using the embargo strategy, as well as when using the dual loss.

Table 4 further shows the reduction in the percentage of interactions when using the embargo strategy. This is calculated as the number of interactions within the embargo period divided by the total number of interactions in the dataset. We can see sizable reduction with a longer embargo, especially for the FakeNewsNet and the PHEME datasets.

Impact of the Retraining Strategy. Table 5 shows the impact of our retraining strategy for incremental learning. We can make several interesting observations:

- *Performance Gains:* Comparing the performance before and after retraining in columns 2 and 3, we can see consistent and sizable performance gains across the five

Model	w/o Temporal Embargo	w/ Temporal Embargo
Strict loss	0.830	0.901
Normal loss	0.760	0.860
Dual (strict+normal) loss	0.847	0.952

Table 3: Impact of the temporal embargo and the loss (F1).

	Datasets																			
TE	PHEME				AntiVax				FNN				CNT				RUW			
	R	Q	C	L	R	Q	C	L	R	Q	C	L	R	Q	C	L	R	Q	C	L
10m	5	4	6	7	3	2	3	2	6	5	7	8	4	3	5	6	2	1	2	2
20m	10	8	12	14	6	4	6	4	12	10	14	16	8	6	10	12	4	2	4	4
30m	15	12	18	21	9	6	9	6	18	15	21	24	12	9	15	18	6	3	6	6
45m	22.5	18	27	31.5	13.5	9	13	9	27	22.5	31.5	36	18	13	22.5	27	9	4.5	9	9
60m	30	24	36	42	18	12	18	12	36	30	42	48	24	18	30	36	12	6	12	12
240m	60	48	72	84	36	24	36	24	72	60	84	96	48	36	60	72	24	12	24	24

Table 4: Reduction in the percentage of interactions when using our framework with temporal embargos (TE) of different lengths: re-tweets (R), quote tweets (Q), comments (C), and likes (L). Here, the unit of temporal embargo (TE) is minute. FNN: FakeNewsNet, CNT: Constraint, RUW: RU War.

datasets when using retraining.

- *Response to New Patterns:* The following column 4 in the table shows to what extent the model can adapt to new misinformation patterns post retraining. For example, in the RU War dataset, a new pattern emerged where misinformation was propagated through manipulated images. Our retrained model successfully identified 88% of these new instances, which is a sizable improvement. Note that we did not use any visual training for the identification of this new pattern.
- *Computational Efficiency:* Column 5 shows that retraining does not incur substantial computational overhead, with time increasing by 12-16%, which is acceptable given the sizable improvement in model performance.
- *Stability vs. Plasticity Trade-off:* Our retraining strategy aimed to maintain a balance between stability (preserving old learnings) and plasticity (adapting to new patterns). On the AntiVax dataset, the retrained model maintained an accuracy of 78% on old instances, while improving the accuracy on new instances from 64% to 81%. (Additional analysis, not shown in the table.)

Related Work (‡)

Early work on misinformation detection focused on extracting various features from content and user behavior; see (Meel and Vishwakarma 2020) for a recent survey. Subsequently, advanced machine learning and deep learning techniques started to dominate the field. Zhou and Zafarani (2018) used Support Vector Machines (SVM) to classify fake news, based on account, social graph, and linguistic features. However, these models faced limitations in their adaptability to new information and to context changes, and they lacked the ability to account for temporal changes.

Ruchansky, Seo, and Liu (2017) introduced the CSI model, which incorporated an LSTM in order to capture the temporal dependencies in the propagation of news. Ma et al. (2016) proposed a deep learning-based model that uses neural networks to learn representations of news articles for fake news detection. However, both models lack an effective mechanism for dynamic updates based on new information or to context changes.

Considering the complex relationship within social networks, graph-based approaches have been proposed. Wu and Liu (2018) developed a Graph Convolutional Network (GCN) based model to detect fake news. Similarly, Monti et al. (2019) presented a geometric deep learning approach to misinformation detection, exploiting the graph structure of social networks. Such models often failed to effectively integrate temporal features, which limited their effectiveness.

Vosoughi, Roy, and Aral (2018) conducted a comprehensive study on the spread of true and false news online, but their work lacked an explicit mechanism for explaining their model's decisions. Shu et al. (2017) proposed an ensemble of multiple specialized detectors to capture different aspects of the fake news. However, the dynamic adaptation of these models to new information or context changes still presents a challenge.

Barnabò et al. (2023) leveraged active learning strategies applied to Graph Neural Networks for misinformation detection where their proposed architecture called Deep Error Sampling (DES) combined with uncertainty sampling, performs equally or better than common active learning strategies and the only existing active learning procedure designed for fake news detection to date. FANG (Nguyen et al. 2020) leveraged a graph-based social context representation, outperforming other models in capturing the social context and

Dataset	Accuracy before Retraining	Accuracy after Retraining	Response to New Patterns	Retraining Time Increase
PHEME	76%	82%	70%	12%
FakeNewsNet	70%	78%	75%	14%
AntiVax	72%	79%	85%	16%
RU War	72%	81%	88%	15%
CONSTRAINT	75%	83%	80%	15%

Table 5: Impact of the retraining strategy in our framework across the five datasets.

achieving significant improvements in fake news detection, even with limited training data. Rumor Gauge (Vosoughi, Mohsenvand, and Roy 2017) predicted the veracity of rumors on Twitter in real-time, but did not provide an intervention strategy. The study of (Barnabò et al. 2022) highlighted the scarcity of high-quality benchmark datasets for online misinformation detection and the potential overestimation of state-of-the-art approaches due to this limitation. They presented FbMultiLingMisinfo, a multilingual benchmark dataset derived from Facebook and augmented with Twitter propagation paths.

Relevant to our objectives, Friggeri et al. (2014) investigated the spread of rumors on Facebook, focusing on the early stages of rumor propagation. They analyzed user engagement and debunking behavior to understand the factors contributing to the persistence of misinformation. Another work (Djenouri et al. 2023) introduced a parallel pattern-mining framework called DMRM-FNA to address the difficulties in the dynamic assessment of misinformation using big data exploration, but the work mainly focused on improving the computational cost and did not mention how temporal intervention can help in mitigating misinformation propagation. Azzimonti and Fernandes (2023) analyzed the impact of social media network structure and fake news on misinformation and polarization in society. They explored the evolution of agents' opinions and the role of internet bots in spreading fake news and they found that even with a relatively small percentage of agents believing fake news, significant misinformation and polarization can occur, emphasizing the importance of network effects. (Agarwal et al. 2022) introduced THINK, a novel framework for network-based time series forecasting that leveraged hypergraph learning and hyperbolic properties. By capturing higher-order relations and scale-free characteristics, THINK outperformed state-of-the-art methods across various tasks. Tardelli et al. (2023) presented the first dynamic analysis of coordinated online behavior by building a multiplex temporal network and employing dynamic community detection. In that direction, (Hristakieva et al. 2022) investigated the interplay between propaganda and coordinated behavior in online debates, specifically focusing on the 2019 UK general election on Twitter. Although the primary objectives of these studies are different, the results demonstrated that coordinated communities exhibit varying levels of temporal instability, emphasizing the need for dynamic analyses.

As mentioned in the survey of (Meel and Vishwakarma 2020), the two major issues are real-time learning for fact-checking to adapt to emerging misinformation and to provide up-to-date detection of false information, and the task of determining the truthfulness or accuracy of information, which becomes challenging due to the complex and dynamic network structure of social platforms. The existing work does not specifically address the continuous grading of misinformation or whether and how temporal intervention can reduce misinformation propagation. To the best of our knowledge, ours is the first work to propose a comprehensive dynamic assessment framework for misinformation monitoring and reducing the propagation by deploying embargo strategies.

Conclusion and Future Work

We presented a dynamic system for real-time evaluation and grading of misinformation on Twitter, leveraging Temporal Graph Network and Recurrent Neural Networks (RNNs), along with an innovative temporal embargo strategy. The effectiveness of our framework was validated on five different social media misinformation datasets, highlighting its robustness and adaptability. In terms of our primary objectives, the model was successful in continuously evaluating and assigning misinformation scores to tweets, demonstrating strong performance at different time points. The model's performance compared favorably to other state-of-the-art misinformation detection models across diverse datasets and timeframes.

In future work, we will focus on further refining these techniques specifically towards multi-modal misinformation detection and inclusion of propaganda in misinformation propagation.

Broader Perspective, Ethics, and Competing Interests

Our framework for real-time evaluation of Twitter misinformation, despite its aim to mitigate harm, may raise ethical issues related to user privacy, censorship, and misuse potential. Improper deployment of the framework could lead to mass surveillance or censorship, infringing on privacy rights and freedom of speech. We advocate for responsible and transparent use, respecting individual privacy and freedom of expression, with clear communication about its deployment and the option for users to opt-out. We acknowledge the possibility for potential false positives and false negatives, and we suggest continuous research, development, and stakeholder feedback for system refinement. We declare no

competing interests. The research was conducted independently, using publicly available datasets, and the framework was developed for academic and public benefit aiming to better understand and fight misinformation online.

References

Agarwal, S.; Sawhney, R.; Thakkar, M.; Nakov, P.; Han, J.; and Derr, T. 2022. THINK: Temporal Hypergraph Hyperbolic Network. In *2022 IEEE International Conference on Data Mining (ICDM)*, 849–854. IEEE.

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.

Azzimonti, M.; and Fernandes, M. 2023. Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76: 102256.

Barnabò, G.; Siciliano, F.; Castillo, C.; Leonardi, S.; Nakov, P.; Da San Martino, G.; and Silvestri, F. 2022. FbMultiLingMisinfo: Challenging Large-Scale Multilingual Benchmark for Misinformation Detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Barnabò, G.; Siciliano, F.; Castillo, C.; Leonardi, S.; Nakov, P.; Da San Martino, G.; and Silvestri, F. 2023. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33: 100244.

Chen, E.; and Ferrara, E. 2022. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia. *arXiv preprint arXiv:2203.07488*.

Derczynski, L.; and Bontcheva, K. 2014. Pheme: Veracity in Digital Social Networks. In *UMAP workshops*.

Djenouri, Y.; Belhadi, A.; Srivastava, G.; and Lin, J. C.-W. 2023. Advanced Pattern-Mining System for Fake News Analysis. *IEEE Transactions on Computational Social Systems*.

Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *proceedings of the international AAAI conference on web and social media*, volume 8, 101–110.

Fung, Y. R.; Huang, K.-H.; Nakov, P.; and Ji, H. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4790–4791.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Ghosh, S.; and Mitra, P. 2023a. Catching Lies in the Act: A Framework for Early Misinformation Detection on Social Media. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, 1–12.

Ghosh, S.; and Mitra, P. 2023b. How Early Can We Detect? Detecting Misinformation on Social Media Using User Profiling and Network Characteristics. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 174–189. Springer.

Ghosh, S.; and Mitra, P. 2023c. Tweeted Fact vs Fiction: Identifying Vaccine Misinformation and Analyzing Dissent. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 136–143.

Hayawi, K.; Shahriar, S.; Serhani, M. A.; Taleb, I.; and Mathew, S. S. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public health*, 203: 23–30.

Hristakieva, K.; Cresci, S.; Da San Martino, G.; Conti, M.; and Nakov, P. 2022. The spread of propaganda by coordinated communities on social media. In *14th ACM Web Science Conference 2022*, 191–201.

Liu, Y.; and Wu, Y.-F. B. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3): 1–33.

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, 5(3): 337–348.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks.

Meel, P.; and Vishwakarma, D. K. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153: 112986.

Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; and Bronstein, M. M. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Nguyen, V.-H.; Sugiyama, K.; Nakov, P.; and Kan, M.-Y. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 1165–1174.

Park, C. Y.; Mendelsohn, J.; Field, A.; and Tsvetkov, Y. 2022. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5209–5235.

Patwa, P.; Sharma, S.; Pykl, S.; Guptha, V.; Kumari, G.; Akhtar, M. S.; Ekbal, A.; Das, A.; and Chakraborty, T. 2021. 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*, 21–29. Springer.

Przybyla, P. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 490–497.

Rossi, E.; Chamberlain, B.; Frasca, F.; Eynard, D.; Monti, F.; and Bronstein, M. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. 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.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36.

Tardelli, S.; Nizzoli, L.; Tesconi, M.; Conti, M.; Nakov, P.; Martino, G. D. S.; and Cresci, S. 2023. Temporal Dynamics of Coordinated Online Behavior: Stability, Archetypes, and Influence. *arXiv preprint arXiv:2301.06774*.

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. 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)*.

Vosoughi, S.; Mohsenvand, M. N.; and Roy, D. 2017. Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4): 1–36.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Wang, Y.; McKee, M.; Torbica, A.; and Stuckler, D. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240: 112552.

Wu, L.; and Liu, H. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, 637–645.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T.; et al. 2017. A Convolutional Approach for Misinformation Identification. In *IJCAI*, 3901–3907.

Yue, Z.; Zeng, H.; Kou, Z.; Shang, L.; and Wang, D. 2022. 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*, 2423–2433.

Zhou, X.; and Zafarani, R. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2.

Paper Checklist

1. Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, because the proposed system strictly uses aggregated data and advanced neural networks to assess misinformation without targeting individual users, thereby avoiding privacy violations, unfair profiling, and socio-economic biases while respecting cultural norms.
2. Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see *Problem Statement* subsection.
3. Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see *Experimental Analysis* section.
4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see *Data* subsection.
5. Did you describe the limitations of your work? Yes, see *Conclusion and Future Work* section.
6. Did you discuss any potential negative societal impacts of your work? NA, there is no negative impact from this work as it primarily aims to enhance the accuracy and reliability of information on social networks.
7. Did you discuss any potential misuse of your work? Yes, see the *Broader Perspective, Ethics, and Competing Interests*.
8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, the paper outlines precautions including the use of data anonymization to protect privacy, a dual model system for accurate information grading, and a robust retraining strategy to maintain model relevance and accuracy, ensuring responsible research practices and reproducibility of findings. Furthermore, the codebase has been released.
9. Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, see the *Broader Perspective, Ethics, and Competing Interests*.