GRACE: Graph-Based Attention for Coherent Explanation in Fake News Detection on Social Media

Orken Mamyrbayev[1], Zhanibek Turysbek*[2], Mariam Afzal[3], Marassulov Ussen Abdurakhimovich[4], Ybytayeva Galiya[5], Muhammad Abdullah[6], Riaz Ul Amin*[7]
Institute of Information and Computational Technologies, Almaty, Kazakhstan[1]
Kazakh National Research Technical University, Kazakhstan[2]
Riphah International University, Faisalabad[3]
International Kazakh-Turkish University named by Khoja Akhmet Yassawi[4]
Department of Technical and Natural Sciences at the International Educational Corporation[5]
School of Computing and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, Henan, China[6]
School of Computing and Information Technology, University of Okara and Edinburgh, Napier University, UK[7]

***Abstract*—Detecting fake news on social media is a critical challenge due to its rapid dissemination and potential societal impact. This paper addresses the problem in a realistic scenario where the original tweet and the sequence of users who retweeted it, excluding any comment section, are available. We propose a Graph-based Attention for Coherent Explanation (GRACE) to perform binary classification by determining if the original tweet is false and provide interpretable explanations by highlighting suspicious users and key evidential words. GRACE integrates user behaviour, tweet content, and retweet propagation dynamics through Graph Convolutional Networks (GCNs) and a dual co-attention mechanism. Extensive experiments conducted on Twitter15 and Twitter16 datasets demonstrate that GRACE outperforms baseline methods, achieving an accuracy improvement of 2.12% on Twitter15 and 1.83% on Twitter16 compared to GCAN. Additionally, GRACE provides meaningful and coherent explanations, making it an effective and interpretable solution for fake news detection on social platforms.**

***Keywords*—Graph neural network; dual attention; NLP; semantics; social network**

I. Introduction

Social media has become integral to everyday life, allowing individuals to share their thoughts, stay updated on current events, and interact with others [1]. These platforms facilitate the fast flow of information across vast networks, where user interactions and feedback shape public opinions and emotions on various topics [2]. This easy and low-cost communication fosters collective intelligence, spreading ideas widely and quickly. However, the very features that make social media so powerful also have significant drawbacks [3]. The speed and reach of these platforms make it easier for misinformation to spread, often without proper checks or regulation [4]. As a result, while social media can be a tool for empowerment and connection, it also amplifies the risk of misinformation, posing challenges to truth and trust in public discourse.

Fake news consists of false stories that are intentionally shared on social media platforms [5]. Its spread can mislead the public opinion, leading to political, economic, or psychological benefits for specific groups [6]. Fake news circulation manipulates opinions, distorts facts, and poses risks to society [7]. Research shows that people often struggle to differentiate between true and false news [8]. Interest in detecting fake news surged after the 2016 U.S. presidential election and COVID-19 vaccination drawing attention from researchers and social media platforms [9], [10], [11].

Detecting fake news is a complex task, primarily when relying solely on the content of news articles [12]. Traditional content-based methods often use features like n-grams and bag-of-words, applying supervised learning models such as random forests or SVM for binary classification [4], [13]. More advanced techniques in natural language processing (NLP) focus on extracting linguistic features like active/assertive verbs, subjectivity, and writing style [14]. Multi-modal approaches also integrate user-profiles and retweet propagation patterns [15]. However, these approaches face several challenges. Social media content, such as tweets, is usually short, leading to data sparsity, which makes it harder to detect fake news effectively [16]. Additionally, many models rely on user comments or retweets for evidence, but most users reshare stories without commenting, reducing the available data for analysis [17].

To address these challenges, researchers have begun focusing on propagation-based methods, which analyze the network of tweets and retweets to detect fake news [18], [19]. These methods are based on the idea that fake news spreads differently than true news. By studying the patterns of information diffusion, researchers can identify inconsistencies and spot fake stories [20]. However, many early approaches rely on static networks, assuming that the entire structure of information propagation is known before applying learning algorithms [21]. Social media networks are dynamic, with new users and content emerging over time, making static models less effective.

Recent research has shown that comprising temporary features, such as the timing of user interactions, can significantly improve fake news detection [22], [18]. For instance, in a temporal graph, the news propagation evolves, while a static graph only apprehends a snapshot of the network at one moment.

Fake and real news often show different temporal patterns, with fake news spreading more quickly or following distinct paths [23]. Regardless, treating these dynamic networks as if they were static limits the effectiveness of current models. To enhance detection, it's crucial to develop models that take into account the continuous changes in how users interact with each other. By doing so, these models can offer a more accurate and reliable way to tell the difference between real and fake news. These time-aware models would tap into the ever-evolving nature of social media. It makes them better equipped to detect misinformation in real-world situations.

This paper concentrates on detecting false content in the Twitter social media environment. The goal is to determine if a tweet is fake based solely on its brief text, the sequence of users who retweeted it, and their profiles. The detection process is approached with three key constraints: (a) analyzing the tweet's short text, (b) excluding user comments, and (c) not using network structures like social or diffusion networks. The model is designed to explain its predictions, meaning it should not only flag fake news but also show the reasoning behind the decision. Specifically, the model should highlight the doubtful users who helped to spread the fake news and identify the particular words or phrases from the source tweet that captured their attention.

Graph-based Attention for Coherent Explanation (GRACE) is proposed for fake news detection that integrates user behavior, tweet content, and retweet propagation dynamics. GRACE begins by feature extracting from user's Twitter profiles and encoding the original tweet's text using word embeddings [24]. A user interaction graph is constructed, and Graph Convolutional Networks (GCNs) [25] generate graph-aware representations of propagation dynamics. The relationship between the original tweet and how it spreads through retweets is identified by dual co-attention mechanism. It's helpful to highlight the users and keywords. By combining these features, GRACE offers an effective and easy-to-understand method for classifying fake news.

The key contributions of this paper are outlined as follows:

1) GRACE model is introduced to improve the understanding of user connection, retweet network, and their relationship with the short text of the source tweet.
2) Clear and meaningful explanations for the predictions are provided through the incorporation of a dual co-attention mechanism.
3) Comprehensive experiments are conducted on publicly available datasets that demonstrate the superior performance of GRACE as compared to baseline models.

This paper is structured as follows:

- Section II provides an overview of existing methods for fake news detection.
- Section III defines the problem and outlines the objectives addressed by the proposed model.
- Section IV details the experimental setup used in this study.
- Section V presents the evaluation metrics and results obtained.
- Finally, Section VI concludes the paper with a summary of findings and contributions.

II. Literature Review

Fake news, though not a new phenomenon, has acquired significant public awareness in recent years, primarily due to its widespread impact on society, politics, and media [26]. As the dissemination of false content continues to evolve, the literature on fake news detection has expanded rapidly, addressing the various challenges posed by this issue. Existing research can be broadly categorized into two main approaches: content-based and network-based methods. Content-based approaches focus on analyzing the textual data of news articles to identify linguistic, syntactic, and semantic features that distinguish fake news from legitimate news [27]. While, network-based methods focus on user's interactions and relationships within social media networks. They explore how news spreads across platforms and how user engagement patterns influence the dissemination of misinformation [21]. This section provides an overview of these two categories of fake news detection techniques and highlight the key developments, strengths, and limitations.

Content-based approaches focus on analyzing the textual data of news articles to evaluate their truthfulness. These methods are especially effective for long range dependencies, as they allow for a deep analysis of linguistic and semantic features to identify signs of misinformation [27]. One widely used technique is TF-IDF, which measures the importance of specific words within a news story [28]. Topic modeling helps to uncover the underlying themes in the content. It offers a structured and meaningful representation of the text [29]. Other linguistic features, such as PoS tags, assertive or factive verbs, and markers of subjectivity, are commonly used to detect subtle language patterns [24]. Further, techniques that assess writing consistency and social emotions are applied to highlight anomalies in news content [30]. The underlying assumption of these content-based methods is that fake news will exhibit detectable differences in linguistic structure, topic, or emotional tone compared to genuine news articles [31].

However, traditional content-based methods face several challenges in detecting fake news, mainly when relying on handcrafted linguistic cues [13]. These cues, such as lexical and syntactic features, are often limited in generalizability across different languages, topics, and domains. These techniques struggle to capture the complex semantic and contextual information embedded in modern news articles [3]. As news articles evolve in structure, content-based approaches that rely solely on traditional methods become less effective. As a result, researchers are increasingly turning to deep learning models to address these limitations [14]. The approaches like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Autoencoders [32] provide a solution by automatically learning hidden representations of text and capturing complex contextual patterns. These models eliminate the need for manually designed features and leverage word embeddings, such as word2vec, to enhance text representation and better identify patterns [33].

To make fake news detection more accurate, researchers have developed multi-modal models that combine different types of information, such as text and visuals, to improve their performance [15]. Visual elements like images and videos often play a significant role in how news is shared and perceived as credible. For instance, Bahad et al. introduced an RNN-based model that uses an attention mechanism to integrate text and visual information, allowing the system to focus on the most relevant features from both [34]. Similarly, Zhao et al. created a model that explores the relationship between text and visuals, which is especially effective in cases where misleading images are used to spread false claims [35].

To make detection systems more adaptable, researchers have also applied multi-task learning, enabling models to transfer knowledge across different types of content and better handle diverse contexts [36]. Since fake news evolves rapidly, new approaches like analyzing temporal patterns, adapting to specific domains, and leveraging weak supervision learning have been explored [37], [38], [10]. These innovations help detection systems stay scalable and flexible, allowing them to adapt to the ever-changing nature of misinformation. By combining these advancements, models are now better equipped to accurately and dynamically detect fake news in real-world scenarios.

Recent advancements in NLP have significantly improved the accuracy and reliability of content-based approaches. Transformer-based models [39], such as BERT (Bidirectional Encoder Representations from Transformers) [40] and GPT (Generative Pre-trained Transformer) [41], have revolutionized text representation and classification tasks by capturing contextual dependencies more effectively than traditional models. For instance, BERT has been fine-tuned for fake news detection by leveraging its bidirectional attention mechanism to understand subtle linguistic cues and context [42]. Similarly, GPT models have been employed to generate synthetic datasets for training effective classifiers and to analyze text for semantic coherence and logical consistency [35]. Hybrid models combining transformers with other neural architectures have also emerged. For example, a recent study integrated BERT with Graph Neural Networks (GNNs) to enhance performance by incorporating relationships between entities within news articles [43]. Other studies have focused on domain-specific adaptations of transformers, such as FakeBERT, which was trained on datasets tailored for misinformation detection [44]. These models not only outperform traditional approaches but also offer better generalization across domains and languages.

Network-based methods for detecting fake news focus on understanding how users interact with content on social media platforms [18]. Actions like commenting, retweeting, and following are critical to how information spreads and provide clues about the fake news propagation [19], [45]. By studying these patterns, researchers gained valuable insights into how to identify fake news and separate it from genuine content [46]. To model how news spreads, both homogeneous networks (where nodes and edges are similar) and heterogeneous networks (where they differ) are used [4].

Homogeneous networks, consisting of uniform nodes and edges, make it easier to study news spread within a unified structure [47]. A notable study by Chang et al. analysed the dispersal of false news on Twitter and found that false news spreads faster, further, and more broadly than true news [19]. This observation highlights the accelerated nature of fake news diffusion. To enhance fake news detection, Huang et al. proposed a tree-structured RNN model that integrates textual features and propagation structure features [48]. Similarly, Gong et al. introduced a bi-directional GCN to learn representations of content semantics and diffusion structures [43].

In difference, heterogeneous networks consist of multiple nodes and edges, offering a more detailed representation of the relationships within the news ecosystem [49]. Kang et al. proposed a model that encodes semantic information and the global structure of the diffusion graph, incorporating posts, comments, and user interactions [50]. Huang et al. developed a meta-path-based heterogeneous graph attention network to capture the semantic relationships among text content in news propagation [48]. Additionally, Xie et al. enhanced the robustness of graph-based fake news detectors by modelling entities through a heterogeneous information network and utilizing graph adversarial learning to ensure more distinctive structural features [51]. Another significant advancement in heterogeneous network models was introduced by Nguyen et al. by developing Factual News Graph (FANG). This framework leverages social structures and user engagement patterns for effective fake news detection [44].

Network-based methods for fake news detection effectively handle multimodal data by leveraging the unique strengths of graph structures to integrate and process text and visual features. Jin et al. [52] proposed a Hierarchical Propagation Network that constructs a hierarchical graph where nodes represent multimodal features such as text embeddings, visual features, and user interactions. These nodes are interconnected through propagation layers that explicitly model the interplay between modalities, enabling a seamless integration of multimodal signals. Wang et al. [53] introduced a Multimodal Fusion Graph where text and image features are processed through graph attention layers, dynamically weighing their contributions to detect fake news. This method effectively links modalities by treating textual and visual embeddings as interconnected nodes in a unified graph. Shu et al. [54] utilized a Graph-based Multimodal Embedding framework, which creates a graph where text and image metadata are nodes, and the relationships between them (e.g. co-occurrence in news items) are edges. The GME approach ensures joint feature learning by allowing intermodal dependencies to be explicitly modeled and updated during training. Zhou et al. [55] extended this concept by employing knowledge-enhanced graphs, incorporating external knowledge from textual and visual data into the graph structure. Here, knowledge graph embeddings serve as additional nodes, creating a richer multimodal representation that enhances the interplay between modalities for accurate fake news detection. These approaches demonstrate how network-based methods construct and leverage graphs to unify and effectively process both modalities.

While these network-based models have shown promise, much previous work has focused on static networks. However, our research takes a dynamic approach by analyzing social media news within temporal diffusion networks, reflecting the continuous evolution of news propagation.

Approaches focusing on user behavior analyze the charac-

teristics of individuals who interact with news content, such as retweeting or commenting on stories. Yang et al. proposed extracting account-based features like the user's gender, hometown, and follower count [56]. Shu et al. found that user profiles associated with fake news differ significantly from those linked to legitimate news [4]. Liu et al. introduced a joint Recurrent and Convolutional Neural Network (CRNN) model that captures more detailed profiles of users, particularly those who retweet news stories [57]. In contrast, session-based heterogeneous graph embedding methods [51] rely on user session data to learn user traits but are not directly applicable to fake news detection.

III. Materials and Methods

A. Preliminaries

Let $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_{|\mathcal{S}|}\}$ represent a collection of tweet stories, and $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{|\mathcal{U}|}\}$ be a group of individuals (users) in the social media network. Each tweet story $\sigma_i \in \mathcal{S}$ is a short-text document, denoted by $\sigma_i = \{w_{i1}, w_{i2}, \ldots, w_{il_i}\}$, where l_i is the number of words in the tweet story σ_i, and w_{ik} represents the k-th word in the story σ_i. Each user $\alpha_j \in \mathcal{A}$ is associated with a feature vector $\mathbf{v}_j \in \mathbb{R}^d$, where d is the dimensionality of the user's feature vector.

When a tweet story σ_i is shared, certain individuals will retweet it, forming a sequence of retweet records, referred to as the *retweet propagation path*. Let the propagation path of story σ_i be denoted by $\mathcal{P}_i = \{(\alpha_j, \mathbf{v}_j, t_j)\}$, where $(\alpha_j, \mathbf{v}_j, t_j)$ indicates that individual α_j with feature vector $\mathbf{v}_j$ retweeted story σ_i at time t_j. Here, $j = 1, 2, \ldots, K$, with $K = |\mathcal{P}_i|$ being the total number of retweets. The set of individuals who retweet story σ_i is denoted as $\mathcal{A}_i \subseteq \mathcal{A}$.

Within the propagation path $\mathcal{P}_i$, the individual α_1 is the original poster of story σ_i at time τ. For all subsequent individuals $j > 1$, individual α_j retweets the story at time τ_j, where $\tau_j > \tau_1$.

The tweet story σ_i is labeled with a binary value $\kappa_i \in \{0, 1\}$ to indicate its truthfulness, where:

$$\kappa_i = \begin{cases} 0 & \text{if the news } \sigma_i \text{ is true,} \\ 1 & \text{if the news } \sigma_i \text{ is fake.} \end{cases}$$

Given a tweet story σ_i and its corresponding propagation path $\mathcal{P}_i$ (which includes individuals α_j who retweet the news and their associated feature vectors $\mathbf{v}_j$), the goal is to predict the authenticity κ_i of the story σ_i, a binary classification task.

The model should highlight a subset of individuals $\alpha_j \in \mathcal{A}_i$ who retweeted σ_i and a subset of words $w_{ik} \in \sigma_i$ that help to explain why σ_i is classified as either true or fake. This interpretability aspect is essential for understanding the reasoning behind the model's prediction.

B. Proposed Model

The GRACE model is developed to tackle the challenge of detecting fake news in social media networks by combining tweet content, user behavior, and the propagation dynamics of retweets. As depicted in Fig. 1, This model consists of several components including user characteristics extraction, news story encoding, user propagation representation, dual co-attention mechanisms, and the final prediction layer. Each component is meticulously crafted to improve the model's ability to predict the truthfulness of a tweet while also providing interpretability by highlighting the users and words contributing to the classification.

The user characteristics extraction component involves creating a feature vector $\mathbf{x}_j \in \mathbb{R}^v$ for each user $u_j \in \mathcal{A}$, where v is the number of features. These features are derived from various aspects of a user's behavior, such as the number of followers, the number of retweets, the time difference between the tweet and retweet, and other profile-related information. This vector allows us to quantify how a user engages with content on social media, which is crucial for identifying fake news spreaders.

The source tweet σ_i is represented as a sequence of words, denoted by $\sigma_i = \{w_{i1}, w_{i2}, \ldots, w_{il_i}\}$, where l_i is the number of words in tweet σ_i. We use a word-level encoder to represent this tweet. Each word w_{ik} in the tweet is initially encoded as a one-hot vector. A FC layer is applied to generate the word embeddings for each tweet, and the resulting embeddings are stored in a matrix $\mathbf{V} = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{d \times m}$, where m is the length of the padded tweet and d is the dimensionality of the word embeddings.

To model the interactions among users who retweet the source tweet σ_i, we construct a graph $\mathcal{H}_i = (\mathcal{V}_i, \mathcal{F}_i)$, where $\mathcal{V}_i$ represents the set of users who retweeted σ_i. The edges $\mathcal{F}_i$ represent the interactions between users. Since the true interactions between users are unknown, we assume that the graph is fully connected. This implies that for every edge $e_{\alpha\beta} \in \mathcal{F}_i$, where $u_\alpha, u_\beta \in \mathcal{V}_i$ and $u_\alpha \neq u_\beta$, the number of edges is given by:

$$|\mathcal{F}_i| = \frac{n \cdot (n-1)}{2} \tag{1}$$

where $n = |\mathcal{V}_i|$ is the number of users who retweeted σ_i.

To incorporate user features into the graph, we assign a weight w_{ab} to each edge $e_{ab} \in \mathcal{F}_i$, which is derived from the cosine similarity between the feature vectors $\mathbf{u}_a$ and $\mathbf{u}_b$. The weight is calculated as:

$$w_{ab} = \frac{\mathbf{u}_a \cdot \mathbf{u}_b}{\|\mathbf{u}_a\| \|\mathbf{u}_b\|} \tag{2}$$

We use the adjacency matrix $\mathbf{W} = [w_{ab}] \in \mathbb{R}^{n \times n}$ to represent the weights between any pair of nodes v_a and v_b in the graph $\mathcal{F}_i$.

C. Graph Convolutional Network (GCN)

A Graph Convolutional Network (GCN) [25] is applied to propagate information through the graph $\mathcal{F}_i$. A GCN layer performs a convolution operation on the graph, updating node embeddings by aggregating information from their neighbors. For the graph $\mathcal{F}_i$, with adjacency matrix $\mathbf{\Pi}$ and feature matrix $\mathbf{\Lambda}$ representing user attributes in $\mathcal{F}_i$, the updated g-dimensional node feature matrix $\mathbf{\Omega}^{(l+1)} \in \mathbb{R}^{n \times g}$ at layer $l+1$ is calculated as:

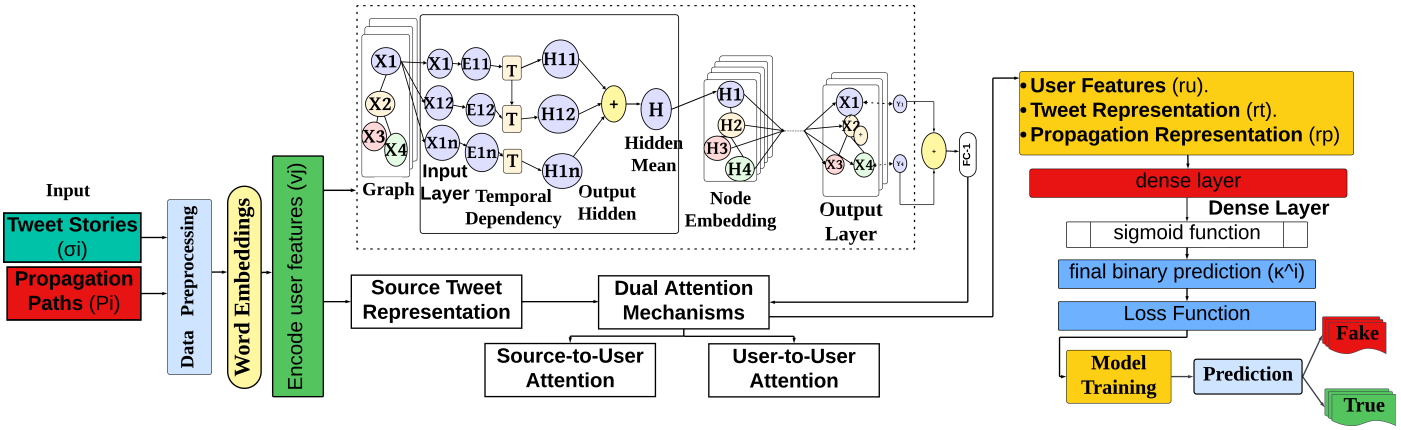

Fig. 1. The proposed model architecture diagram.

$$\mathbf{\Omega}^{(l+1)} = \phi\left(\tilde{\mathbf{\Pi}}\mathbf{\Omega}^{(l)}\mathbf{\Gamma}_l\right) \tag{3}$$

Here, $\tilde{\mathbf{\Pi}} = \mathbf{\Sigma}^{-1/2}\mathbf{\Pi}\mathbf{\Sigma}^{-1/2}$ represents the normalized adjacency matrix, $\mathbf{\Sigma}$ is the diagonal degree matrix, and ϕ is a non-linear activation function. This process is repeated iteratively over multiple layers, allowing information to propagate and be refined across the graph.

D. Co-attention Mechanisms

The correlation between the source tweet and users' interactions, including retweets, is captured using a dual co-attention mechanism. This mechanism simultaneously models the relationship between the source tweet and its retweets, as well as interactions between users within the propagation graph.

1) Tweet-Retweet Correlation: The first attention mechanism focuses on the relationship between the source tweet ($\mathbf{Q}_\sigma$) and the embeddings of retweets ($\mathbf{Q}_u$), which are derived from user propagation embeddings. The attention weights, representing the correlation between the content of the tweet and retweets, are computed as:

$$\mathbf{A}_\sigma = \text{softmax}(\mathbf{Q}_\sigma^T\mathbf{Q}_u) \tag{4}$$

These weights capture how strongly each retweet relates to the source tweet, refining both the source tweet and retweet representations for improved feature learning.

2) User-User Correlation: The second attention mechanism captures interactions between users by modeling the relationship between user embeddings across the propagation graph. This is achieved through:

$$\mathbf{A}_u = \text{softmax}(\mathbf{Q}_u^T\mathbf{Q}_u) \tag{5}$$

Here, the attention weights emphasize connections between users who share similar propagation behaviors, enabling the model to better understand the dynamics of retweet propagation.

By combining these two mechanisms, the model learns attention-driven representations that reflect the content and propagation dynamics of the source tweet and retweets. These refined representations are used as inputs for the final prediction stage.

E. Final Prediction

The final prediction $\hat{\kappa}_i$ is obtained by combining the learned user features, source tweet embeddings, and propagation representations. The concatenated vector is passed through a fully connected layer with a sigmoid activation, producing a probability between 0 and 1 that represents the likelihood of the source tweet σ_i being fake. This process can be expressed as:

$$\hat{\kappa}_i = \sigma\left(\mathbf{W}_f \cdot [\mathbf{r}_u, \mathbf{r}_t, \mathbf{r}_p] + b_f\right) \tag{6}$$

where $\mathbf{r}_u$ is the learned representation of user characteristics, $\mathbf{r}_t$ is the learned embedding of the source tweet, and $\mathbf{r}_p$ is the learned propagation representation of the users. The vector $[\mathbf{r}_u, \mathbf{r}_t, \mathbf{r}_p]$ is the concatenation of these representations, $\mathbf{W}_f$ is the weight matrix, and b_f is the bias term. The sigmoid function $\sigma(\cdot)$ is applied to ensure that the output is a probability between 0 and 1.

F. Loss Function

The binary cross-entropy loss function is used for model training. It measures the difference between the predicted probability $\hat{\kappa}_i$ and the true label κ_i:

$$\mathcal{L}(\hat{\kappa}_i, \kappa_i) = -\kappa_i \log(\hat{\kappa}_i) - (1 - \kappa_i)\log(1 - \hat{\kappa}_i) \tag{7}$$

The loss function is minimized using the Adam optimizer, ensuring that the model's parameters are updated to reduce the classification error over time. The optimization process helps the model improve its predictions by adjusting weights, thereby minimizing the loss and enhancing the performance of the fake news detection system.

Algorithm 1 GRACE (Graph-based Attention for Coherent Explanation)

Input: Tweet stories $\mathcal{S} = \{\sigma_1, \ldots, \sigma_{|\mathcal{S}|}\}$, user profiles $\mathcal{A}$, propagation paths $\mathcal{P}_i$, truthfulness labels κ_i.
Output: Predicted labels $\hat{\kappa}_i$ and explanation (highlighted users α_j and words w_{ik}).

Initialize: Pre-trained word embeddings, user feature vectors $\mathbf{v}_j$, graph adjacency matrices $\mathbf{A}$, and model parameters.
for each tweet $\sigma_i \in \mathcal{S}$ **do**
  Encode tweet σ_i as word embeddings $\mathbf{V} \in \mathbb{R}^{d \times m}$.
  Extract user feature vectors $\mathbf{v}_j \in \mathbb{R}^d$ for users in $\mathcal{P}_i$.
  Construct a graph $\mathcal{H}_i = (\mathcal{V}_i, \mathcal{F}_i)$:
  for each pair of users $(\alpha_\alpha, \alpha_\beta) \in \mathcal{V}_i$ **do**
    if users are connected **then**
      Compute edge weight:

$$\omega_{\alpha\beta} = \frac{\mathbf{x}_\alpha \cdot \mathbf{x}_\beta}{\|\mathbf{x}_\alpha\| \|\mathbf{x}_\beta\|}.$$

    end if
  end for
  Apply GCN to propagate embeddings over $\mathcal{H}_i$:

$$\mathbf{H}^{(l+1)} = \rho\left(\mathbf{A}\tilde{\mathbf{H}}^{(l)}\mathbf{W}_l\right),$$

  where ρ is a non-linear activation function.
  Compute dual co-attention:
    Source-to-user attention:

$$\mathbf{A}_\sigma = \text{softmax}(\mathbf{Q}_\sigma^T \mathbf{Q}_u).$$

    User-to-user attention:

$$\mathbf{A}_u = \text{softmax}(\mathbf{Q}_u^T \mathbf{Q}_u).$$

  Concatenate learned embeddings $\mathbf{r}_u$, $\mathbf{r}_t$, and $\mathbf{r}_p$:

$$\mathbf{r} = [\mathbf{r}_u, \mathbf{r}_t, \mathbf{r}_p].$$

  Predict truthfulness:

$$\hat{\kappa}_i = \sigma\left(\mathbf{W}_f \cdot \mathbf{r} + b_f\right).$$

  Highlight key users α_j and words w_{ik} based on $\mathbf{A}_\sigma$ and $\mathbf{A}_u$.
end for
Optimize model parameters by minimizing the binary cross-entropy loss:

$$\mathcal{L}(\hat{\kappa}_i, \kappa_i) = -\kappa_i \log(\hat{\kappa}_i) - (1 - \kappa_i) \log(1 - \hat{\kappa}_i).$$

IV. Experimental Setup

The GRACE model is implemented using the PyTorch framework. The tweet text is represented using pre-trained word embeddings. Each word in the tweet is mapped to its corresponding vector representation. These embeddings help transform the raw text into a meaningful numerical format suitable for further processing. GCN layers capture the interactions among users who retweet the source tweet. The graph represents users as nodes, and the interactions between users (such as retweeting) form the edges. Each node's feature vector is updated based on its neighbours, allowing the model to learn user-specific representations in the context of retweet propagation. The number of GCN layers is set to 3, with each layer processing information from the node's direct and indirect neighbours. These features are concatenated after obtaining the embeddings from the tweet content, user characteristics, and user propagation representations. The concatenated vector is passed through fully connected (dense) layers to make the final classification decision. The hidden layers in the fully connected section use ReLU activation, while the output layer employs a sigmoid activation function to predict the probability of a fake tweet.

A. Hyperparameters

The proposed model is designed with several key hyperparameters that allow for efficient and effective training as described in Table I. A learning rate of 0.001 was selected after a grid search of several potential values. This choice balances convergence speed and stability, ensuring that the model trains effectively without overshooting the optimal solution. The batch size was set to 64, a commonly used value in graph-based models like GCNs. A batch size of this allows for efficient computation and good convergence properties while maintaining memory efficiency during training.

The model architecture is developed with three GCN layers, which strike a balance between capturing the interactions within the retweet network and avoiding overfitting caused by excessive depth. Each GCN layer contains 128 hidden units, which are sufficient to learn rich user interaction features without making the model too large and prone to overfitting. A dropout rate of 0.3 is applied to mitigate overfitting, meaning

30% of the neurons are randomly dropped during training, helping the model avoid reliance on specific features.

Following the GCN layers, fully connected (FC) layers were added with 256 hidden units to combine and process features from tweet content, user characteristics, and propagation patterns. To reduce overfitting in these layers, a higher dropout rate of 0.5 was applied, randomly dropping 50% of the neurons during training to improve generalization.

ReLU activation is used throughout the hidden layers to introduce non-linearity, enabling the model to learn more complex patterns and decision boundaries effectively. The output layer uses the sigmoid activation function, which maps the final output to a probability between 0 and 1. This value is interpreted as the likelihood that a given tweet is fake. The Adam optimizer was chosen for optimisation, known for its efficiency in handling sparse gradients and large datasets. The binary cross-entropy loss function was used as the loss criterion, as it is well-suited for binary classification tasks like fake news detection. The model was trained for 20 epochs, which is sufficient for convergence without overfitting. These hyperparameters were carefully chosen to ensure the model performs well on the fake news detection task, balancing model complexity, training efficiency, and the ability to generalize to unseen data.

TABLE I. HYPERPARAMETERS FOR GCAN MODEL

Hyperparameter	Value
Learning Rate	0.001
Batch Size	64
GCN's Layers	3
Hidden Units	128
Dropout Rate	0.3
Hidden Units in FC Layers	256
Dropout Rate (FC layers)	0.5
Activation Function (Hidden)	ReLU
Activation Function (Output)	Sigmoid
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Epochs	20

B. Datasets

This study utilizes two widely used datasets, Twitter15 and Twitter16, compiled by Ma et al. [58], which are recognized benchmarks in the field of fake news detection. These datasets provide a comprehensive basis for evaluating propagation-based modeling approaches, as they include tweets along with the corresponding sequences of retweeting users, which are essential for capturing propagation dynamics.

The Twitter15 dataset includes 1,490 claims, while Twitter16 contains 818 claims. Both datasets are annotated with four ground truth veracity labels: True News (T), Fake News (F), Non-Fake News (NF), and Unverified News (U). For our binary classification experiments, we focus only on True News (T) and Fake News (F) labels, aligning with the scope of our study.

These datasets are particularly suitable for evaluating our proposed model as they include rich propagation structures that allow us to assess the effectiveness of graph-based approaches. Additionally, they represent real-world social media interactions, offering realistic challenges and scenarios for fake news detection.

To enrich the data, we collected user profile information using the Twitter API, as the original datasets do not include user profiles. This additional data allows us to incorporate user-specific features, such as activity patterns and engagement metrics, which are crucial for analyzing user behavior in the context of fake news propagation.

The datasets are divided into three parts: 70% for training, 15% for testing, and 15% for validation. This ensures a balanced and rigorous evaluation of the model. Table II and Fig. 2 provide a summary of the key statistics and label distributions, illustrating the diversity and scale of the datasets.

These choices ensure that our approach is validated against reliable, well-established benchmarks, offering a fair comparison with prior works and a robust demonstration of the proposed model's effectiveness.

TABLE II. DATASET STATISTICS

Feature	Twitter15	Twitter16
Total Claims	1,490	818
True News (T)	370	205
Fake News (F)	374	204
Non-Fake News (NF)	374	203
Unverified News (U)	372	206
Total Postings	331,612	204,820
Users	190,868	115,036
Avg. Retweets per Story	292.19	308.70
Avg. Words per Source	13.25	12.81
# Total Nodes	912,638	501,032
# Total Edges	697,523	382,936

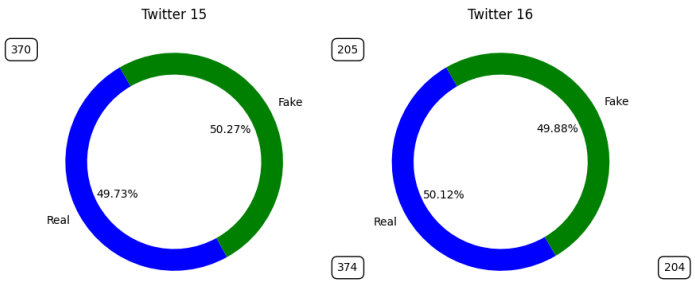

Fig. 2. Label distribution for Twitter15 and Twitter16 datasets.

C. Evaluation Metrics

To assess the proposed model's performance for fake news detection, we use several key metrics that provide insights into its effectiveness. These metrics include Accuracy, Precision, Recall, F1 Score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

Accuracy is the most straightforward metric, measuring the overall correctness of the model across all predictions. It is the ratio of correct predictions to the total number of predictions. **Precision** evaluates the proportion of positive predictions (predicted fake news) that are actually correct. A high Precision indicates that the model is accurate when it predicts fake news. **Recall** focuses on the model's ability to capture all actual positive instances (actual fake news). It is the ratio of true positives to the sum of true positives and false negatives. A high Recall means that the model successfully identifies most of the true fake news instances. **F1 Score** is the

harmonic mean of Precision and Recall. It provides a balanced measure and offers a single number that evaluates the model's performance in relevance and completeness.

V. Results

Results are reported in Table III. The GRACE model demonstrated notable accuracy and F1 score improvements across the Twitter15 and Twitter16 datasets. The F1 score increased by 2.42% from baseline models, reaches at 84.50, while accuracy improved by 2.08%, achieving 89.50 on the Twitter15 dataset. On the Twitter16 dataset, the F1 score saw a 2.07% improvement, reaching 77.50, and accuracy increased by 1.83%, reaching 92.50. On average, the GRACE model showed a 2.24% improvement in F1 score and a 1.95% improvement in accuracy across both datasets. These results reflect the model's consistent enhancement in both key performance metrics. The GRACE model's improvements indicate its strong capacity to achieve higher classification precision and accuracy than baseline models, showcasing its ability to generalize well across different datasets. The bigger improvements in Twitter15 suggest the model's adaptability in handling diverse data characteristics, while its solid performance in Twitter16 further emphasizes its robustness in real-world, noisy data scenarios.

A. Baseline Models

The proposed model is compared with several baseline methods on the Twitter15 and Twitter16 datasets, as shown in Table III. The GCAN (Graph-aware Co-attention Network) predicts fake news by considering the orignal tweet and its propagation, with a variant, GCAN-G, which excludes the graph convolution component to evaluate the effectiveness of graph-aware representations [21]. SVM-TS combines a Support Vector Machine with heuristic rules and a time-series structure to classify posts as fake or real, leveraging handcrafted features. While effective initially, deep learning models now outperform traditional approaches due to superior feature extraction capabilities [59]. DTC is a rumor detection method that uses a Decision Tree classifier and leverages various handcrafted features to evaluate information credibility [60]. It focuses on extracting and analyzing features related to content, user behavior, and network interactions to improve detection accuracy. CRNN, a deep residual network, integrates four cascading graph convolutional networks to capture long-range dependencies and nonlinear features effectively [61]. RFC is a ranking method based on Random Forest that refines and elaborates the inquiry phrases within posts. By leveraging this approach, it aims to enhance the analysis and prioritization of relevant information [62]. dEFEND represents tweet contents and interaction graphs in a latent space, capturing multi-level features of fake news through claim-aware and inference-based attention mechanisms [63]. The CSI model stands out as an advanced fake news detection model that integrates both article content and the collective behaviour of users propagating fake news [64]. This model uses LSTM to capture sequential dependencies and computes user-specific scores to evaluate the likelihood of a tweet being fake. The tCNN model introduces a modified Convolutional Neural Network (CNN) to learn local variations in user profile sequences, combining them with features from the source tweet [65]. This approach effectively captures intricate variations in user behaviour. The CRNN merges CNN and RNN to learn local and global user profile variations [66]. This hybrid technique enables the model to capture temporal and spatial dependencies in retweet propagation. mGRU is a modified gated recurrent unit (GRU) model designed for rumor detection. It captures temporal patterns by leveraging retweet user profiles in combination with the source tweet's features [58].

The confusion matrices are presented in Fig. 3. These metrics show the model's performance in classifying news across multiple categories. For the Twitter15 dataset, the model correctly identifies True News, with 109 instances accurately classified, while only four are misclassified as False News and seven as Unverified News. This indicates the model's proficiency in distinguishing authentic information. The model successfully classifies 36 instances for False News, with minimal misclassifications (6 as True News and one as Unverified News). In Twitter16 dataset, The model accurately identifies True News, classifying 56 instances correctly, while only three are misclassified as False News and four as Unverified News. It also performs well in detecting False News, correctly classifying 23 instances, with just a few misclassifications (2 instances each into True News and Unverified News). The

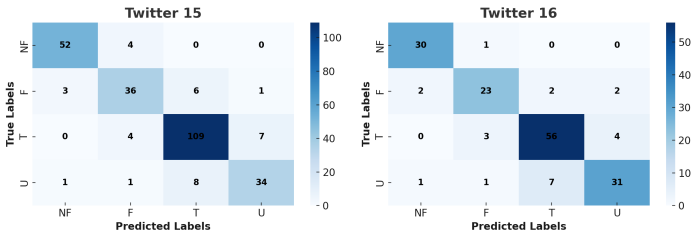

Fig. 3. Confusion matrices for Twitter15, Twitter16 on test dataset.

differences in classification accuracy across these two datasets can be attributed to the varying complexity of the classification tasks. While both datasets include multiple categories, the Twitter15 and Twitter16 datasets introduce the additional challenge of distinguishing Unverified News from True and False News, resulting in a higher degree of misclassification, especially between False News and Unverified News, which share similar content characteristics. These results underscore the model's adaptability in handling both binary and multi-class classification challenges, demonstrating its effectiveness across diverse datasets.

The performance of the proposed model is evaluated in terms of accuracy in Fig. 4 for early detection. It is analyzed under varying conditions by altering the number of observed retweet users per source story, ranging from 10 to 50. The results demonstrate that GRACE consistently and significantly outperforms all competing methods across all scenarios. Despite as few as 10 observed retweeters, GRACE achieves an impressive 82% accuracy on Twitter16, underscoring its robustness and reliability. These findings highlight GRACE's capability to deliver accurate and early detection of fake news dissemination, which is critical for effectively combating misinformation and mitigating its impact.

We assess the effectiveness of proposed approach and baseline models for early stage fake news detection. Early

TABLE III. COMPARISON OF PROPOSED MODEL WITH BASELINE AND STATE-OF-THE-ART MODELS ON TWITTER15 AND TWITTER16 DATASETS

Method	Twitter15				Twitter16			
	F1	Recall	Precision	Accuracy	F1	Recall	Precision	Accuracy
DTC	49.48	48.06	49.63	49.49	56.16	53.69	57.53	56.12
SVM-TS	51.90	51.86	51.95	51.95	69.15	69.10	69.28	69.32
mGRU	51.04	51.48	51.45	55.47	55.63	56.18	56.03	66.12
GCAN-G	79.38	79.90	79.59	86.36	67.54	68.02	67.85	79.39
RFC	46.42	53.02	57.18	53.85	62.75	65.87	73.15	66.20
tCNN	51.40	52.06	51.99	58.81	62.00	62.62	62.48	73.74
CRNN	52.49	53.05	52.96	59.19	63.67	64.33	64.19	75.76
CSI	71.74	68.67	69.91	69.87	63.04	63.09	63.21	66.12
GCAN	82.50	82.95	82.57	87.67	75.93	76.32	75.94	90.84
dEFEND	65.41	66.11	65.84	73.83	63.11	63.84	63.65	70.16
GRACE	84.17	84.95	84.74	89.53	77.51	78.09	77.73	79.11

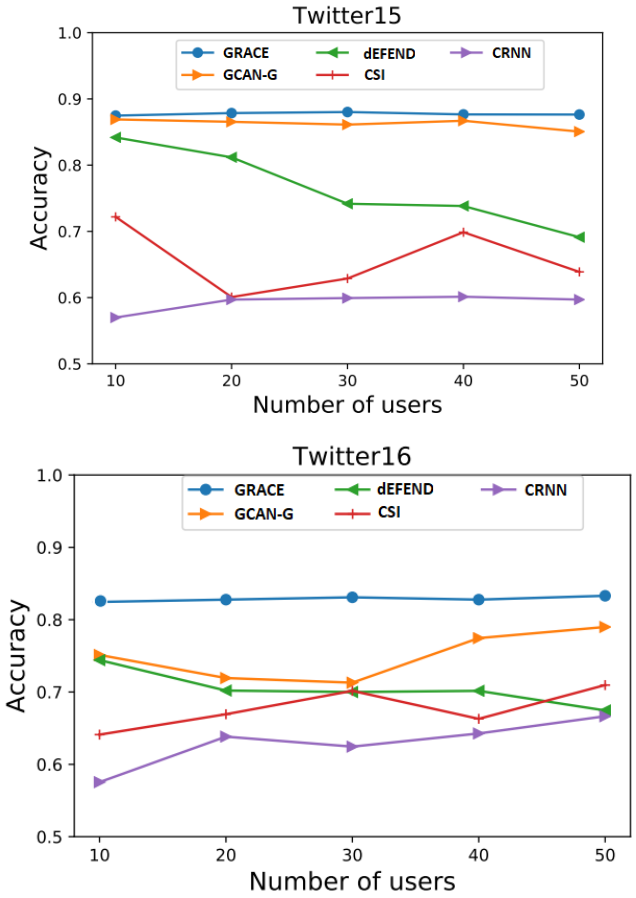

Fig. 4. Accuracy over retweet users on Twitter15 and Twitter16 datasets.

identification of fake news is essential to curbing its spread and minimizing its harmful societal impacts. For this evaluation, we use a specific tweet's propagation time or initial release time within a news event as the detection deadline. Tweets posted beyond this deadline are excluded from consideration. To compare the performance of various detection methods, we vary the detection time points within a specific range and analyze their performance.

Fig. 5 presents the accuracy of all methods at different time intervals across three datasets. The results indicate that GRACE consistently performs better than baseline models in early-stage fake news detection. Across all datasets, the accuracy of all methods improves rapidly during the early stages of information diffusion. Notably, our model exhibits a distinct performance advantage as the propagation progresses, demonstrating its ability to sustain high accuracy over time and effectively adapt to the dynamics of fake news dissemination.

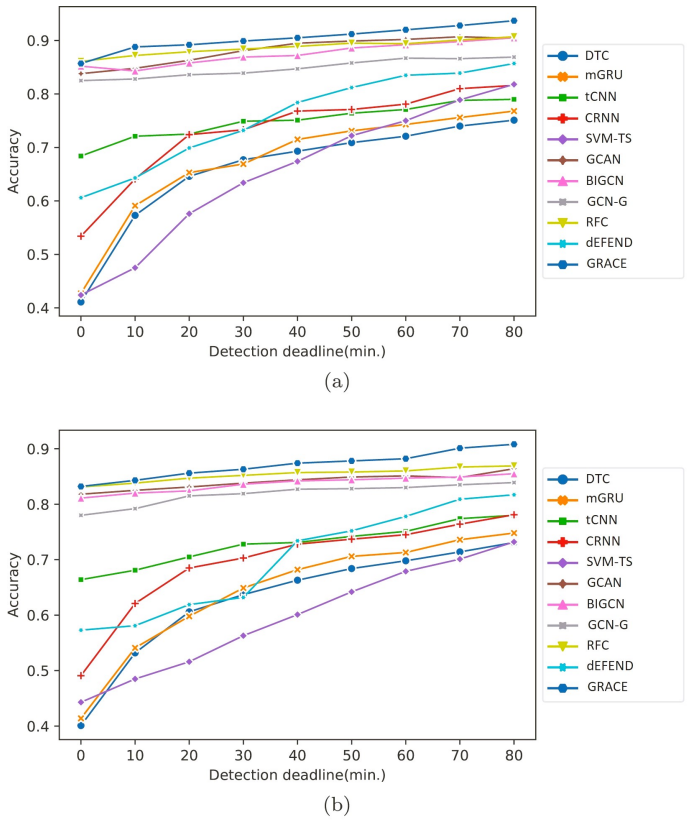

Fig. 5. (a) Early detection of fake news on Twitter15; (b) early detection of fake news on Twitter16.

The source-propagation co-attention mechanism embedded in our proposed model offers meaningful insights into identifying the characteristics of suspicious users and the linguistic cues they emphasize during the spread of information. As Fig. 6 demonstrates, the model highlights several distinct traits commonly associated with suspicious retweeters. These

include unverified accounts, newly created profiles with shorter account lifespans, minimal user descriptions, and shorter graph path lengths connecting them to the source tweet's author.

Moreover, the analysis reveals that these users focus disproportionately on specific words, such as "breaking" and "pipeline," often used in sensationalized or misleading content. By leveraging these observations, the model enhances its ability to detect fake news and provides interpretability by uncovering suspicious accounts' behavioural patterns and language preferences. Such explainability is crucial for understanding the underlying mechanisms of fake news dissemination and developing strategies to mitigate its spread effectively.

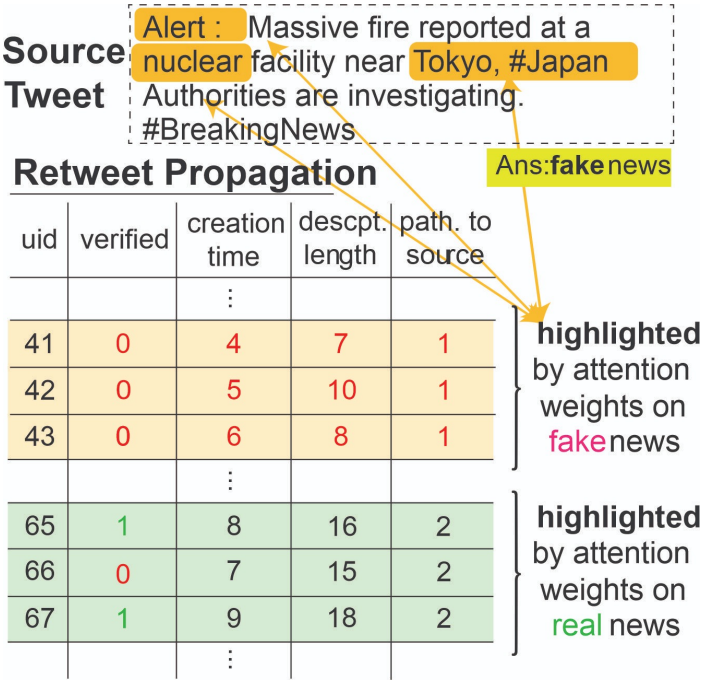

Fig. 6. Key evidential words identified by the GRACE model in the source tweet (top) and suspicious users flagged during the retweet propagation process (bottom). Each column corresponds to a specific user characteristic, providing deeper insights into user behaviours. For simplicity, only a select number of user characteristics are presented.

B. Ablation Study

The ablation study is conducted in Table IV. It highlights the significance of each component in the proposed model. Removing the dual co-attention mechanism ("-A") leads to a noticeable drop in performance, which underscores its role in linking tweet content with user interactions and propagation dynamics. Excluding the graph-aware representation ("-G") also affects the model's accuracy, as it captures the structural relationships between users in the retweet network. Similarly, removing the user characteristics module ("-U"), which captures behavioural patterns like account age and retweet frequency, significantly reduces the model's ability to detect suspicious users. The absence of source tweet embeddings ("-S") results in a substantial decline, showing the importance of semantic content in distinguishing fake news. The most severe performance degradation occurs when the source tweet embeddings and dual co-attention mechanism are removed ("-S-A"), demonstrating that integrating content-based and interaction-based features is crucial for achieving high accuracy. These results confirm that each component contributes meaningfully to the overall effectiveness of the GRACE model.

C. Discussion

The findings from our study highlight the robustness and interpretability of the GRACE model in detecting fake news across various datasets and scenarios. By leveraging multiple data modalities [58], such as user characteristics, tweet content, and propagation dynamics, GRACE achieves superior performance compared to existing baseline models. This discussion contextualizes these results, explores their implications, and addresses the model's broader applicability and potential limitations.

One of the most significant insights from our work is the importance of integrating user behavior and propagation dynamics into fake news detection. Traditional models often focus solely on tweet content, neglecting the behavioral and relational cues that can provide essential context [36], [54]. GRACE fills this gap by incorporating graph-aware propagation modeling and user embedding representation, which allows it to capture the underlying social dynamics in retweet propagation. This synergy between components is evident from our ablation study, where removing key elements, such as the dual co-attention mechanism or graph-based user representations, led to noticeable drops in performance.

The results also reveal GRACE's adaptability in both the early and advanced stages of fake news propagation. For instance, GRACE's ability to maintain high accuracy with limited early-stage data (e.g. as few as 10 retweeters) underscores its potential for real-time applications. This early detection capability is crucial for mitigating the spread of misinformation, as even a small delay in identification can result in widespread dissemination and societal harm.

Another strength of GRACE lies in its explainability. The co-attention mechanism enables the model to highlight the specific words in tweets and user behaviors contributing to its predictions. For instance, the model identified linguistic patterns, such as emotionally charged words like "breaking", and behavioural traits, including unverified accounts and recently created profiles, as key indicators of suspicious activity. This interpretability is vital for building trust with end-users, particularly in applications where automated decisions must be transparent and defensible.

Understanding the characteristics of suspicious users and the propagation patterns of fake news provides actionable insights for social media platforms and policymakers. By identifying high-risk accounts and content early, GRACE can assist in designing targeted interventions, such as flagging or debunking misleading posts before they gain significant traction.

D. Limitation and Future Work

While GRACE demonstrates strong performance and interpretability, it is not without limitations. One of the primary challenges is the reliance on user interaction data to build propagation graphs. The model's effectiveness could

TABLE IV. ABLATION STUDY RESULTS OF GRACE ON TWITTER15 AND TWITTER16 DATASETS

Method	Twitter15				Twitter16			
	F1	Rec	Precision	Accuracy	F1	Recall	Precision	Accuracy
Full Model	84.17	84.95	84.74	89.53	77.51	78.09	77.73	79.11
-A	81.45	82.13	80.97	87.12	74.89	75.12	74.45	76.45
-G	82.03	82.67	81.45	87.67	75.43	75.87	75.01	77.02
-U	80.12	80.89	79.68	85.34	73.25	73.98	72.87	74.34
-S	78.65	79.02	78.30	83.21	72.11	72.56	71.43	72.89
-S-A	75.34	75.89	74.12	80.78	70.34	70.92	69.87	71.21

be reduced if user data is incomplete or anonymized due to privacy concerns. Additionally, while GRACE assumes a fully connected graph without explicit user relationships, this assumption may not always reflect real-world interactions, potentially leading to inaccuracies in propagation modelling. Future work could explore incorporating more advanced graph representation techniques, such as dynamic graph neural networks, to better model evolving user interactions over time to enhance GRACE further. Additionally, leveraging external knowledge bases or fact-checking databases could improve the model's ability to validate content credibility, particularly for previously unseen news stories. Finally, expanding GRACE to handle multilingual content and adapting it to different cultural contexts would increase its applicability on a global scale.

VI. CONCLUSION

In this study, we introduced Graph- based Attention for Coherent Explanation (GRACE) approach for detecting fake news on social media platforms. GRACE addresses the complex and dynamic nature of misinformation by leveraging tweet content, user behaviour, and retweet propagation dynamics, making it capable of identifying fake news with high accuracy and interpretability. Unlike traditional methods, GRACE operates in a more realistic and challenging setting by focusing on short-text tweets and their retweeter sequences, closely aligning with the real-world propagation of misinformation. The evaluation results underscore GRACE's robustness and effectiveness, demonstrating its ability to deliver accurate predictions while maintaining explainability through its co-attention mechanism. Notably, GRACE excels in early-stage detection, achieving satisfying performance even with limited propagation data. This early detection capability is critical for minimizing the spread of misinformation and reducing its societal impact.

Beyond fake news detection, GRACE has broader applications for other short length text classification tasks in social media, such as sentiment analysis and tweet popularity prediction. Its flexible and modular design makes it a promising candidate for addressing various social media challenges. Future work will enhance the model's generalization capabilities to accommodate different platforms and contexts.

REFERENCES

[1] L. Humphreys, *The qualified self: Social media and the accounting of everyday life*. MIT press, 2018.

[2] S. K. Rathi, B. Keswani, R. K. Saxena, S. K. Kapoor, S. Gupta, and R. Rawat, *Online Social Networks in Business Frameworks*. John Wiley & Sons, 2024.

[3] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, "Evidence-aware Fake News Detection with Graph Neural Networks," *arXiv e-prints*, p. arXiv:2201.06885, Jan. 2022.

[4] 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*, vol. 8, no. 3, pp. 171–188, 2020.

[5] B. D. Horne, D. Nevo, and S. L. Smith, "Ethical and safety considerations in automated fake news detection," *Behaviour & Information Technology*, pp. 1–22, 2023.

[6] F. Odum, "Covid conspiracy narratives: Dissecting the origins of misinformation in digital space," 2021.

[7] M. Farokhian, V. Rafe, and H. Veisi, "Fake news detection using dual bert deep neural networks," *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 43 831–43 848, 2024.

[8] D. O. Ong'ong'a, "The role of online news consumers in lessening the extent of misinformation on social media platforms," *Journal Communication Spectrum: Capturing New Perspectives in Communication*, vol. 12, no. 2, pp. 96–111, 2022.

[9] A. Damisa, "Fake news: Finding truth in strategic communication," 2024.

[10] A. Bruns, E. Hurcombe, and S. Harrington, "Covering conspiracy: Approaches to reporting the covid/5g conspiracy theory," *Digital Journalism*, vol. 10, no. 6, pp. 930–951, 2022.

[11] Y.-P. Chen, Y.-Y. Chen, K.-C. Yang, F. Lai, C.-H. Huang, Y.-N. Chen, and Y.-C. Tu, "The prevalence and impact of fake news on covid-19 vaccination in taiwan: retrospective study of digital media," *Journal of Medical Internet Research*, vol. 24, no. 4, p. e36830, 2022.

[12] A. A. Abd El-Mageed, A. A. Abohany, A. H. Ali, and K. M. Hosny, "An adaptive hybrid african vultures-aquila optimizer with xgb-tree algorithm for fake news detection," *Journal of Big Data*, vol. 11, no. 1, p. 41, 2024.

[13] V. U. Gongane, M. V. Munot, and A. Anuse, "Machine learning approaches for rumor detection on social media platforms: a comprehensive survey," *Advanced machine intelligence and signal processing*, pp. 649–663, 2022.

[14] A. Yadav and A. Gupta, "An emotion-driven, transformer-based network for multimodal fake news detection," *International Journal of Multimedia Information Retrieval*, vol. 13, no. 1, pp. 1–16, 2024.

[15] S. Tufchi, A. Yadav, and T. Ahmed, "A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 28, 2023.

[16] K. Soga, S. Yoshida, and M. Muneyasu, "Exploiting stance similarity and graph neural networks for fake news detection," *Pattern Recognition Letters*, vol. 177, pp. 26–32, 2024.

[17] A. Ali and M. Gulzar, "An improved fakebert for fake news detection," *Applied Computer Systems*, vol. 28, no. 2, pp. 180–188, 2023.

[18] Z. Zhang, Q. Lv, X. Jia, W. Yun, G. Miao, Z. Mao, and G. Wu, "Gbca: Graph convolution network and bert combined with co-attention for fake news detection," *Pattern Recognition Letters*, 2024.

[19] Q. Chang, X. Li, and Z. Duan, "Graph global attention network with memory: A deep learning approach for fake news detection," *Neural Networks*, vol. 172, p. 106115, 2024.

[20] Y. Zhang, S. Li, J. Weng, and B. Liao, "Gnn model for time-varying matrix inversion with robust finite-time convergence," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[21] Y.-J. Lu and C.-T. Li, "Gcan: Graph-aware co-attention networks for explainable fake news detection on social media," 2020.

[22] S. Xu, X. Liu, K. Ma, F. Dong, B. Riskhan, S. Xiang, and C. Bing, "Rumor detection on social media using hierarchically aggregated feature via graph neural networks," *Applied Intelligence*, vol. 53, no. 3, pp. 3136–3149, 2023.

[23] L. Wei, D. Hu, W. Zhou, Z. Yue, and S. Hu, "Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection," 2021.

[24] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial intelligence review*, vol. 56, no. 9, pp. 10 345–10 425, 2023.

[25] P. Veličković, "Everything is connected: Graph neural networks," *Current Opinion in Structural Biology*, vol. 79, p. 102538, 2023.

[26] R. Rodríguez-Ferrándiz, "The plandemic and its apostles: Conspiracy theories in pandemic mode," in *Digital totalitarianism*. Routledge, 2022, pp. 62–83.

[27] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, "Content-based fake news detection with machine and deep learning: a systematic review," *Neurocomputing*, vol. 530, pp. 91–103, 2023.

[28] A. Widianto, E. Pebriyanto, F. Fitriyanti, and M. Marna, "Document similarity using term frequency-inverse document frequency representation and cosine similarity," *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 4, no. 2, pp. 149–153, 2024.

[29] L.-C. Chen, "An extended tf-idf method for improving keyword extraction in traditional corpus-based research: An example of a climate change corpus," *Data & Knowledge Engineering*, p. 102322, 2024.

[30] M. H. Al-Tai, B. M. Nema, and A. Al-Sherbaz, "Deep learning for fake news detection: Literature review," *Al-Mustansiriyah Journal of Science*, vol. 34, no. 2, pp. 70–81, 2023.

[31] I. Alshubaily, "Textcnn with attention for text classification," *arXiv preprint arXiv:2108.01921*, 2021.

[32] A. R. Merryton and M. Gethsiyal Augasta, "An attribute-wise attention model with bilstm for an efficient fake news detection," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 38 109–38 126, 2024.

[33] A. Mallik and S. Kumar, "Word2vec and lstm based deep learning technique for context-free fake news detection," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 919–940, 2024.

[34] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional lstm-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2019.

[35] M. Zhao, Y. Zhang, and G. Rao, "Fake news detection based on dual-channel graph convolutional attention network," *The Journal of Supercomputing*, pp. 1–22, 2024.

[36] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 549–556.

[37] H. Thakar and B. Bhatt, "Fake news detection: recent trends and challenges," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 176, 2024.

[38] B. Das *et al.*, "Multi-contextual learning in disinformation research: A review of challenges, approaches, and opportunities," *Online Social Networks and Media*, vol. 34, p. 100247, 2023.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need.(nips), 2017," *arXiv preprint arXiv:1706.03762*, vol. 10, p. S0140525X16001837, 2017.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] S. Gong, R. O. Sinnott, J. Qi, and C. Paris, "Fake news detection through graph-based neural networks: A survey," *arXiv preprint arXiv:2307.12639*, 2023.

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

[45] Q. Chang, X. Li, and Z. Duan, "A novel approach for rumor detection in social platforms: Memory-augmented transformer with graph convolutional networks," *Knowledge-Based Systems*, vol. 292, p. 111625, 2024.

[46] C. Cui and C. Jia, "Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 73–81.

[47] Y. Zhao, W. Li, F. Liu, J. Wang, and A. M. Luvembe, "Integrating heterogeneous structures and community semantics for unsupervised community detection in heterogeneous networks," *Expert Systems with Applications*, vol. 238, p. 121821, 2024.

[48] Q. Huang, C. Zhou, J. Wu, L. Liu, and B. Wang, "Deep spatial–temporal structure learning for rumor detection on twitter," *Neural Computing and Applications*, vol. 35, no. 18, pp. 12 995–13 005, 2023.

[49] G. Zhang, D. Li, H. Gu, T. Lu, and N. Gu, "Heterogeneous graph neural network with personalized and adaptive diversity for news recommendation," *ACM Transactions on the Web*, vol. 18, no. 3, pp. 1–33, 2024.

[50] M. Kang, G. F. Templeton, E. T. Lee, and S. Um, "A method framework for identifying digital resource clusters in software ecosystems," *Decision Support Systems*, vol. 177, p. 114085, 2024.

[51] B. Xie, X. Ma, J. Wu, J. Yang, S. Xue, and H. Fan, "Heterogeneous graph neural network via knowledge relations for fake news detection," in *Proceedings of the 35th International Conference on Scientific and Statistical Database Management*, 2023, pp. 1–11.

[52] Z. Jin, J. Ma, S. Wang, J. Tang, and J. Luo, "Hierarchical propagation network for fake news detection," in *Proceedings of the 29th International Conference on Information and Knowledge Management (CIKM)*. ACM, 2020, pp. 802–811.

[53] S. Wang, Y. Zhang, X. Wang, and J. Li, "Multimodal fusion graph neural networks for fake news detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 4397–4407, 2021.

[54] K. Shu, D. Mahudeswaran, and H. Liu, "Graph-based multimodal embedding for fake news detection," in *Proceedings of The Web Conference (WWW)*. ACM, 2019, pp. 291–300.

[55] X. Zhou, W. Lin, J. Zhang, and Y. Sun, "Incorporating knowledge graphs in multimodal fake news detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2022, pp. 5678–5685.

[56] P. Yang, J. Leng, G. Zhao, W. Li, and H. Fang, "Rumor detection driven by graph attention capsule network on dynamic propagation structures," *The Journal of Supercomputing*, vol. 79, no. 5, pp. 5201–5222, 2023.

[57] T. Liu, Q. Cai, C. Xu, Z. Zhou, F. Ni, Y. Qiao, and T. Yang, "Rumor detection with a novel graph neural network approach," *arXiv preprint arXiv:2403.16206*, 2024.

[58] 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.

[59] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1751–1754.

[60] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[61] N. Ye, D. Yu, Y. Zhou, K.-k. Shang, and S. Zhang, "Graph convolutional-based deep residual modeling for rumor detection on social media," *Mathematics*, vol. 11, no. 15, p. 3393, 2023.

[62] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1395–1405.

[63] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[64] 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*, 2017, pp. 797–806.

[65] J. Yang and G. Yang, "Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer," *Algorithms*, vol. 11, no. 3, p. 28, 2018.

[66] M. A. Khan, "Hcrnnids: Hybrid convolutional recurrent neural network-based network intrusion detection system," *Processes*, vol. 9, no. 5, p. 834, 2021.