GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media

Yi-Ju Lu
Department of Statistics
National Cheng Kung University
Tainan, Taiwan
l1852888@gmail.com

Cheng-Te Li
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
chengte@mail.ncku.edu.tw

Abstract

This paper solves the fake news detection problem under a more realistic scenario on social media. Given the source short-text tweet and the corresponding sequence of retweet users without text comments, we aim at predicting whether the source tweet is fake or not, and generating explanation by highlighting the evidences on suspicious retweeters and the words they concern. We develop a novel neural network-based model, Graph-aware Co-Attention Networks (GCAN), to achieve the goal. Extensive experiments conducted on real tweet datasets exhibit that GCAN can significantly outperform state-of-the-art methods by 16% in accuracy on average. In addition, the case studies also show that GCAN can produce reasonable explanations.

1 Introduction

Social media is indispensable in people's daily life, where users can express themselves, access news, and interact with each other. Information can further spread through the social network. Opinions and sentiments on source stories can be reflected by user participation and interaction. The convenient and low-cost essence of social networking brings collective intelligence, but at the same time leads to a negative by-product, the propagation of misinformation such as *fake news*.

Fake news is a kind of news story possessing intentionally false information on social media (Rashkin et al., 2017; Allcott and Gentzkow, 2017). The widespread of fake news can mislead the public, and produce unjust political, economic, or psychological profit for some parties (Horne and Adali, 2017; Allcott and Gentzkow, 2017). Data mining and machine learning techniques were utilized to detect fake news (Shu et al., 2017; Cha et al., 2020). Typical approaches rely on the content of new articles to extract textual features, such as n-gram and bag of words, and apply supervised learning (e.g., random forest and support vector machine) for binary classification (Shu et al., 2017). NLP researchers also learn advanced linguistic features, such as factive/assertive verbs and subjectivity (Popat, 2017) and writing styles and consistency (Potthast et al., 2018). Multi-modal context information is also investigated, such as user profiles (Yang et al., 2012; Liu and Wu, 2018) and retweet propagation (Ruchansky et al., 2017; Shu et al., 2019a).

Nevertheless, there are still critical challenges in detecting fake news online. First, existing content-based approaches (Castillo et al., 2011; Potthast et al., 2018; Shu et al., 2019a) require documents to be *long* text, e.g., news articles, so that the representation of words and sentences can be better learned. However, tweets on social media are usually *short* text (Yan et al., 2015), which produces severe data sparsity problem. Second, some state-of-the-art models (Ruchansky et al., 2017; Liu and Wu, 2018; Shu et al., 2019a) require a rich collection of *user comments* for every news story, to learn the opinions of retweeters, which usually provide strong evidences in identifying fake news. However, most users on social media tend to simply reshare the source story without leaving any comments (Kwak et al., 2010). Third, some studies (Ma et al., 2018) consider that the pathways of information cascade (i.e., retweets) in the social network are useful for classifying misinformation, and thus learn the representations of the tree-based propagation structures. However, it is costly to obtain the diffusion structure of retweets at most times due to privacy concerns (Li et al., 2018). Many users choose to hide or delete the records of social interactions. Fourth, if the service providers or the government agencies desire to inspect who are the suspicious users who support the fake news, and which topics do they concern in producing fake

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 505–514
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

news (Reis et al., 2019), existing models cannot provide explanations. Although dEFEND (Shu et al., 2019a) can generate reasonable explanation, it requires both long text of source articles and text of user comments.

This paper deals with fake news detection under a more realistic scenario on social media. We predict whether a source tweet story is fake, given only its *short text* content and its *retweet sequence of users*, along with *user profiles*. That said, we detect fake news under three settings: (a) short-text source tweet, (b) no text of user comments, and (c) no network structures of social network and diffusion network. Moreover, we require the fake news detection model to be capable of *explainability*, i.e., highlighting the evidence when determining a story is fake. The model is expected to point out the suspicious retweeters who support the spreading of fake news, and highlight the words they especially pay attention to from the source tweet.

To achieve the goal, we propose a novel model, **G**raph-aware **C**o-**A**ttention **N**etwork (**GCAN**) [1]. We first extract user features from their profiles and social interactions, and learn word embeddings from the source short text. Then we use convolutional and recurrent neural networks to learn the *representation of retweet propagation* based on user features. A graph is constructed to model the potential interactions between users, and the graph convolution network is used to learn the *graph-aware representation of user interactions*. We develop a *dual co-attention mechanism* to learn the correlation between the source tweet and retweet propagation, and the co-influence between the source tweet and user interaction. The binary prediction is generated based on the learned embeddings.

We summarize the contributions as follows. (1) We study a novel and more realistic scenario of fake news detection on social media. (2) For accurate detection, we develop a new model, GCAN, to better learn the representations of user interactions, retweet propagation, and their correlation with source short text. (3) Our dual co-attention mechanism can produce reasonable explanations. (4) Extensive experiments on real datasets demonstrate the promising performance of GCAN, comparing to state-of-the-art models. The GCAN explainability is also exhibited in case studies.

[1]The Code of GCAN model is available and can be accessed via: https://github.com/l852888/GCAN

We organize this paper as follows. Section 2 reviews the relevant approaches to fake news detection in social media. We describe the problem statement in Section 3. Then in Section 4, the details of our proposed GCAN model will be elaborated. Section 5 demonstrates the evaluation settings and results. We conclude this work in Section 6.

2 Related Work

Content-based approaches rely on the text content to detect the truthfulness of news articles, which usually refer to long text. A variety of text characteristics are investigated for supervised learning, including TF-IDF and topic features (Castillo et al., 2011), language styles (e.g., part of speech, factive/assertive verbs, and subjectivity) (Popat, 2017), writing styles and consistency (Potthast et al., 2018), and social emotions (Guo et al., 2019). Zhao et al. (2015) find the enquiry phrases from user responses are useful, and Ma et al. (2016) use recurrent neural networks to learn better representations of user responses.

User-based approaches model the traits of users who retweet the source story. Yang et al. (2012) extract account-based features, such as "is verified", gender, hometown, and number of followers. Shu et al. (2019b) unveil user profiles between fake and real news are significantly different. CRNN (Liu and Wu, 2018) devise a joint recurrent and convolutional network model (CRNN) to better represent retweeter's profiles. Session-based heterogeneous graph embedding (Jiang et al., 2018) is proposed to learn the traits of users so that they can be identified in shared accounts. However, since such a method relies on session information, it cannot be directly applied for fake news detection.

Structure-based approaches leverage the propagation structure in the social network to detect fake news. Sampson et al. (2016) leverage the implicit information, i.e., hashtags and URLs, to connect conversations whose users do not have social links, and find such implicit info can improve the performance of rumor classification. Ma et al. (2017) create a kernel-based method that captures high-order patterns differentiating different types of rumors. Ma et al. (2018) develop a tree-structured recursive neural networks to learn the embedding of rumor propagation structure. Although multi-relational graph embedding methods (Feng et al., 2019; Wang and Li, 2019) are able to effectively learn how different types of entities (related to source news ar-

Table 1: Comparison of related studies. Column notations: news story texts (NS), response comments (RC), user characteristics (UC), propagation structure (PS), social network (SN), and model explainability (ME). For the NS column, "S" and "L" indicates short and long text, respectively.

	NS	RC	UC	PS	SN	ME
Ma et al. (2016)	✓(S)	✓				
Ma et al. (2018)	✓(S)	✓		✓	✓	
Liu and Wu (2018)	✓(S)		✓	✓		
Ruchansky et al. (2017)	✓(S)	✓	✓			
Shu et al. (2019a)	✓(L)	✓			✓	✓
Our work	✓(S)		✓	✓	✓	✓

ticles) interact with each other in a heterogeneous information network for classification tasks, they cannot be applied for the inductive setting, i.e., detecting the truthfulness of new-coming tweets.

Hybrid-based approaches consider and fuse multi-modal context information regarding the source tweets. CSI (Ruchansky et al., 2017) learns the sequential retweet features by incorporating response text and user profiles, and generates suspicious scores of users based on their social interactions. Wang et al. (2018) develop an event adversarial neural network to learn transferable features by removing the event-specific features, along with convolutional neural networks to extract textual and visual features. dEFEND (Shu et al., 2019a) jointly learns the sequential effect of response comments and the correlation between news content and comments, and use an attention mechanism to provide explainability.

We compare our work and the most relevant studies in Table 1. The uniqueness of our work lies in: targeting at short text, requiring no user response comments, and allow model explainability.

3 Problem Statement

Let $\Psi = \{s_1, s_2 ... s_{|\Psi|}\}$ be a set of tweet stories, and $U = \{u_1, u_2 ... u_{|U|}\}$ be a set of users. Each $s_i \in \Psi$ is a short-text document (also called the *source tweet*), given by $s_i = \{q_1^i, q_2^i, ..., q_{l_i}^i\}$ indicating l_i words in story s_i. Each $u_j \in U$ is associated with a user vector $\mathbf{x}_j \in \mathbb{R}^d$ representing the user feature with d dimensions. When a news story s_i is posted, some users will share s_i and generate a sequence of retweet records, which is termed a *propagation path*. Given a news story s_i, we denote its propagation path as $R_i = \{..., (u_j, \mathbf{x}_j, t_j), ...\}$, where $(u_j, \mathbf{x}_j, t_j)$ depicts j-th user u_j (with their feature vector $\mathbf{x}_j$) who retweets story s_i, and $j = 1, 2, ..., K$ (i.e., $K = |R_i|$). We denote the set of users who retweet story s_i as U_i. In R_i, we denote the user who originally shares s_i as u_1 at time t_1. For $j > 1$, user u_j retweets s_i at t_j ($t_j > t_1$). Each story s_i is associated with a binary label $y_i \in \{0, 1\}$ to represent its truthfulness, where $y_i = 0$ indicates story s_i is true, and $y_i = 1$ means s_i is fake.

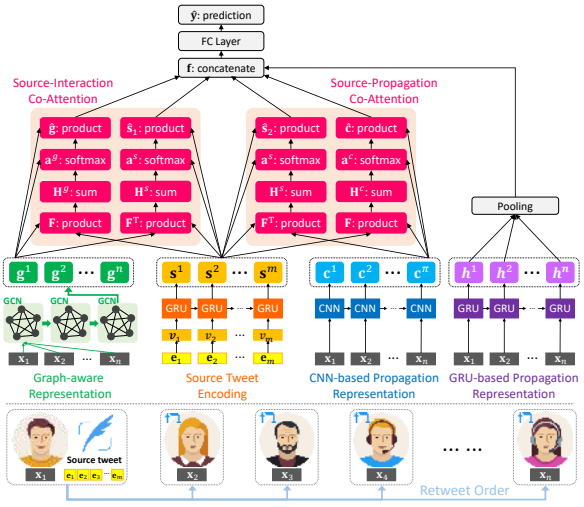

Figure 1: The architecture of our GCAN model.

Given a source tweet s_i, along with the corresponding propagation path R_i containing users u_j who retweet s_i as well as their feature vectors $\mathbf{x}_j$, our goal is to predict the truthfulness y_i of story s_i, i.e., binary classification. In addition, we require our model to highlight few users $u_j \in U_i$ who retweet s_i and few words $q_k^i \in s_i$ that can interpret why s_i is identified as a true or fake one.

4 The Proposed GCAN Model

We develop a novel model, Graph-aware Co-Attention Networks (GCAN), to predict fake news based on the source tweet and its propagation-based users. GCAN consists of five components. The first is *user characteristics extraction*: creating features to quantify how a user participates in online social networking. The second is *new story encoding*: generating the representation of words in the source tweet. The third is *user propagation representation*: modeling and representing how the source tweet propagates by users using their extracted characteristics. The fourth is *dual co-attention mechanisms*: capturing the correlation between the source tweet and users' interactions/propagation. The last is *making prediction*: generating the detection outcome by concatenating all learned representations.

4.1 User Characteristics Extraction

To depict how users participate in social networking, we employ their metadata and profiles to define the feature vector $\mathbf{x}_j$ of every user u_j. The extracted features are listed as follows: (1) number of words in a user's self-description, (2) number of words in u_j's screen name, (3) number of users who follows u_j, (4) number of users that u_j is following, (5) number of created stories for u_j, (6) time elapsed after u_j's first story, (7) whether the u_j account is verified or not, (8) whether u_j allows the geo-spatial positioning, (9) time difference between the source tweet's post time and u_j's retweet time, and (10) the length of retweet path between u_j and the source tweet (1 if u_j retweets the source tweet). Eventually, every user feature vector $\mathbf{x}_j \in \mathbb{R}^v$ is generated, where v is the number of features.

4.2 Source Tweet Encoding

The given source tweet is represented by a word-level encoder. The input is the one-hot vector of each word in story s_i. Since the length of every source story is different, we perform zero padding here by setting a maximum length m. Let $\mathbf{E} = [e_1, e_2, ..., e_m] \in \mathbb{R}^m$ be the input vector of source story, in which e_m is the one-hot encoding of the m-th word. We create a fully-connected layer to generate word embeddings, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m] \in \mathbb{R}^{d \times m}$, where d is the dimensionality of word embeddings. The derivation of $\mathbf{V}$ is given by:

$$\mathbf{V} = \tanh(\mathbf{W}_w \mathbf{E} + \mathbf{b}_w) \tag{1}$$

where $\mathbf{W}_w$ is the matrix of learnable weights, and $\mathbf{b}_c$ is the bias term. Then, we utilize Gating Recurrent Units (GRU) (Chung et al., 2014) to learn the words sequence representation from $\mathbf{V}$. The source tweet representation learning can be depicted by: $\mathbf{s}_t = GRU(\mathbf{v}_t)$, $t \in \{1, ..., m\}$, where m is the GRU dimensionality. We denote the source tweet representation as $\mathbf{S} = [\mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^m] \in \mathbb{R}^{d \times m}$.

4.3 User Propagation Representation

The propagation of source tweet s_i is triggered by a sequence of users as time proceeds. We aim at exploiting the extracted user feature vectors $\mathbf{x}_j$, along with the user sequence spreading s_i, to learn user propagation representation. The underlying idea is that the user characteristics in real news propagations are different from those of fake ones. We make use of Gating Recurrent Units (GRU) and Convolutional Neural Network (CNN) to learn propagation representations.

Here the input is the sequence of feature vectors of users retweeting s_i, denoted by $PF(s_i) = \langle \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_n \rangle$, where n is the fixed length of observed retweets. If the number of users sharing s_i is higher than n, we take the first n users. If the number is lower than n, we resample users in $PF(s_i)$ until its length equals to n.

GRU-based Representation. Given the sequence of feature vectors $PF(s_i) = \langle ..., \mathbf{x}_t, ..., \rangle$, we utilize GRU to learn the propagation representation. Each GRU state has two inputs, the current feature vector $\mathbf{x}_t$ and the previous state's output vector $\mathbf{h}_{t-1}$, and one output vector $\mathbf{h}_t$. The GRU-based representation learning can be depicted by: $\mathbf{h}_t = GRU(\mathbf{x}_t)$, $t \in \{1, ..., n\}$, where n is the dimensionality of GRU. We generate the final GRU-based user propagation embedding $\mathbf{h} \in \mathbb{R}^d$ by average pooling, given by $\mathbf{h} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{h}_t$.

CNN-based Representation. We take advantage of 1-D convolution neural network to learn the sequential correlation of user features in $PF(s_i)$. We consider λ consecutive users at one time to model their sequential correlation, i.e., $\langle \mathbf{x}_t, ..., \mathbf{x}_{t+\lambda-1} \rangle$. Hence the filter is set as $\mathbf{W}_f \in \mathbb{R}^{\lambda \times v}$. Then the output representation vector $\mathbf{C} \in \mathbb{R}^{d \times (t+\lambda-1)}$ is given by

$$\mathbf{C} = \text{ReLU}(\mathbf{W}_f \cdot \mathbf{X}_{t:t+\lambda-1} + b_f) \tag{2}$$

where $\mathbf{W}_f$ is the matrix of learnable parameters, $ReLU$ is the activation function, $\mathbf{X}_{t:t+\lambda-1}$ depicts sub-matrices whose first row's index is from $t = 1$ to $t = n - \lambda + 1$, and b_f is the bias term.

4.4 Graph-aware Propagation Representation

We aim at creating a graph to model the potential interaction among users who retweet source story s_i. The idea is that some correlation between users with particular characteristics can reveal the possibility that the source tweet is fake. To fulfill such an idea, a graph $\mathcal{G}^i = (U_i, \mathcal{E}_i)$ is constructed for the set of users who share source story s_i (i.e., U_i), where $\mathcal{E}_i$ is the corresponding edge set. Since the true interactions between users are unknown, we consider $\mathcal{G}^i$ is a fully-connected graph, i.e., $\forall e_{\alpha\beta} \in \mathcal{E}_i$, $u_\alpha \in U_i, u_\beta \in U_i$, and $u_\alpha \neq u_\beta$, $|\mathcal{E}_i| = \frac{n \times (n-1)}{2}$. To incorporate user features in the graph, each edge $e_{\alpha\beta} \in \mathcal{E}_i$ is associated with

a weight $\omega_{\alpha\beta}$, and the weight is derived based on cosine similarity between user feature vectors $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$, given by $\omega_{\alpha\beta} = \frac{\mathbf{x}_\alpha \cdot \mathbf{x}_\beta}{\|\mathbf{x}_\alpha\|\|\mathbf{x}_\beta\|}$. We use matrix $\mathbf{A} = [\omega_{\alpha\beta}] \in \mathbb{R}^{n\times n}$ to represent weights between any pair of nodes u_α and u_β in graph $\mathcal{G}^i$.

A graph convolution network (GCN) layer (Kipf and Welling, 2017) is created based on the constructed graph $\mathcal{G}^i$ for source tweet s_i. A GCN is a multi-layer neural network that performs on graph data and generates embedding vectors of nodes according to their neighborhoods. GCN can capture information from a node's direct and indirect neighbors through stacking layer-wise convolution. Given the matrix $\mathbf{A}$ for graph $\mathcal{G}^i$, and $\mathbf{X}$ depicting the matrix of feature vectors for users in $\mathcal{G}^i$, the new g-dimensional node feature matrix $\mathbf{H}^{(l+1)} \in \mathbb{R}^{n\times g}$ can be derived by

$$\mathbf{H}^{(l+1)} = \rho(\tilde{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}_l), \tag{3}$$

where l is the layer number, $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the normalized symmetric weight matrix ($\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$), and $\mathbf{W}_l \in \mathbb{R}^{d\times g}$ is the matrix of learnable parameters at the l-th GCN layer. ρ is an activation function, i.e., a ReLU $\rho(x) = \max(0, x)$. Here $\mathbf{H}^{(0)}$ is set to be $\mathbf{X}$. We choose to stack two GCN layers in derive the learned graph-aware representation, denoted as $\mathbf{G} \in \mathbb{R}^{g\times n}$.

4.5 Dual Co-attention Mechanism

We think the evidence of fake news can be unveiled through investigating which parts of the source story are concerned by which kinds of retweet users, and fake clues can be reflected by how retweet users interact with each other. Therefore, we develop a *dual co-attention mechanism* to model the mutual influence between the source tweet (i.e., $\mathbf{S} = [\mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^m]$) and user propagation embeddings (i.e., $\mathbf{C} = [\mathbf{c}^1, \mathbf{c}^2, ..., \mathbf{c}^{n-\lambda+1}]$ from Section 4.3), and between the source tweet and graph-aware interaction embeddings (i.e., $\mathbf{G} = [\mathbf{g}^1, \mathbf{g}^2, ..., \mathbf{g}^n]$ from Section 4.4). Equipped with co-attention learning, our model is capable of the explainability by looking into the attention weights between retweet users in the propagation and words in the source tweet. In other words, by extending the co-attention formulation (Lu et al., 2016), the proposed dual co-attention mechanism aims to attend to the source-tweet words and graph-aware interaction users simultaneously (source-interaction co-attention), and also attend to the source-tweet words and propagated users simultaneously (source-propagation co-attention).

Source-Interaction Co-attention. We first compute a proximity matrix $\mathbf{F} \in \mathbb{R}^{m\times n}$ as: $\mathbf{F} = \tanh(\mathbf{S}^\top \mathbf{W}_{sg} \mathbf{G})$, where $\mathbf{W}_{sg}$ is a $d \times g$ matrix of learnable parameters. By treating the proximity matrix as a feature, we can learn to predict source and interaction attention maps, given by

$$\begin{aligned} \mathbf{H}^s &= \tanh(\mathbf{W}_s\mathbf{S} + (\mathbf{W}_g\mathbf{G})\mathbf{F}^\top) \\ \mathbf{H}^g &= \tanh(\mathbf{W}_g\mathbf{G} + (\mathbf{W}_s\mathbf{S})\mathbf{F}) \end{aligned} \tag{4}$$

where $\mathbf{W}_s \in \mathbf{R}^{k\times d}, \mathbf{W}_g \in \mathbf{R}^{k\times g}$ are matrices of learnable parameters. The proximity matrix $\mathbf{F}$ can be thought to transforming user-interaction attention space to source story word attention space, and vice versa for its transpose $\mathbf{F}^\top$. Then we can generate the attention weights of source words and interaction users through the softmax function:

$$\begin{aligned} \mathbf{a}^s &= \text{softmax}(\mathbf{w}_{hs}^\top \mathbf{H}^s) \\ \mathbf{a}^g &= \text{softmax}(\mathbf{w}_{hg}^\top \mathbf{H}^g) \end{aligned} \tag{5}$$

where $\mathbf{a}^s \in \mathbb{R}^{1\times m}$ and $\mathbf{a}^g \in \mathbb{R}^{1\times n}$ are the vectors of attention probabilities for each word in the source story and each user in the interaction graph, respectively. $\mathbf{w}_{hs}, \mathbf{w}_{hg} \in \mathbb{R}^{1\times k}$ are learnable weights. Eventually we can generate the attention vectors of source story words and interaction users through weighted sum using the derived attention weights, given by

$$\hat{\mathbf{s}}_1 = \sum_{i=1}^{m} \mathbf{a}_i^s \mathbf{s}^i \ , \quad \hat{\mathbf{g}} = \sum_{j=1}^{n} \mathbf{a}_j^g \mathbf{g}^j \tag{6}$$

where $\hat{\mathbf{s}}_1 \in \mathbb{R}^{1\times d}$ and $\hat{\mathbf{g}} \in \mathbb{R}^{1\times g}$ are the learned co-attention feature vectors that depict how words in the source tweet are attended by users who interact with one another.

Source-Propagation Co-attention. The process to generate the co-attention feature vectors, $\hat{\mathbf{s}}_2 \in \mathbb{R}^{1\times d}$ and $\hat{\mathbf{c}} \in \mathbb{R}^{1\times d}$, for the source story and user propagation, respectively, is the same as source-interaction co-attention, i.e., creating another proximity matrix to transform them into each other's space. We skip the repeated details due to the page limit.

Note that the GRU-based user representations are not used to learn the interactions with the source tweet. The reason is that how user profiles in the retweet sequence look like is also important, as suggested by CRNN (Liu and Wu, 2018), and should

Table 2: Statistics of two Twitter datasets.

	Twitter15	Twitter16
# source tweets	742	412
# true	372	205
# fake	370	207
# users	190,868	115,036
avg. retweets per story	292.19	308.70
avg. words per source	13.25	12.81

be emphasized separately. Nevertheless, the CNN-based user representations (i.e., features that depict the sequence of user profiles) has been used in the co-attention mechanism to learn their interactions with source tweet.

4.6 Make Prediction

We aim at predicting fake news using the source-interaction co-attention feature vectors $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{g}}$, the source-propagation feature vectors $\hat{\mathbf{s}}_2$ and $\hat{\mathbf{c}}$, and the sequential propagation feature vector $\mathbf{h}$. Let $\mathbf{f} = [\hat{\mathbf{s}}_1, \hat{\mathbf{g}}, \hat{\mathbf{s}}_2, \hat{\mathbf{c}}, \mathbf{h}]$ which is then fed into a multi-layer feedforward neural network that finally predicts the label. We generate the binary prediction vector $\hat{\mathbf{y}} = [\hat{y}_0, \hat{y}_1]$, where $\hat{y}_0$ and $\hat{y}_1$ indicate the predicted probabilities of label being 0 and 1, respectively. It can be derived through

$$\hat{\mathbf{y}} = \text{softmax}(\text{ReLU}(\mathbf{f}\mathbf{W}_f + \mathbf{b}_f)), \quad (7)$$

where $\mathbf{W}_f$ is the matrix of learnable parameters, and $\mathbf{b}_f$ is the bias term. The loss function is devised to minimize the cross-entropy value:

$$\mathcal{L}(\Theta) = -y \log(\hat{y}_1) - (1-y)\log(1-\hat{y}_0) \quad (8)$$

where Θ denotes all learnable parameters in the entire neural network. We choose the Adam optimizer to learn Θ as it can determine the learning rate abortively.

5 Experiments

We conduct experiments to answer three questions: (1) whether our GCAN model is able to achieve satisfactory performance of fake news detection, compared to state-of-the-art methods? (2) how does each component of GCAN contribute to the performance? (3) can GCAN generate a convincing explanation that highlights why a tweet is fake?

5.1 Datasets and Evaluation Settings

Data. Two well-known datasets compiled by Ma et al. (2017), Twitter15 and Twitter16, are utilized. Each dataset contains a collection of source tweets, along with their corresponding sequences of retweet users. We choose only "true" and "fake" labels as the ground truth. Since the original data does not contain user profiles, we use user IDs to crawl user information via Twitter API.

Competing Methods. We compare our GCAN with the state-of-the-art methods and some baselines, as listed below. (1) **DTC** (Castillo et al., 2011): a decision tree-based model combining user profiles and the source tweet. (2) **SVM-TS** (Ma et al., 2015): a linear support vector machine classifier that utilizes the source tweet and the sequence of retweet users' profiles. (3) **mGRU** (Ma et al., 2016): a modified gated recurrent unit model for rumor detection, which learns temporal patterns from retweet user profile, along with the source's features. (4) **RFC** (Kwon et al., 2017): an extended random forest model combining features from retweet user profiles and the source tweet. (5) **CSI** (Ruchansky et al., 2017): a state-of-the-art fake news detection model incorporating articles, and the group behavior of users who propagate fake news by using LSTM and calculating the user scores. (6) **tCNN** (Yang et al., 2018): a modified convolution neural network that learns the local variations of user profile sequence, combining with the source tweet features. (7) **CRNN** (Liu and Wu, 2018): a state-of-the-art joint CNN and RNN model that learns local and global variations of retweet user profiles, together with the resource tweet. (8) **dEFEND** (Shu et al., 2019a): a state-of-the-art co-attention-based fake news detection model that learns the correlation between the source article's sentences and user profiles.

Model Configuration. Our model is termed "**GCAN**". To examine the effectiveness of our graph-aware representation, we create another version "**GCAN-G**", denoting our model without the graph convolution part. For both our models and competing methods, we set the number of training epochs to be 50. The hyperparameter setting of GCAN is: number of retweet users = 40, word embedding dim = 32, GRU output dim = 32, 1-D CNN output filter size = 3, 1-D CNN output dim = 32, and GCN output dim = 32. The hyperparameters of competing methods are set by following the settings mentioned in respective studies.

Metrics & Settings. The evaluation metrics include Accuracy, Precision, Recall, and F1. We randomly choose 70% data for training and 30% for testing. The conducted train-test is repeated 20

Table 3: Main results. The best model and the best competitor are highlighted by **bold** and underline, respectively.

	Twitter15				Twitter16			
Method	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
DTC	0.4948	0.4806	0.4963	0.4949	0.5616	0.5369	0.5753	0.5612
SVM-TS	0.5190	0.5186	0.5195	0.5195	0.6915	0.6910	0.6928	0.6932
mGRU	0.5104	0.5148	0.5145	0.5547	0.5563	0.5618	0.5603	0.6612
RFC	0.4642	0.5302	0.5718	0.5385	0.6275	0.6587	0.7315	0.6620
tCNN	0.5140	0.5206	0.5199	0.5881	0.6200	0.6262	0.6248	0.7374
CRNN	0.5249	0.5305	0.5296	0.5919	0.6367	0.6433	0.6419	0.7576
CSI	0.7174	0.6867	0.6991	0.6987	0.6304	0.6309	0.6321	0.6612
dEFEND	0.6541	0.6611	0.6584	0.7383	0.6311	0.6384	0.6365	0.7016
GCAN-G	0.7938	0.7990	0.7959	0.8636	0.6754	0.6802	0.6785	0.7939
GCAN	**0.8250**	**0.8295**	**0.8257**	**0.8767**	**0.7593**	**0.7632**	**0.7594**	**0.9084**
Improvement	**15.0%**	**20.8%**	**18.1%**	**18.7%**	**19.3%**	**15.9%**	**3.8%**	**19.9%**

times, and the average values are reported.

5.2 Experimental Results

Main Results. The main results are shown in Table 3. We can clearly find that the proposed GCAN significantly outperforms the best competing methods over all metrics across two datasets, improving the performance by around 17% and 15% on average in Twitter15 and Twitter16, respectively. Even without the proposed graph-aware representation, GCAN-G can improve the best competing method by 14% and 3% on average in Twitter15 and Twitter16, respectively. Such promising results prove the effectiveness of GCAN for fake news detection. The results also imply three insights. First, GCAN is better than GCAN-G by 3.5% and 13% improvement in Twitter15 and Twitter16, respectively. This exhibits the usefulness of graph-aware representation. Second, the dual co-attention mechanism in GCAN is quite powerful, as it clearly outperforms the best non-co-attention state-of-the-art model CSI. Third, while both GCAN-G and dEFEND are co-attention-based, additional sequential features learned from the retweet user sequence in GCAN-G can significantly boost the performance.

Early Detection. We further report the performance (in only Accuracy due to page limit) by varying the number of observed retweet users per source story (from 10 to 50), as exhibited in Figure 2 and Figure 3. It can be apparently found that our GCAN consistently and significantly outperforms the competitors. Even with only ten retweeters, GCAN can still achieve 90% accuracy. Such results tell GCAN is able to generate accurate early detection of the spreading fake news, which is crucial when defending misinformation.

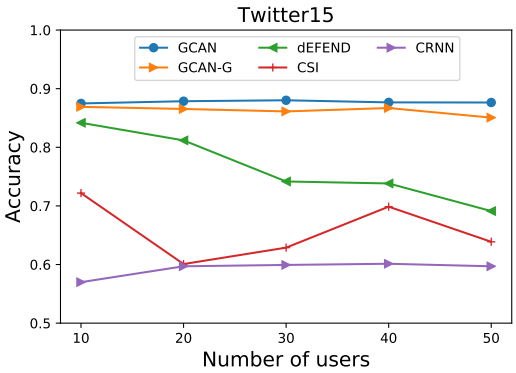

Figure 2: Accuracy by # retweet users in Twitter15.

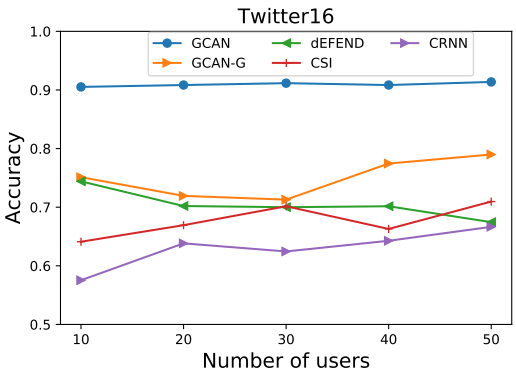

Figure 3: Accuracy by # retweet users in Twitter16.

Ablation Analysis. We report how each of GCAN component contributes by removing each one from the entire model. Below "ALL" denotes using all components of GCAN. By removing dual co-attention, GRU-based representation, graph-aware representation, and CNN-based representation, we have sub-models "-A", "-R", "-G",

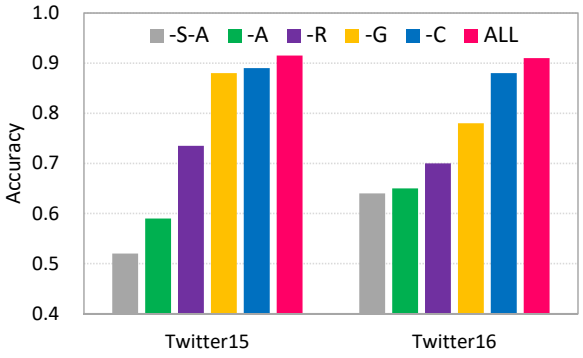

Figure 4: GCAN ablation analysis in Accuracy.

Figure 5: Highlighting evidential words via word cloud. Larger font sizes indicate higher co-attention weights.

and "-C", respectively. Sub-model "-S-A" denotes the one without both source tweet embeddings and dual co-attention. The results are presented in Figure 4. We can find every component indeed plays a significant contribution, especially for dual co-attention ("-A") and the representation learning of user propagation and interactions ("-R" and "-G"). Since the source tweet provides fundamental clues, the accuracy drops significantly without it ("-S-A").

5.3 GCAN Explainability

The co-attention weights derived from Section 4.5 attended on source tweet words and retweet users (source-propagation co-attention) allow our GCAN to be capable of explainability. By exhibiting where attention weights distribute, evidential words and users in predicting fake news can be revealed. Note that we do not consider source-interaction co-attention for explainability because user interaction features learned from the constructed graph cannot be intuitively interpretable.

Explainability on Source Words. To demonstrate the explainability, we select two source tweets in the test data. One is **fake** ("*breaking: ks patient at risk for ebola: in strict isolation at ku med center in kansas city #kwch12*"), and the other is **real** ("*confirmed: this is irrelevant. rt @ksdknews: confirmed: #mike-brown had no criminal record. #ferguson*"). We highlight evidential words with higher co-attention weights in font sizes of word clouds, as exhibited in Figure 5. GCAN predicts the former to be fake with stronger attention on words "breaking" and "strict", and detects the latter as real since it contains "confirmed" and "irrelevant." Such results may correspond to the common knowledge (Rashkin et al., 2017; Horne and Adali, 2017) that fake news tends to use dramatic and obscure words while real news is attended by confirmed and fact checking-related words.

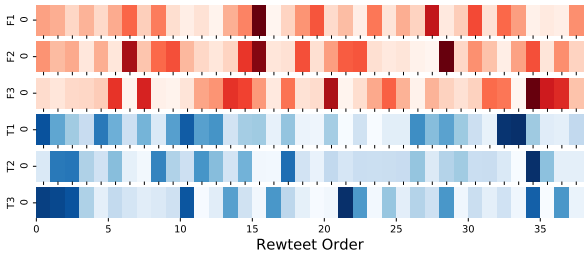

Figure 6: Visualization of attention weights for user propagations of 3 fake (upper F1-F3) and 3 true source tweets. From left to right is retweet order. Dark colors refer to higher attention weights.

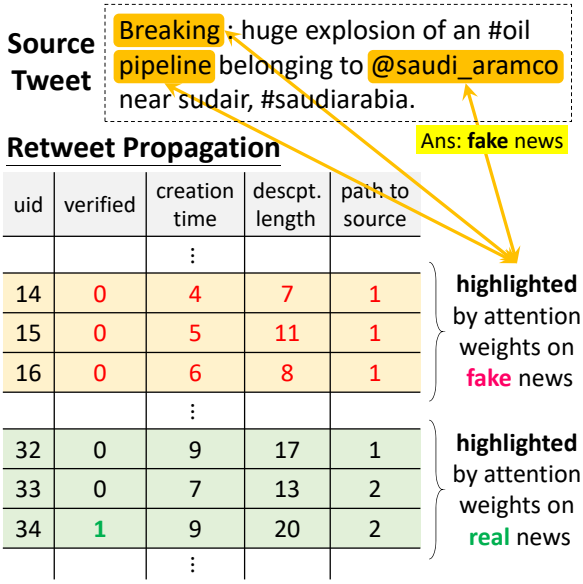

Figure 7: Evidential words highlighed by GCAN in source tweet (upper) and suspicious users highlighed by GCAN in retweet propagation (bottom), in which each column is a user characteristic. Note that only few user characteristics are presented.

Explainability on Retweet Propagation. We aim to exploit the retweet order in propagations to unfold the behavior difference between fake and real news. We randomly pick three fake (F1-F3) and three true (T1-T3) source stories, and plot their

weights from source-propagation co-attention (Section 4.5), as exhibited in Figure 6, in which the horizontal direction from left to right denotes the order of retweet. The results show that to determine whether a story is fake, one should first examine the characteristics of users who **early** retweet the source story. The evidences of fake news in terms of user characteristics may be evenly distributed in the propagation.

Explainability on Retweeter Characteristics. The source-propagation co-attention of our GCAN model can further provide an explanation to unveil the traits of suspicious users and the words they focus on. A case study is presented in Figure 7. We can find that the traits of suspicious users in retweet propagation can be: accounts are not verified, shorter account creation time, shorter user description length, and shorter graph path length to the user who posts the source tweet. In addition, what they highly attend are words "breaking" and "pipeline." We think such kind of explanation can benefit interpret the detection of fake news so as to understand their potential stances.

6 Conclusion

In this study, we propose a novel fake news detection method, Graph-aware Co-Attention Networks (GCAN). GCAN is able to predict whether a short-text tweet is fake, given the sequence of its retweeters. The problem scenario is more realistic and challenging than existing studies. Evaluation results show the powerful effectiveness and the reasonable explainability of GCAN. Besides, GCAN can also provide early detection of fake news with satisfying performance. We believe GCAN can be used for not only fake news detection, but also other short-text classification tasks on social media, such as sentiment detection, hate speech detection, and tweet popularity prediction. We will explore model generalization in the future work. Besides, while fake news usually targets at some events, we will also extend GCAN to study how to remove event-specific features to further boost the performance and explainability.

Acknowledgments

This work is supported by Ministry of Science and Technology (MOST) of Taiwan under grants 109-2636-E-006-017 (MOST Young Scholar Fellowship) and 108-2218-E-006-036, and also by Academia Sinica under grant AS-TP-107-M05.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *The Journal of Economic Perspectives*, 31:211–235.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684.

Meeyoung Cha, Wei Gao, and Cheng-Te Li. 2020. Detecting fake news in social media: An asia-pacific perspective. *Commun. ACM*, 63(4):68–71.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Ming-Han Feng, Chin-Chi Hsu, Cheng-Te Li, Mi-Yen Yeh, and Shou-De Lin. 2019. Marine: Multi-relational network embeddings with relational proximity and node attributes. In *The World Wide Web Conference*, WWW '19, pages 470–479.

Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. Exploiting emotions for fake news detection on social media. *CoRR*, abs/1903.01728.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of AAAI International Conference on Web and Social Media*, pages 759–766.

Jyun-Yu Jiang, Cheng-Te Li, Yian Chen, and Wei Wang. 2018. Identifying users behind shared accounts in online streaming services. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 65–74.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLOS ONE*, 12(1):1–19.

Cheng-Te Li, Yu-Jen Lin, and Mi-Yen Yeh. 2018. Forecasting participants of information diffusion on social networks with its applications. *Information Sciences*, 422:432 – 446.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *AAAI Conference on Artificial Intelligence*, pages 254–261.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 289–297.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. *IJCAI International Joint Conference on Artificial Intelligence*, pages 3818–3824.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. 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*, CIKM '15, pages 1751–1754.

Jing Ma, Wei Gao, and Kam Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 708–717.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1989.

Kashyap Popat. 2017. Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 735–739.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 231–240.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.

Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, pages 17–26.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 797–806.

Justin Sampson, Fred Morstatter, Liang Wu, and Huan Liu. 2016. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2377–2382.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 395–405.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019b. The role of user profile for fake news detection. *CoRR*, abs/1904.13355.

Pei-Chi Wang and Cheng-Te Li. 2019. Spotting terrorists by learning behavior-aware heterogeneous network embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 2097–2100.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 849–857.

Rui Yan, Ian E.H. Yen, Cheng-Te Li, Shiqi Zhao, and Xiaohua Hu. 2015. Tackling the achilles heel of social networks: Influence propagation based language model smoothing. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1318–1328.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. 2018. Ti-cnn: Convolutional neural networks for fake news detection.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1395–1405.