SCARLET: Explainable Attention Based Graph Neural Network for Fake News Spreader Prediction

Bhavtosh Rath[1(✉)], Xavier Morales[2(✉)], and Jaideep Srivastava[1(✉)]

[1] University of Minnesota, Minneapolis, USA
{rathx082,srivasta}@umn.edu
[2] Harvard College, Cambridge, USA
xavier_morales@college.harvard.edu

Abstract. False information and true information fact checking it, often co-exist in social networks, each competing to influence people in their spread paths. An efficient strategy here to contain false information is to proactively identify if nodes in the spread path are likely to endorse false information (i.e. further spread it) or refutation information (thereby help contain false information spreading). In this paper, we propose SCARLET (truSt andCredibility bAsed gRaph neuraLnEtwork model using aTtention) to predict likely action of nodes in the spread path. We aggregate trust and credibility features from a node's neighborhood using historical behavioral data and network structure and explain how features of a spreader's neighborhood vary. Using real world Twitter datasets, we show that the model is able to predict false information spreaders with an accuracy of over 87%.

1 Introduction

Social network platforms like Twitter, Facebook and Whatsapp are used by millions around the world to share information and opinions. Often, the veracity of content shared on these platforms is not confirmed. This gives rise to scenarios where information having conflicting veracity, i.e. false information and its refutation, co-exist. Refutation can be defined as true information which fact checks claims made by a false information. A typical scenario is that false information originates at time t_1, and starts propagating. Once it is identified, its refutation information is created at time t_2 ($t_1 < t_2$). Both pieces of information propagate simultaneously, with many nodes lying in their common spreading paths.

While detecting false information is an important and widely researched problem, an equally important problem is that of preventing the impact of false information spreading. Techniques involve containment/suppression of false information, as well as accelerating the spread of its refutation. *Being able to predict the likely action of such users before they are exposed to false information is an important aspect of such a strategy.* Nodes identified as vulnerable to believing false information can thus 1) be cautioned about the presence of the

K. Karlapalem et al. (Eds.): PAKDD 2021, LNAI 12712, pp. 714–727, 2021.
https://doi.org/10.1007/978-3-030-75762-5_56

false information so that they do not propagate it, and 2) be urged to propagate its refutation. While optimization models based on information diffusion theories have been proposed in the past for misinformation containment, recent advancements in deep learning on graphs serve as the motivation to explore false information control models which use components that exist even before false information starts spreading, namely the underlying network structure and people's historical behavioral data.

Trust and *Credibility* are important psychological and sociological concepts respectively, that have subtle differences in their meanings. While trust represents the confidence one person has in another person, credibility represents generalized confidence in a person based on their perceived performance record [14]. Thus, in a graph representation of a social network, trust is a property of a (directed) edge, while credibility is a property of an individual node. Metzger et al. [7] showed that *the interpretation of a neighbor's credibility by a node relies on its perception of the neighbor based on their trust dynamics.* Motivated with this idea, we propose a graph neural network model that integrates people's credibility and interpersonal trust features in a social network to predict whether a node is likely to spread false information or not. We make the following contributions in this paper:

1) We propose *SCARLET*, a novel user-centric model using graph neural network with attention mechanism to predict whether a node will most likely spread false information, its refutation or be a non-spreader.
2) We demonstrate that a person's decision to spread a false information is sensitive to its perception of neighbor's credibility, and this perception is a function of trust dynamics with the neighbor.
3) To the best of our knowledge, this is the first model being evaluated on real world Twitter datasets of co-existing false and refutation information.

Related Work: Social science research in the past has explored the aspects of people's behavior that cause false information spreading. Jaeger et al. [5] was one of the first to study what makes rumors believable when told by peers instead of authority figures. While it focused on modelling people's anxiety, it served as motivation to explore other sociological features that are relevant to information spreading. Petty and Cacioppo [10] found credibility perception to be an important factor for believing false information. Rosnow et al. [15] proposed that *interpersonal trust* also played an important role in rumor transmission. The idea was further enforced by Morris et al. [8] where they claimed that people assess credibility based on trust relationships with their neighbors in a social network. Motivated by these ideas, there has been much interest in computational models for false information spreader detection using trust, which has shown promising results [12,13]. Many computational techniques to combat false information spreading have been explored over the past decade, as summarized by Sharma et al. [17]. Most models rely on generating relevant features from the information that help distinguish false information from true. Our proposed model is based on recent advances in graph neural networks [22]. In addition, our work proposes an explainable attention based model, inspired from recent work [23,24].

Qui et al. [11] focuses on influence in general, while our model integrates people's psychological and sociological features to identify false information spreaders.

Models inspired by information diffusion models for false information mitigation have also been proposed. Budak et al. [1] proposed an optimization strategy to identify false information spreaders in a network who, when convinced by its refutation, would minimize the number of people receiving the false information. Nguyen et al. [9] proposed greedy approaches to a similar problem of limiting the spread of false information in social networks. More recently, Tong et al. [19] studied the problem as a multiple cascade diffusion problem.

2 Interpersonal Trust and User Credibility Features

2.1 Trust-Based Features

1. **Global Trust (Tr^G):** Global trust are trust scores that are computed on the directed follower-followee network around information spreaders. It is called global because an individual's trust score is sensitive to changes in the network structure. Using the Trust in Social Media (TSM) algorithm [16], we quantify the likelihood of *trusting others* and being *trusted by others*. The TSM algorithm uses a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ as input, together with a specified convergence criteria, and computes trustingness and trustworthiness scores using the equations: $ti(v) = \sum_{\forall x \in out(v)} \left(\frac{w(v,x)}{1+(tw(x))^s} \right)$ and $tw(u) = \sum_{\forall x \in in(u)} \left(\frac{w(x,u)}{1+(ti(x))^s} \right)$ where $u, v, x \in \mathcal{V}$ are nodes, $ti(v)$ and $tw(u)$ are the trustingness and trustworthiness scores of v and u, respectively, $w(v,x)$ is the weight of edge from v to x, $out(v)$ is the set of out-edges of v, $in(u)$ is the set of in-edges of u, and s is the involvement score of the network. The involvement score is basically the potential risk an actor takes when creating a link in the network. Details of the algorithm are excluded due to space constraints and can be found in [16].
2. **Local Trust (Tr^L):** Local trust is computed based on the retweeting behavior of an individual. It is termed local because the trust score depends on node's behavior, and not on the network structure. We consider the proxy for *trusting others* as the fraction of tweets of x that are retweets (RT_x) denoted by $\sum_{\forall i \in t}\{1 \text{ if } i = RT_x \text{ else } 0\}/n(t)$. Meanwhile, we consider the proxy for *trusted by others* as the average number of times x's tweets are retweeted ($n(RT)$) denoted by $\sum_{\forall i \in t} i_{n(RT_x)}/n(t)$. ($t$ represents the most recent tweets posted in x's timeline).

2.2 Credibility-Based Features

Credibility of users is generalized based on features extracted from information posted on their timeline and are obtained from [2]. We generate relevant credibility features for nodes in the network, which can be categorized into two types: user-based and content-based.

1. **User-based Credibility (Cr^U):** User credibility features are extracted from user metadata of nodes in the network. Features used in our model are summarized below:
   A. Registration age (U1): Registration age denotes the time that has transpired since a user created their account. Older accounts tend to be associated with more credible users.
   B. Overall activity count (U2): Activity or statuses count is the number of tweets issued by a user. Low credibility is associated with users who have less activity on their timeline.
   C. Is verified (U3): This label suggests whether a user account is marked as authentic or not by Twitter. Verified accounts are more likely to be credible.
2. **Content-based Credibility (Cr^C):** These features are obtained by aggregating a user's timeline activity. It is important to note that, unlike Castillo's assumption, we do not make a distinction between information that is specifically related to news or not, as that process would require manually assessing newsworthiness of the tweets. The following relevant features are extracted:
   A. Emotions conveyed by user (M1): Emotions represent positive or negative sentiments associated with a tweet. Content with negative sentiments is usually associated with non-credible users [2].
   B. Level of uncertainty (M2): Level of uncertainty is quantified as the fraction of user's tweets that are questioning in nature. Tweets with a high level of uncertainty tend to be less credible.
   C. External source citation (M3): External source citation is quantified as the fraction of user's tweets that cite an external URL. tweets which do not include URLs tend to be related to non-credible news [2].

3 Proposed Approach

This section explains how we integrate both credibility and trust features in an attention based graph neural network model to predict whether a person would likely be a spreader of false information or its refutation. The problem formulation is as follows:

Problem formulation: Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a directed social network containing false information spreaders ($\mathcal{V}_F$), refutation information spreaders ($\mathcal{V}_T$) and non-spreaders ($\mathcal{V}_{\hat{Sp}}$) at a time instance t ($\{\mathcal{V}_F \cup \mathcal{V}_T \cup \mathcal{V}_{\hat{Sp}}\} \subset \mathcal{V}$). By assigning importance score using global (Tr^G) and local (Tr^L) trust features ($Tr = Tr^G||Tr^L$), and aggregating user-based (Cr^U) and content-based (Cr^C) credibility features ($Cr = Cr^U||Cr^C$) of node i and its neighborhood nodes ($\mathcal{N}_i{}^K$) sampled till depth K, we predict whether i is more likely to spread false information, refutation information or be non-spreader at future time $t + \Delta t$.

The proposed graph neural network framework can be broadly divided into two steps:

1. We assign an importance score to neighborhood nodes ($\mathcal{N}_i{}^K$) sampled till depth K based on trust (Tr) features. This is done using an attention mechanism.

2. We learn representations using Graph Convolutional Networks by aggregating credibility (Cr) features proportional to the importance scores assigned for the neighborhood nodes based on step 1.

An overview of the proposed model architecture is shown in Fig. 1. The following subsections explain the framework in detail.

3.1 Importance Score Using Attention:

We apply a graph attention mechanism [21] which attends over the neighborhood of i and, based on their trust features, assigns an importance score to every j ($j \in \mathcal{N}_i$). First, every node is assigned a parameterized weight matrix ($\mathbf{W}$) to perform linear transformation. Then, self-attention is performed using a shared attention mechanism a (a single layer feed-forward neural network) which computes trust-based importance scores. The unnormalized trust score between i,j is represented as:

$$e_{ij} = a(\mathbf{W}_{Tr_i}, \mathbf{W}_{Tr_j}) \tag{1}$$

where e_{ij} quantifies j's importance to i in the context of interpersonal trust. We perform masked attention by only considering nodes in $\mathcal{N}_i$. This way we aggregate features based only on the neighborhood's structure. To make the importance scores comparable across all neighbors we normalize them using the softmax function:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} exp(e_{ik})} \tag{2}$$

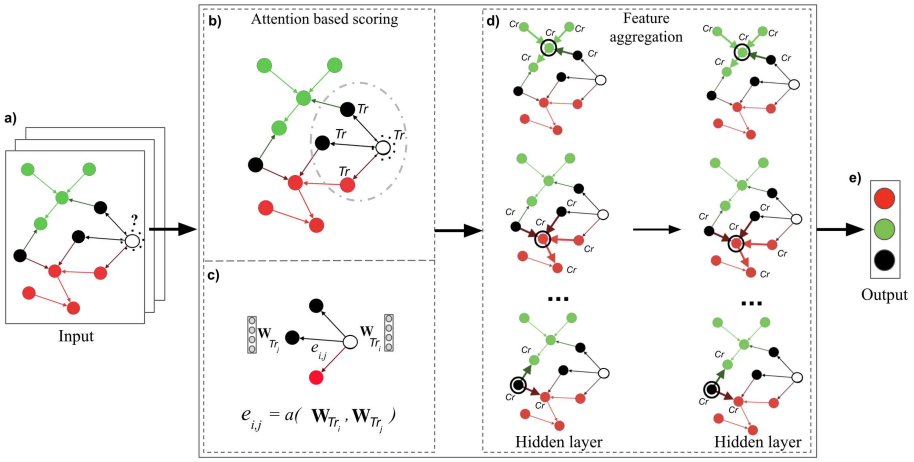

Fig. 1. Architecture overview. Importance score e is assigned to neighbors based on trust features (Tr). Credibility (Cr) features are aggregated proportional to neighbors' importance scores using graph convolution networks for node classification.

The attention layer a is parameterized by weight vector $\mathbf{a}$ and applied using LeakyReLU nonlinearity. Normalized neighborhood edge weights can be represented as:

$$\alpha_{ij} = \frac{exp(LeakyReLU(\mathbf{a}^T[\mathbf{W}_{Tr_i}||\mathbf{W}_{Tr_j}]))}{\sum_{k\in\mathcal{N}_i} exp(LeakyReLU(\mathbf{a}^T[\mathbf{W}_{Tr_i}||\mathbf{W}_{Tr_k}]))} \tag{3}$$

α_{ij} thus represents trust between i and j with respect to all nodes in $\mathcal{N}_i$. Each α_{ij} obtained for the edges is used to create an attention-based adjacency matrix $\hat{A}_{atn} = [\alpha_{ij}]_{|\mathcal{V}|\times|\mathcal{V}|}$ which is later used to aggregate credibility features.

3.2 Feature Aggregation

The Graph Convolution Network [6] is a graph neural network model that efficiently aggregates features from a node's neighborhood. It consists of multiple neural network layers where the information propagation between layers can be generalized by Eq. 4. Here, H represents the hidden layer and A represents the adjacency matrix representation of the subgraph ($A = \hat{A}_{atn}$). $H^{(0)} = Cr$ and $H^{(L)} = Z$, where Z denotes node-level output during transformation.

$$H^{(l+1)} = f(H^{(l)}, A) \tag{4}$$

We implement a Graph Convolution Network with two hidden layers using a propagation rule as explained in [6].

$$H^{(l+1)} = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)}) \tag{5}$$

Here, $\hat{A} = A + I$, where I is the identity matrix of the neighborhood subgraph. This operation ensures that we include self-features during aggregation of neighbor's credibility features. $\hat{D}$ is the diagonal matrix of node degrees for $\hat{A}$, where $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$. $W^{(l)}$ is the layer weight matrix, and σ denotes the activation function. Symmetric normalization of $\hat{D}$ ensures our model is not sensitive to varying scale of the features being aggregated.

3.3 Node Classification

Using credibility features and network structure for nodes in i's neighborhood, node representations are learned from the graph using a symmetric adjacency matrix with attention-based edge weights ($\hat{A} = \hat{D}^{-1/2}\hat{A}_{atn}\hat{D}^{-1/2}$). Following forward propagation model is applied:

$$Z = f(X, \hat{A}_{atn}) = softmax(\hat{A}ReLU(\hat{A}XW^{(0)})W^{(1)}) \tag{6}$$

X represents the credibility features. $W^{(0)}$ and $W^{(1)}$ are input-to-hidden and hidden-to-output weight matrices respectively, and are learnt using gradient descent learning. Classification is performed using the following cross entropy loss function:

$$\mathcal{L} = \sum_{l\in\mathcal{Y}_L}\sum_{f\in Cr} Y_{lf} ln Z_{lf} \tag{7}$$

where $\mathcal{Y}_L$ represents indices of labeled vertices, f represents each of the credibilty features being used in the model, and $Y \in R^{|\mathcal{Y}_L| \times |Cr|}$ is the label indicator matrix.

Table 1. Network dataset statistics for news events N1-N10.

	N1			N2			N3			N4			N5		
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$
F	1,797,059	5,316,114	2,584	885,598	1,824,585	943	1,228,479	2,477,986	1,313	2,607,629	7,146,454	4,552	2,150,820	5,215,120	3,344
T	1,164,162	2,283,160	437	453,537	879,854	403	1,169,681	1,988,576	425	433,616	773,778	467	1,168,820	1,543,513	305
F ∪ T	2,677,924	7,562,503	3,017	1,230,559	2,641,513	1,337	2,198,524	4,458,228	1,738	2,900,925	7,882,019	5,015	3,019,066	6,631,032	3,627
F ∩ T	283,297	8,956	4	108,576	59,912	9	199,636	376	0	140,320	3,273	5	300,574	112,098	22
	N6			**N7**			**N8**			**N9**			**N10**		
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$
F	2,387,610	5,356,288	3,498	627,147	1,071,120	696	2,036,162	2,876,783	894	1,197,935	2,139,912	2,317	2,174,023	4,280,962	2,323
T	1,297,371	1,727,503	481	1,166,528	2,524,907	847	1,058,482	1,513,404	489	2,999,865	6,317,032	1,833	704,006	1,314,996	741
F ∪ T	2,449,434	5,691,728	3,769	1,606,924	3,577,449	1,534	2,663,392	4,082,373	1,365	4,064,545	8,443,888	4,151	2,729,312	5,584,915	3,063
F ∩ T	1,235,547	1,379,510	212	186,751	11,131	9	431,252	305,358	20	133,255	722	1	148,717	699	1

4 Experimental Analysis

4.1 Data Collection

We evaluate our proposed model using real world Twitter datasets. The ground truth of false information and the refuting true information was obtained from *www.altnews.in*, a popular fact checking website based in India and are based around politics in India. The source tweet related to the information was obtained directly as a tweet embedded in the website. From that source tweet, we used the Twitter API to determine the source tweeter and retweeters (proxy for spreaders), the follower-following network of the spreaders (proxy for social network), and user activity data (100 most recent tweets) for all nodes in the network. Trust and credibility scores extracted from the activity data are summarized in Fig. 2 are directly used as feature vectors. Besides evaluating our model on the false information (F) and true information (T) spreading networks separately, we also evaluated our model on the combined information spreading networks (F ∪ T). Details regarding the number of nodes ($|\mathcal{V}|$), edges ($|\mathcal{E}|$), and spreaders ($|Sp|$) for the networks of 10 different news events (N1-N10) is detailed in Table 1.

4.2 Analysis of F ∩ T

F ∩ T in Table 1 denotes the section of the network that was exposed to both the false and its refutation information. An interesting observation is the spreaders who decided to spread both types of information. Figure 3 (a) denotes the distribution of spreaders in F ∩ T who spread false information followed by its refutation (FT) and those whose spread refutation followed by the false information (TF). N1 and N9 is excluded from the analysis as our dataset as we did not have the spreaders' timestamp information. An interesting observation is that the majority of spreaders belong to FT. Intuitively, these are spreaders

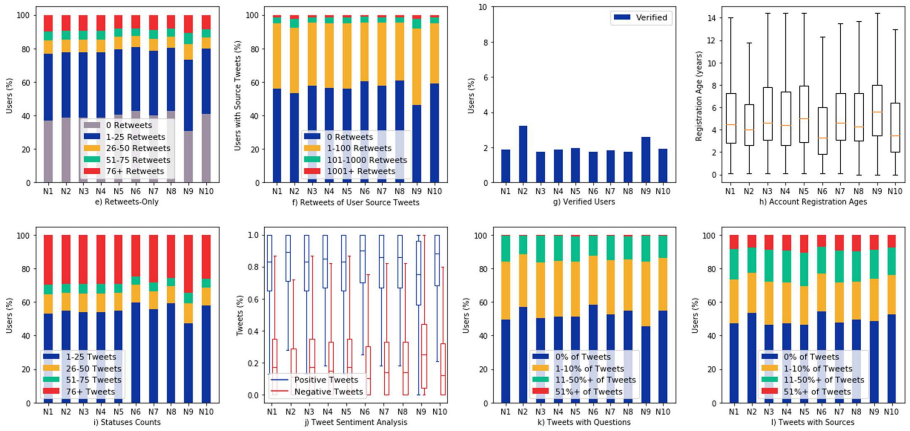

Fig. 2. Trust and credibility feature analysis from networks N1-N10.

who trusted the endorser without verifying the information and later corrected their position, thereby implying that they did not intentionally want to spread false information. Consequently, the proposed model can help identify such people proactively in order to take measures to prevent them from endorsing false information in the first place. While spreaders belonging to TF are comparatively fewer (whose intentions are not certain) the proposed model can help identify them and effective containment strategies can be adopted. Figure 3 (b) shows the time that transpired between spreading refutation and false information for FT spreaders. Once the false information is endorsed, large portions of the network must have already been exposed to false information before the endorser corrected themselves after a significant amount of time (∼1 day). This serves as a strong motivation to have a spreader prediction model which proactively identifies likely future spreaders.

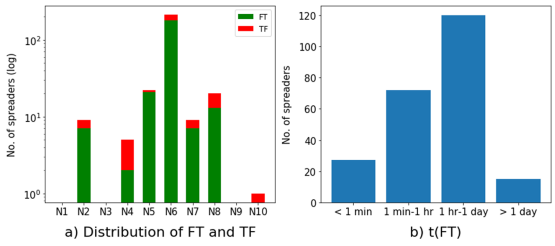

Fig. 3. Analysis of spreaders in F ∩ T.

4.3 Models and Metrics

We compare our proposed attention based model with 10 baseline models. Among the baselines, 3 models use node features only (SVM_{Tr}, SVM_{Cr}, $SVM_{Tr,Cr}$), 1 model uses network structure only ($LINE$) and 6 models integrate both node features and the network structure ($SAGE_{Tr}$, $SAGE_{Cr}$, $SAGE_{Tr,Cr}$, GCN_{Tr}, GCN_{Cr}, $GCN_{Tr,Cr}$).

1. **Node Feature-Based Models:**
   i). SVM_{Tr}: This model applies Support Vector Machines (SVM) [3] on node's trust based features Tr to find an optimal classification threshold.
   ii). SVM_{Cr}: This model applies SVM on node's credibility based features Cr.
   iii). $SVM_{Tr,Cr}$: This model applies SVM by combining node's trust based and credibility based features.
2. **Network Structure-Based Models:**
   iv). $LINE$: Applies the Large-scale Information Network Embedding [18] as a transductive representation learning baseline, where node embeddings are generated after optimization is performed on the entire graph structure.
3. **Network Structure + Node Feature-Based Models:**
   v). $SAGE_{Tr}$: GraphSAGE [4] serves as the inductive learning baseline where node embeddings are generated by aggregating Tr features from neighborhoods.
   vi). $SAGE_{Cr}$: This inductive representation learning baseline generates node embeddings by aggregating Cr features from neighborhoods.
   vii). $SAGE_{Tr,Cr}$: This inductive representation learning baseline generates node embeddings by aggregating both Tr and Cr features from neighborhoods.
   viii). GCN_{Tr}: This model applies Graph Convolution Networks [6] to learn node embeddings by aggregating Tr features from neighborhoods.
   ix). GCN_{Cr}: This model applies Graph Convolution Networks by aggregating Cr features from neighborhoods.
   x). $GCN_{Tr,Cr}$: This model applies Graph Convolution Networks by aggregating both Tr and Cr features from neighborhoods.

$SCARLET$ is the proposed model in this paper, which aggregates a node neighborhood's Cr features based on attention based importance scores assigned using Tr. For evaluation, we did an 80-10-10 train-validation-test split of the dataset. We used 5-fold cross validation and four common metrics: Accuracy, Precision, Recall, and F1 score.

4.4 Implementation Details

We obtained Global Trust features by running the TSM algorithm on the follower-following network of the spreaders. We used the generic settings for

TSM parameters (number of iterations = 100, involvement score = 0.391) based on [16]. The size of sampled neighborhood was set to 50 and depth was set to 1. We considered neighbors with higher degrees in order to generate denser adjacency matrices. The number of epochs, batch size, learning rate and dropout rate were set to 200, 64, 0.001 and 0.2, respectively. The code implementation is also available[1].

Table 2. Model performance evaluation ($\mathcal{V}_F$): False information spreader, ($\mathcal{V}_T$): Refutation spreader.

	F ($\mathcal{V}_F$)				T ($\mathcal{V}_T$)				F $\cup$ **T** ($\mathcal{V}_F$)			
	Accu.	Prec.	Rec.	F1	Accu.	Prec.	Rec.	F1	Accu.	Prec.	Rec.	F1
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

4.5 Performance Evaluation

Classification results of the baselines and proposed model are summarized in Table 2. The results are averaged over the 10 news events. We report the precision, recall, and F1 scores of the false information spreaders class ($\mathcal{V}_F$) in F and F $\cup$ T networks, and of the refutation spreaders class ($\mathcal{V}_T$) in T network. Due to class imbalance, we undersample the majority class to obtain balanced class distribution. We observe that structure only baseline performs better than feature only baselines, and models that combine both node features and network structure show further improvement in performance. Additionally, we observe that Cr features perform better than Tr features (because there are more number of Cr features than Tr features) and the model performance increases when we use Tr and Cr features together. $LINE$, the structure only baseline, performs better than feature only baselines by a substantial margin, which suggests that network structure plays an important role in identifying false information spreaders. In terms of accuracy, the $LINE$ model shows an increase of 32.9%, 22.1% and 54.9% for F, T and F $\cup$ T networks, respectively, over $SVM_{Tr,Cr}$. Graph neural network baselines that combine both network structure and node features show a significant improvement in performance. GCN models perform better than $GraphSAGE$ models on all metrics for F networks, while that is not the case for T and F $\cup$ T networks. This is because Tr and Cr features

[1] https://github.com/BhavtoshRath/GAT-GCN-SpreaderPrediction.

for neighborhood of refutation information spreaders and non-spreaders do not differ much from each other. Our proposed model *SCARLET* shows an increase in performance for all three networks. However, $SAGE_{Tr,Cr}$ shows better accuracy and precision on T networks because the specific news events on which it performed better involved religious tones, and so decision to refute them is more sensitive to neighborhood's Cr than Tr. Precision on F ∪ T networks is highest for $GCN_{Tr,Cr}$, though it is still comparable to the proposed model's performance. More importantly, in the F ∪ T network we observe highest accuracy and F1 scores of 78.9% and 86.6% , thus supporting our hypothesis that false information spreading is very sensitive to trust and credibility.

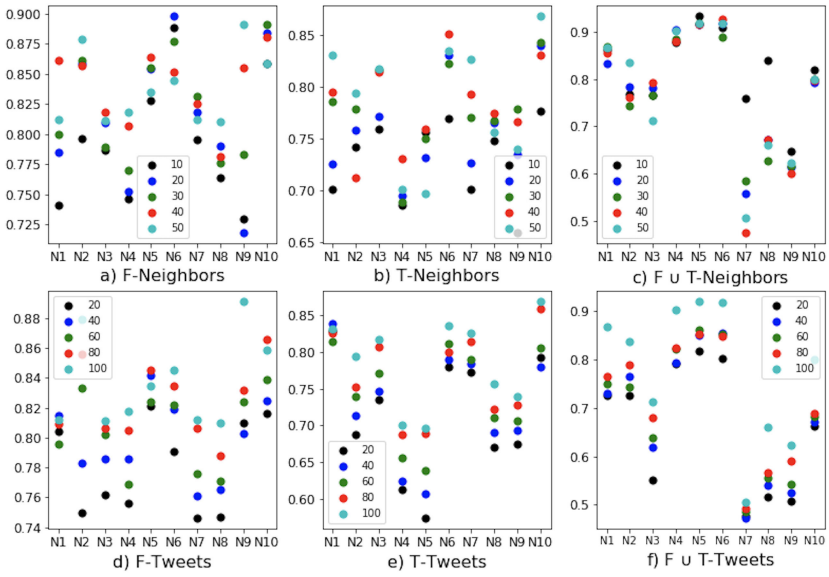

Fig. 4. Sensitivity analysis: Neighborhood size (Neighbors) and features (Tweets). (x-axis: News events N1-N10, y-axis: F1 scores for spreader prediction).

4.6 Sensitivity Analysis

Figure 4 shows the sensitivity analysis of F1 scores of the proposed model on two important parameters: the size of neighborhoods (Neighbors), and the number of recent tweets from user timeline (Tweets).

Neighbors: We evaluated our model on n-neighbors, where n = 10, 20, 30, 40, 50. Figure 4(a), (b), and (c) show results on F, T and F ∪ T networks, respectively. We observe that model performance is not very sensitive to varying neighborhood size, which could be attributed to the fact that since we have only the immediate follower-following network (sampling depth=1) we are not able to entirely capture meaningful dynamics (i.e. the decision to retweet might depend less on the immediate neighbors, and more on the source tweeter).

Tweets: We also evaluated our model on the n-most recent timeline tweets, where n = 20, 40, 60, 80, 100. Figure 4(d), (e), and (f) shows results on F, T and F ∪ T networks, respectively. We observe that for all three networks, prediction performance tends to increase as the number of timeline tweets used to aggregate features increases. This is probably because using more behavioral data helps us estimate trust and credibility features better.

4.7 Explainability Analysis of Trust and Credibility

Figure 5 shows importance scores that false ($\mathcal{V}_F$) and refutation ($\mathcal{V}_T$) spreader's neighbors (size = 10) assign each other based on trust dynamics (softmax attention score) and credibility score (euclidean norm of normalized feature vector) for neighbors with both high and low modularity. Node 0 is the neighbor that the spreader endorses. We observe that $\mathcal{V}_T$'s neighbors have higher credibility than $\mathcal{V}_F$'s neighbors because of network homophily. Also low magnitude of importance scores for neighbors of node 0 of $\mathcal{V}_F$ suggest that it's neighbors trust each other less compared to $\mathcal{V}_T$'s neighbors. We observe in Fig. 5(a) and (b) that node 0 in $\mathcal{V}_F$'s neighbor has strong trust dynamics with its followers (i.e. incoming edges) because it has more incoming edges than outgoing edges and also retweets and gets retweeted substantially more by the neighbors, unlike who $\mathcal{V}_T$ endorses in Fig. 5 c) and d), because $\mathcal{V}_T$'s decision to endorse depends more on information source, which is usually a fact checker.

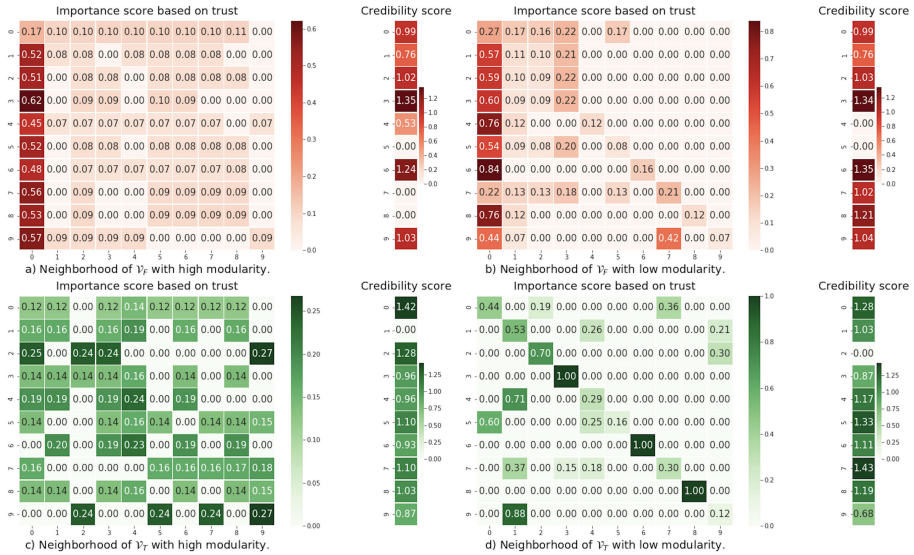

Fig. 5. Explainability analysis. (0–9: Ten highest degree neighbors the spreader follows.)

5 Conclusions and Future Work

We propose *SCARLET*, an attention-based explainable graph neural network model to predict whether a node is likely to spread false information or not. The model learns node embeddings by first assigning trust-based importance scores and then aggregating its neighborhood's credibility features proportionally. What makes this model different from most existing research is that it does not rely on features extracted from the information itself. Thus it can be used to predict spreaders even before information spreading begins. As part of future work, we would like to analyze our model on more news events comprising larger networks in order to sample and aggregate features at greater sampling depths.

References

1. Budak, C., Agrawal, D., Abbadi, A.: Limiting the spread of misinformation in social networks. In: WWW (2011)
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: WWW (2011)
3. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
4. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: NeurIPS (2017)
5. Jaeger, M., Anthony, S., Rosnow, R.: Who hears what from whom and with what effect: a study of rumor. Pers. Soc. Psychol. Bull. **6**, 473–478 (1980)
6. Kipf, T., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
7. Metzger, M., Flanagin, A.: Credibility and trust of information in online environments: the use of cognitive heuristics. J. Pragmatics **59**, 210–220 (2013)
8. Morris, M., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing? Understanding microblog credibility perceptions. In: CSCW (2012)
9. Nguyen, N., Yan, G., Thai, M., Eidenbenz, S.: Containment of misinformation spread in online social networks. In: WebSci (2012)
10. Petty, R., Cacioppo, J.: Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer, Heidelberg (2012). https://doi.org/10.1007/978-1-4612-4964-1
11. Qiu, J., Tang, J., Ma, H., Dong, Y., Wang, K., Tang, J.: Deepinf: social influence prediction with deep learning. In: KDD (2018)
12. Rath, B., Gao, W., Ma, J., Srivastava, J.: Utilizing computational trust to identify rumor spreaders on Twitter. Soc. Netw. Anal. Min. **8**(1), 1–16 (2018). https://doi.org/10.1007/s13278-018-0540-z
13. Rath, B., Gao, W., Srivastava, J.: Evaluating vulnerability to fake news in social networks: a community health assessment model. In: ASONAM (2019)
14. Renn, O., Levine, D.: Credibility and trust in risk communication. In: Kasperson, R.E., Stallen, P.J.M. (eds.) Communicating Risks to the Public. Technology, Risk, and Society (An International Series in Risk Analysis), vol. 4, pp. 175–217. Springer, Dordrecht (1991). https://doi.org/10.1007/978-94-009-1952-5_10
15. Rosnow, R.: Inside rumor: a personal journey. Am. Psychol. **46**, 484 (1991)
16. Roy, A., Sarkar, C., Srivastava, J., Huh, J.: Trustingness & trustworthiness: a pair of complementary trust measures in a social network. In: ASONAM (2016)

17. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: a survey on identification and mitigation techniques. In: TIST (2019)
18. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: large-scale information network embedding. In: WWW (2015)
19. Tong, A., Du, D., Wu, W.: On misinformation containment in online social networks. In: NeurIPS (2018)
20. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
21. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: ICLR (2018)
22. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.: A comprehensive survey on graph neural networks. Trans. Neural Netw. Learn. Syst. (2020)
23. Lu, Y., Li, C.: GCAN: graph-aware co-attention networks for explainable fake news detection on social media. In: ACL (2020)
24. Shu, K., Cui, L., Wang, S., Lee, D., Liu, H.: defend: explainable fake news detection. In: KDD (2019)