A Generalized Deep Markov Random Fields Framework for Fake News Detection

Yiqi Dong[1], **Dongxiao He**[1,2], **Xiaobao Wang**[2 *], **Yawen Li**[3], **Xiaowen Su**[2] and **Di Jin**[1,2]
[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China
{dongyiqi, hedongxiao, wangxiaobao}@tju.edu.cn, warmly0716@bupt.edu.cn,
{suxiaowen, jindi}@tju.edu.cn

Abstract

Recently, the wanton dissemination of fake news on social media has adversely affected our lives, rendering automatic fake news detection a pressing issue. Current methods are often fully supervised and typically employ deep neural networks (DNN) to learn implicit relevance from labeled data, ignoring explicitly shared properties (e.g., inflammatory expressions) across fake news. To address this limitation, we propose a graph-theoretic framework, called Generalized Deep Markov Random Fields Framework (GDMRFF), that inherits the capability of deep learning while at the same time exploiting the correlations among the news articles (including labeled and unlabeled data). Specifically, we first leverage a DNN-based module to learn implicit relations, which we then reveal as the unary function of MRF. Pairwise functions with refining effects to encapsulate human insights are designed to capture the explicit association among all samples. Meanwhile, an event removal module is introduced to remove event impact on pairwise functions. Note that we train GDMRFF with the semi-supervised setting, which decreases the reliance on labeled data while maximizing the potential of unlabeled data. We further develop an Ambiguity Learning Guided MRF (ALGM) model as a concretization of GDMRFF. Experiments show that ALGM outperforms the compared methods significantly on two datasets, especially when labeled data is limited.

1 Introduction

Nowadays, online social media such as Twitter and Weibo have been woven into the fabric of people's lives, providing a convenient platform for users to acquire instantaneous information and share personal opinions. Unfortunately, with this prevalent trend, massive false information known as fake news also proliferates extensively across social media. As fake news may manipulate major public events and affect social safety by misleading users' views [Allcott and Gentzkow, 2017], developing automatic fake news detectors has been in the limelight over the past few years.

Thus far, many methods have achieved impressive performance with extra social contexts such as news propagation networks [Bian *et al.*, 2020] and user comments [Cheng *et al.*, 2020], but those additional social features do not always exist especially when a news article is just emerging, which inspire us to focus on mining the news content itself. Lately, online news content includes a wealth of multimodal resources, involving texts, images, etc. So many works aggregating multimodal content information have emerged. Besides, with the advancement of neural network technology, numerous content-based methods utilize deep neural networks (DNN) to extract unimodal or multimodal features for identifying fake news [Zhou *et al.*, 2020; Chen *et al.*, 2022].

However, these existing DNN-based approaches merely learn the implicit relevance of labeled news. Actually, there are obvious shared characteristics such as negative emotional words and inflammatory expressions among all labeled and unlabeled fake news [Ajao *et al.*, 2019], which is often ignored by existing works. Therefore, we aim at a framework that leverages the implicit relationship modeling capabilities of DNN while effectively incorporating the explicit correlations among news when determining their credibility.

Intuitively, similar news articles are inclined to share the same labels. Nevertheless, a pair of news may be irrelevant even though their labels are consistent if they discuss totally irrelevant events [Zhang *et al.*, 2021]. We believe the event-specific information could contradict the intuitive judgment somehow. To investigate the impact of event information, we calculate the percentage of label consistency news pairs with varying cosine similarity scores on the Twitter dataset [Boididou *et al.*, 2018]. Specifically, the four curves in Fig. 1 correspond to three event-specific news set E_1, E_2, and E_3 (E_1: Boston Marathon, E_2: Hurricane Sandy A, E_3: Hurricane Sandy B) and the union of these three sets E_{All} respectively. From Fig. 1, our major findings are two-fold. Firstly, the increasing trend of all curves demonstrates that as the similarity of news pairs rises, it is likely that have the same labels. Secondly, E_{All} curve is much lower than the E_1, E_2, and E_3 curves, with an average ratio of 11.7%, 10.2% and 34.5%, indicating that event-specific information would weaken the

*Corresponding author

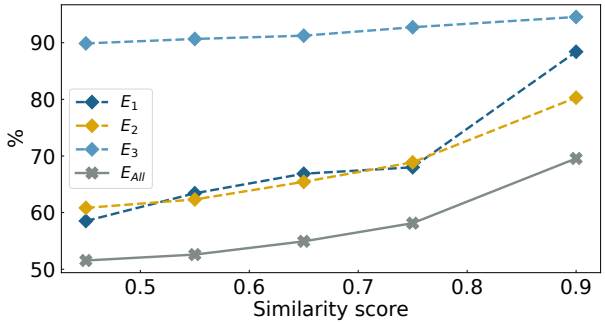


Figure 1: Label consistency percentage of news pairs under different similarity scores. The four curves correspond to three event-specific news sets (E_1, E_2, E_3) and the union of these three sets (E_{All}) respectively.

correlation between label consistency percentage and similarity scores.

Having the aforementioned insights, we would like to model our intuition using Markov Random Fields (MRF) [Jin *et al.*, 2019] while minimizing the impact of event-specific information on the correlation coefficient. Thus, we develop a Generalized Deep Markov Random Fields Framework (GDMRFF) based on graph theory for fake news detection, which inherits the power of deep learning and captures dependencies among news articles jointly. In our framework, we first learn the implicit relations for all news with the superior capability of DNN, which are unfolded to the unary function of MRF. The pairwise potential functions encapsulating our intuition are designed to model the explicit corrections among all samples (including labeled and unlabeled samples), which helps refine the initial results made by DNN. Furthermore, we introduce an event removal module to progressively remove event features during the training process. Note that the proposed GDMRFF is trained in a semi-supervised setting, which decreases the reliance on labeled data while maximizing the potential of unlabeled data. To clearly analyze the effectiveness of our framework, we further propose an Ambiguity Learning Guided MRF (ALGM) model for multimodal fake news detection as a concretization of GDMRFF, which is derived by replacing the content features extraction module using DNN in our framework.

To summarize, the contributions of this paper include:

- To the best of our knowledge, this is the first semi-supervised fake news detection framework that explicitly explores the characteristics shared among fake news only utilizing news content at the early stage of news dissemination.
- We first develop GDMRFF, a general semi-supervised framework for fake news detection, in which a MRF layer capturing shared characteristics is stacked to refine initial predictions made by DNN while removing event-specific features during training. Most existing DNN-based supervised models can be adapted to our framework, lowering the amount of labeled data required and enhancing performance. To verify this, we further present an Ambiguity Learning Guided MRF (ALGM) model as a materialization of GDMRFF.
- Experimental results on two real-world datasets demonstrate ALGM outperforms the state-of-the-art models. In addition, by displaying the latent embeddings, we highlight the improvement effects of GDMRFF's refinement.

2 Methodology

2.1 Task Definition

We focus on identifying fake news in social media using its contents only. Specially, given a news dataset $\mathcal{D}_{all} = \{X, Y\}$ containing news articles X and their ground truth labels $Y \in \{0, 1\}$. Each news item in X can incorporate features from one or more modalities. In our ALGM, we consider two primary modalities, i.e., text and image for each news. The dataset $\mathcal{D}_{all}$ could be divided into a training set $\mathcal{D}_{tra}$ and a test set $\mathcal{D}_{tes}$ with a specific ratio. Given a training set $\mathcal{D}_{tra}$ and a test set $\mathcal{D}_{tes}$ without labels, our goal is to assign label to each news in test set by investigating its own content information and associations across all data. Mathematically,

$$(X, Y) \in \mathcal{D}_{tra}, X \in \mathcal{D}_{tes} \longrightarrow Y \in \mathcal{D}_{tes}. \quad (1)$$

2.2 Model Overview

Our proposed ALGM model aims to explicitly explore the universal features shared among all the fake news. As illustrated in Fig. 2, it integrates three major modules: a) *Multimodal Feature Extraction* that extracts features from two modalities and derives a fused embedding for each news item (This module can be replaced by any other DNN-based methods, allowing our Generalized Deep Markov Random Fields Framework (GDMRFF) to improve their performance); b) *Event Removal* that includes an event discriminator to minimize event-specific feature in the joint news representation; and c) *MRF Inference* that explicitly captures shared fake news properties based on the entire datatset.

Given a dataset $\mathcal{D}_{all}$ with N articles, we begin by extracting features from their texts and attached images separately, then aligning and fusing the unimodal features in an adaptive manner. After obtaining joint multimodal representations, they are passed into an event removal module to eliminate event-related characteristics as much as possible. In the meanwhile, they are also converted into preliminary predicted probabilities, which are then served as the unary potential functions for MRF. At last, we construct a global news similarity graph utilizing the multimodal representations of all news, upon which our pairwise potential functions of MRF are designed to model our intuition.

2.3 Multimodal Feature Extraction

Unimodal Feature Extraction. To exploit underlying semantic information of each news $x_i, \forall i \in \{1, ..., N\}$, we employ the pre-trained BERT [Kenton and Toutanova, 2019] and ResNet34 [He *et al.*, 2016] to extract latent text and image features $\boldsymbol{f}_i^t$ and $\boldsymbol{f}_i^v$ in x_i separately. Subsequently, we feed $\boldsymbol{f}_i^t$ and $\boldsymbol{f}_i^v$ into linear layers to have a fixed dimension d_f.

Cross-modal Alignment. Before fusing text and image features, an issue that needs to be considered is the inherent information gaps persisting across heterogeneous modalities

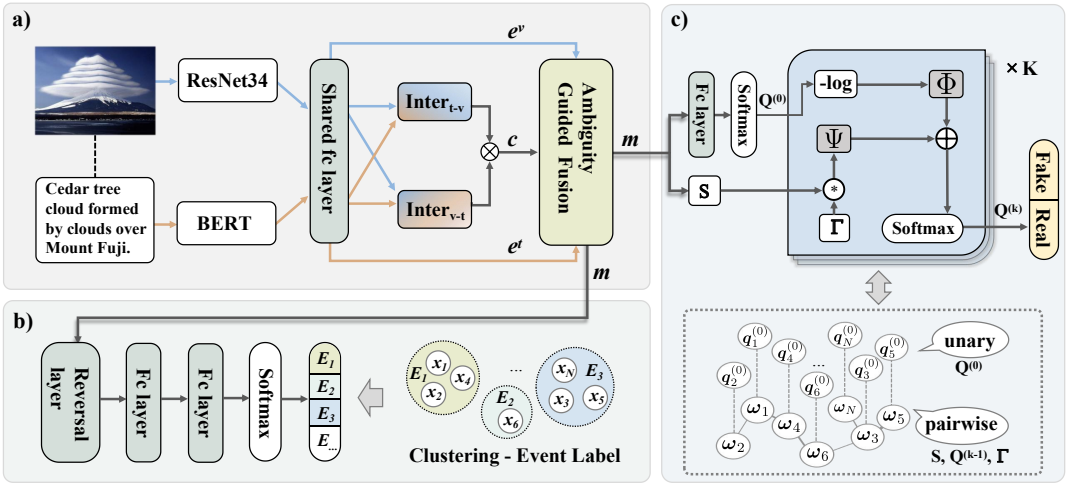


Figure 2: The architecture of the proposed ALGM based on GDMRFF. Its three main modules are in different background colors. Note that the upper part of module 'c)' is the specific operation process inside the MRF, while the lower part is the overall structure of the MRF. These two parts are equivalent.

[Hazarika *et al.*, 2020]. We handle this by introducing a cross-modal alignment component, which embodies a shared fully connected layer coupled with a Leaky Rectified Linear Units (LeakyReLU) to map features from different spaces into a common semantic space as follows:

$$\begin{aligned} \boldsymbol{e}_i^t &= \text{LeakyReLU}\left(\boldsymbol{W}_1 \boldsymbol{f}_i^t + \boldsymbol{b}_1\right), \\ \boldsymbol{e}_i^v &= \text{LeakyReLU}\left(\boldsymbol{W}_1 \boldsymbol{f}_i^v + \boldsymbol{b}_1\right), \end{aligned} \tag{2}$$

where $\boldsymbol{e}_i^t, \boldsymbol{e}_i^v \in \mathbb{R}^{d_e}$ denote aligned representations, d_e is the feature dimension after alignment, $\boldsymbol{W}_1 \in \mathbb{R}^{d_e \times d_f}, \boldsymbol{b}_1 \in \mathbb{R}^{d_e}$ are learnable parameters of the shared layer.

Ambiguity Guided Feature Fusion. To adaptively incorporate features from different modalities, we utilize an ambiguity guided feature fusion component following [Chen *et al.*, 2022], which considers the inherent ambiguity (i.e., degree of information gap) between different content modalities. The ambiguity score is estimated by learning the Kullback-Leibler (KL) divergence between two unimodal distributions that are approximated by two variational autoencoders separately. For each news x_i with aligned textual and visual representations, the variational posteriors could be denoted as:

$$\begin{aligned} g\left(\boldsymbol{z}_i^t \mid \boldsymbol{e}_i^t\right) &= \mathcal{N}\left(\boldsymbol{z}_i^t \mid \mu\left(\boldsymbol{e}_i^t\right), \sigma^2\left(\boldsymbol{e}_i^t\right)\right), \\ g\left(\boldsymbol{z}_i^v \mid \boldsymbol{e}_i^v\right) &= \mathcal{N}\left(\boldsymbol{z}_i^v \mid \mu\left(\boldsymbol{e}_i^v\right), \sigma^2\left(\boldsymbol{e}_i^v\right)\right), \end{aligned} \tag{3}$$

where μ is the mean and σ^2 is the variance. Taking the entire dataset into account, they are

$$\begin{aligned} g\left(\boldsymbol{z}^t\right) &= \frac{1}{N}\sum_{i=1}^{N} g\left(\boldsymbol{z}_i^t \mid \boldsymbol{e}_i^t\right), \\ g\left(\boldsymbol{z}^v\right) &= \frac{1}{N}\sum_{i=1}^{N} g\left(\boldsymbol{z}_i^v \mid \boldsymbol{e}_i^v\right). \end{aligned} \tag{4}$$

Then, an averaged KL divergence is served as the ambiguity score η_i of different modalities as:

$$\begin{aligned} \eta_i^1 &= \left(\frac{\mathrm{D}_{KL}\left(g\left(\boldsymbol{z}_i^t \mid \boldsymbol{e}_i^t\right) \| g\left(\boldsymbol{z}_i^v \mid \boldsymbol{e}_i^v\right)\right)}{\mathrm{D}_{KL}\left(g\left(\boldsymbol{z}^t\right) \| g\left(\boldsymbol{z}^v\right)\right)}\right), \\ \eta_i^2 &= \left(\frac{\mathrm{D}_{KL}\left(g\left(\boldsymbol{z}_i^v \mid \boldsymbol{e}_i^v\right) \| g\left(\boldsymbol{z}_i^t \mid \boldsymbol{e}_i^t\right)\right)}{\mathrm{D}_{KL}\left(g\left(\boldsymbol{z}^v\right) \| g\left(\boldsymbol{z}^t\right)\right)}\right), \end{aligned} \tag{5}$$

$$\eta_i = \text{sigmoid}\left(\frac{1}{2}\left(\eta_i^1 + \eta_i^2\right)\right), \tag{6}$$

where D_{KL} represents the calculation of KL divergence and sigmoid is a normalization function.

The interaction between modalities is conducive to induce the complementary features from each others, especially when unimodal features alone emerge as strongly incongruous. So an interaction vertor $\boldsymbol{c}_i \in \mathbb{R}^{d_e}$ is computed as follows:

$$\begin{aligned} \hat{\boldsymbol{e}}_i^t &= \text{softmax}\left(\left[\boldsymbol{e}_i^t\right]\left[\boldsymbol{e}_i^v\right]^T / \sqrt{d_e}\right) \times \boldsymbol{e}_i^t, \\ \hat{\boldsymbol{e}}_i^v &= \text{softmax}\left(\left[\boldsymbol{e}_i^v\right]\left[\boldsymbol{e}_i^t\right]^T / \sqrt{d_e}\right) \times \boldsymbol{e}_i^v, \end{aligned} \tag{7}$$

$$\boldsymbol{c}_i = \left[\boldsymbol{W}_2\left(\hat{\boldsymbol{e}}_i^t \otimes \hat{\boldsymbol{e}}_i^v\right) + \boldsymbol{b}_2\right]^T, \tag{8}$$

where softmax is an activation function for normalizing, $[\cdot]^T$ indicates matrix transposition, $\otimes$ denotes the outer product, and $\boldsymbol{W}_2 \in \mathbb{R}^{1 \times d_e}, \boldsymbol{b}_2 \in \mathbb{R}^{1 \times d_e}$ are learnable parameters.

To get the final feature representation, unimodal and multimodal features are concatenated adaptively guided by the ambiguity score as follows:

$$\boldsymbol{m}_i = \left(\eta_i \times \boldsymbol{c}_i\right) \oplus \left(\left(1-\eta_i\right) \times \boldsymbol{e}_i^t\right) \oplus \left(\left(1-\eta_i\right) \times \boldsymbol{e}_i^v\right), \tag{9}$$

where $\oplus$ denotes the concatenation of vectors.

2.4 Event Removal

Due to event-specific information has an adverse impact on the correlation between news label consistency percentage

and similarity scores statistically, we intend to remove as many event related features as possible in each news vector $\boldsymbol{m}_i$. We first adopt an event discriminator containing two fully connected layers followed by a softmax function, whose original purpose is to learn event-specific features to distinguish among different events, while our aim is exactly the opposite. Accordingly, we apply a gradient reversal layer [Ganin and Lempitsky, 2015; Wang *et al.*, 2018] before the discriminator to maximize the cross-entropy loss function over the whole dataset defined as below:

$$\mathcal{L}_b\left(\psi^a, \psi^b\right) = -\sum_{i=1}^{N}\sum_{d=1}^{d_b} \tau_i^d \log \hat{\tau}_i^d, \tag{10}$$

where ψ^a and ψ^b are all parameters to be trained in the multimodal feature extraction module and the event removal module respectively. τ_i is the ground-truth event label distribution of news x_i, which is obtained by k-means clustering using the text embeddings $\boldsymbol{f}^t = \{\boldsymbol{f}_1^t, \boldsymbol{f}_2^t, ..., \boldsymbol{f}_N^t\}$ of all data from the BERT extractor, $\hat{\tau}_i$ is the predicted event distribution of news x_i, and d_b is the number of event categories.

2.5 MRF Inference

Considering the explicit relations among all the news, we aim to capture those relations utilizing a Markov Random Field with its ability in modeling joint probability distribution over dependent random variables. Our MRF is built on top of the multimodal feature extraction module, thus taking the final feature representations $\boldsymbol{m} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_N\}$ as input. Above all, we construct a global news similarity graph $\mathcal{S} = (\mathcal{V}, \mathcal{E})$ based on $\boldsymbol{m}$, where the i-th node in $\mathcal{V}$ ($|\mathcal{V}| = N$) denotes the news x_i and edges in $\mathcal{E}$ link similar news pairs. Then the pairwise potential functions of MRF are designed over this graph.

In our MRF, each x_i is associated with a random variable ω_i, which represents its label. The set of all nodes' label assignments is represented as random variables $\omega_{\mathcal{V}}$ with domain $L = \{0, 1\}$. The MRF can be estimated in this Gibbs distribution form:

$$P(\omega_{\mathcal{V}}) = \frac{1}{Z}\exp(-E(\omega_{\mathcal{V}})), \tag{11}$$

$$E(\omega_{\mathcal{V}}) = \sum_{i\in\mathcal{V}} \Phi(\omega_i^u) + \alpha \sum_{(i,j)\in\mathcal{E}} \Psi(\omega_i^u, \omega_j^r), \tag{12}$$

where Z is a constant for normalizing, $(u, r) \in L$ are the labels of news items. Energy function $E(\omega_{\mathcal{V}})$ includes the unary potential $\Phi(\omega_i^u)$ that measures the cost that assign label u to node x_i and the pairwise potential $\Psi(\omega_i^u, \omega_j^r)$ denoting the cost that nodes x_i and x_j have labels u and r respectively. α is a trainable parameter.

As the exact distribution of $P(\omega_{\mathcal{V}})$ is infeasible, we employ the mean-field theory [Koller and Friedman, 2009] to approximate $P(\omega_{\mathcal{V}})$ by a factorizable distribution $Q(\omega_{\mathcal{V}}) = \prod_{i\in\mathcal{V}} Q_i(\omega_i)$. We define $Q_i(\omega_i = u)$ as the probability that the node x_i has the label u. Here we use q_i^u to indicate $Q_i(\omega_i = u)$, which can be iteratively updated as follows:

$$q_i^u = \frac{1}{Z_i}\exp\left\{-\left(\Phi(\omega_i^u) + \alpha \sum_{(i,j)\in\mathcal{E}}\sum_{r\in L} q_j^r \Psi(\omega_i^u, \omega_j^r)\right)\right\}, \tag{13}$$

where Z_i is a normalization term.

We use the unary potential function $\Phi(\omega_i^u)$ to serve as an interface between DNN-based module (i.e., the multimodal feature extraction module) and MRF, fully absorbing the ability of DNN in learning high-level representations into our MRF layer. Concretely, the unary term is calculated by:

$$\Phi(\omega_i^u) = -\log p(\omega_i = u) = -\log\left(\boldsymbol{q}_i^{(0)}[u]\right), \tag{14}$$

where the probability $\boldsymbol{q}_i^{(0)}$ that inherits the power of DNN is derived by projecting the final representation $\boldsymbol{m}_i$ of node x_i from the DNN-based feature extraction module into two dimensions, and $\boldsymbol{q}_i^{(0)}[u]$ is the component value of the u-th dimension of $\boldsymbol{q}_i^{(0)}$, which is used as the prior probability that node x_i has label u.

Since we have experimentally proved that similar news pairs tend to share the same labels, we now model this correlation with pairwise potentials of MRF. In order to measure the similarity between news pairs, we introduce a parameter matrix-guided cosine distance metric function whose computed values are taken as edge weights in the global news similarity graph $\mathcal{S}$:

$$s_{i,j} = \cos\left(\mathbf{w}_{i,j} \odot \boldsymbol{m}_i, \mathbf{w}_{i,j} \odot \boldsymbol{m}_j\right), \tag{15}$$

where $\odot$ represents the Hadamard product. $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ are final vectors of node x_i and node x_j from the multimodal feature extraction module respectively. The parameter $\mathbf{w}_{i,j}$ has the same dimension as $\boldsymbol{m}_i$. Besides, a threshold δ is preset to control the sparsity of graph $\mathcal{S}$, that is, once $s_{i,j}$ is lower than δ, its value will be displaced by 0.

Based on this, $s_{i,j}$ reflects how strong the similarity between two nodes is, then the pairwise term is designed as:

$$\Psi(\omega_i^u, \omega_j^r) = s_{i,j}\gamma(u, r), \tag{16}$$

where $\gamma(u, r)$ is the label compatibility, if $u! = r$, it equals 1, otherwise it equals 0. In our pairwise, similar node pairs with inconsistent labels will be penalized. Moreover, the more similar the node x_i and the node x_j are, the larger the values of $s_{i,j}$ and $\Psi(\omega_i^u, \omega_j^r)$, yet considering the negative sign in Equation (14), the value of q_i^u will decrease, i.e., the penalty will increase accordingly. In this way, the relation between label consistency percentage and similarity is embodied in the MRF layer.

Then we extend Equation (13) to the entire dataset and stack K MRF layers, and we get the following matrix form that can be updated iteratively:

$$\boldsymbol{Q}^{(k)} = \mathrm{softmax}\left(\log(\boldsymbol{Q}^{(0)}) - \alpha \boldsymbol{S}\boldsymbol{Q}^{(k-1)}\boldsymbol{\Gamma}\right), \tag{17}$$

where k denotes the k-th MRF layer. The output of the previous layer $\boldsymbol{Q}^{(k-1)}$ is sent to the k-th layer as an input. $\boldsymbol{Q}^{(0)} \in \mathbb{R}^{N\times d_c}$ is the preliminary probability matrix from the multimodal feature extraction module, whose i-th row vector equals to $\boldsymbol{q}_i^{(0)}$, the adjacency matrix $\boldsymbol{S} \in \mathbb{R}^{N\times N}$ comprises all the edge weights in the news similarity graph, and $\Gamma \in \mathbb{R}^{d_c\times d_c}$ is the label compatibility matrix. d_c is the dimensionality of news authenticity labels.

Algorithm 1 Model Training of ALGM

Input: Dataset: $\mathcal{D}$, hyper-parameter: $d_f, d_e, d_b, d_c, \delta, K, \lambda$, and learning rate: ρ
Output: Parameters: ψ^a, ψ^b, ψ^c

1: **while** not converge **do**
2: update parameters ψ^a in the multimodal feature extraction module: $\psi^a \leftarrow \psi^a - \rho\left(\frac{\partial \mathcal{L}_c}{\partial \psi^a} - \lambda \frac{\partial \mathcal{L}_b}{\partial \psi^a}\right)$
3: update parameters ψ^b in the event removal module: $\psi^b \leftarrow \psi^b - \rho \frac{\partial \mathcal{L}_b}{\partial \psi^b}$
4: update parameters ψ^c in the MRF inference module: $\psi^c \leftarrow \psi^c - \rho \frac{\partial \mathcal{L}_c}{\partial \psi^c}$
5: **end while**

We regard the output $\boldsymbol{Q}^{(K)}$ of the last MRF layer as the predicted probability of news credibility. For fake news detection, a cross-entropy loss function is devised as follows:

$$\mathcal{L}_c\left(\psi^a, \psi^c\right) = -\sum_{i=1}^{N}\sum_{d=1}^{d_c} y_i^d \log \hat{y}_i^d, \tag{18}$$

where ψ^c absorbs all parameters to be trained in the MRF inference module, y_i and $\hat{y}_i$ are the ground-truth and predicted label distribution of the node x_i separately.

2.6 Model Learning

Combining the event removal loss $\mathcal{L}_b$ and the fake news detection loss $\mathcal{L}c$, the overall loss function is defined as:

$$\mathcal{L}_{all}\left(\psi^a, \psi^b, \psi^c\right) = \mathcal{L}_c\left(\psi^a, \psi^c\right) - \lambda \mathcal{L}_b\left(\psi^a, \psi^b\right), \tag{19}$$

where λ is a trade-off parameter between the two terms. The parameters $\hat{\psi}^a, \hat{\psi}^b, \hat{\psi}^c$ that we desire are the saddle point of the overall loss function:

$$\begin{aligned}\left(\hat{\psi}^a, \hat{\psi}^c\right) &= \arg\min_{\psi^a, \psi^c} \mathcal{L}_{all}\left(\psi^a, \hat{\psi}^b, \psi^c\right), \\ \hat{\psi}^b &= \arg\max_{\psi^b} \mathcal{L}_{all}\left(\hat{\psi}^a, \psi^b, \hat{\psi}^c\right).\end{aligned} \tag{20}$$

Here a gradient reversal layer is introduced to achieve the objective in Equation (20), which multiplies the gradient by $-\lambda$ during the backpropagation process. The overall training process of our ALGM is outlined in Algorithm 1.

3 Experiments

3.1 Experimental Setup

Datasets. We use two widely-used datasets collected from Twitter and Weibo for fair evaluation. The Twitter dataset is part of MediaEval released for Verifying Multimedia Use task [Boididou *et al.*, 2018], which is spilt into the development set and test set. The Weibo dataset was released by [Jin *et al.*, 2017]. The real news is collected from Xinhua News Agency, an authoritative news source of China, while the fake one is verified by the official rumor debunking system of Weibo. We provide comprehensive statistics for these two datasets in Table 1. Before training, we carry out a series of pre-processing steps including removing posts with attached videos, deleting special symbols and hyperlinks in texts and normalizing images for each dataset.

Datasets	Train	Test	Total
Twitter	11847	1406	13253
Weibo	6151	1697	7848

Table 1: Statistics of the two datasets used in our experiments.

Compared Baselines. To evaluate the performance of the proposed ALGM, we choose two categories of compared models, namely: 1) **Unimodal methods**, i.e., Image and Text, which adopts a pre-trained ResNet34 model and BERT model coupled with a fully connected layer, respectively; 2) **Multimodal methods**, i.e., NeuralTalk [Vinyals *et al.*, 2015], att_RNN [Jin *et al.*, 2017], EANN [Wang *et al.*, 2018], MVAE [Khattar *et al.*, 2019], SAFE [Zhou *et al.*, 2020], BTIC [Zhang *et al.*, 2021], CAFE [Chen *et al.*, 2022]. These methods typically apply deep neural networks such as CNN, RNN, and attention mechanisms to extract multimodal features, combined with some well-designed sub-task or strategy like cross-modal alignment and contrastive learning, then they are trained to suitable models based on general rules of supervised classification.

Implementation Details. In the multimodal feature extraction module, we fix the dimension $d_f = 128$, $d_e = 64$. We set the number of event categories d_b as 20 on Twitter and 50 on Weibo. The sparsity threshold δ of the news graph $\mathcal{S}$ is set to 0.4. We optimize parameters in our model with Adam [Kingma and Ba, 2014] optimizer. The learning rates on the two datasets are equal to 10^{-4} and 10^{-2} respectively. To prevent over-fitting, we adopt a random dropout with a probability of 0.5. Our model is trained with a maximum epoch of 200 and we use the early stopping strategy.

3.2 Performance Comparison

The performances of baselines and our proposed ALGM are presented in Table 2. Experimental results clearly exhibit that our ALGM achieves the best performance on both datasets in terms of accuracy, precision, recall, and macro F_1 score. Specifically, ALGM reaches the highest accuracy of 89.1% and 84.6%, surpassing the state-of-the-art method's 3.2% and 2.0% on Twitter and Weibo datasets respectively.

We can observe many commonalities between the two datasets. Considering the unimodal methods, Text (BERT) performs much better than Image (ResNet34). This implies that textual features in news provide more evidence for determining the veracity of news. In addition, multimodal techniques beat unimodal approaches on nearly all metrics, proving that incorporating multimodal features can provide supplementary support in detecting fake news. On both datasets, the proposed ALGM outperforms all compared models significantly. We believe this is the result of our realistic simulation of the general characteristics of fake news and the full use of the advantages of the semi-supervised setting, i.e., in addition to labeled data, we extensively mine the information from unlabeled data (which is impossible to achieve by the

Models	Twitter				Weibo			
	Accuracy	Precision	Recall	Macro F_1	Accuracy	Precision	Recall	Macro F_1
Image	0.559	0.544	0.543	0.543	0.522	0.521	0.521	0.521
Text	0.636	0.625	0.622	0.623	0.723	0.722	0.723	0.723
NeuralTalk	0.610	0.631	0.628	0.610	0.726	0.739	0.777	0.723
att_RNN	0.664	0.669	0.672	0.664	0.772	0.787	0.656	0.773
EANN	0.644	0.637	0.639	0.630	0.778	0.779	0.777	0.778
MVAE	0.745	0.745	0.748	0.744	0.824	0.828	0.822	0.823
SAFE	0.762	0.763	0.767	0.761	0.816	0.817	0.816	0.817
BTIC	0.842	0.859	0.853	0.842	0.802	0.803	0.803	0.802
CAFE	0.859	0.872	0.862	0.859	0.826	0.834	0.827	0.825
ALGM	**0.891**	**0.909**	**0.893**	**0.890**	**0.846**	**0.846**	**0.846**	**0.845**

Table 2: Performance comparison between our model (ALGM) and the considered baselines on the two datasets in terms of accuracy, precision, recall, and macro F_1 score. The best results are in bold.

Datasets	Models	Acc	P	R	F_1
Twitter	ALGM	**0.891**	**0.908**	**0.893**	**0.890**
	w/o event	0.878	0.891	0.880	0.877
	w/o mrf	0.842	0.849	0.843	0.841
	w/o both	0.828	0.829	0.829	0.828
Weibo	ALGM	**0.846**	**0.846**	**0.846**	**0.845**
	w/o event	0.840	0.841	0.841	0.840
	w/o mrf	0.839	0.839	0.838	0.839
	w/o both	0.837	0.838	0.836	0.837

Table 3: Main experimental results of ablation study. The best results are in bold.

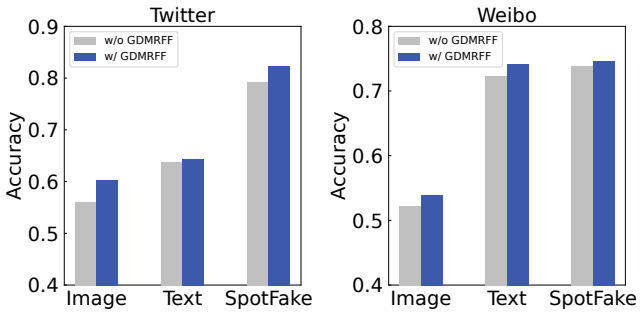


Figure 3: The performance comparison for three supervised DNN-based models without and with modification using our generalized deep Markov Random Fields framework (GDMRFF).

full-supervised model) by the generalized deep Markov Random Fields framework (GDMRFF).

3.3 Ablation Study

Effectiveness of Key Components. Firstly, we investigate the impact of each key component in ALGM on performance by removing it from the entire model. Concretely, we get the ALGM (w/o event) and ALGM (w/o mrf) by omitting the event removal and MRF inference module respectively. Then we eliminate the above two modules together and obtain the ALGM (w/o both). When the MRF inference module is omitted from these variants, we simply append a fully connected layer with a softmax function to the end of the feature extraction module for final prediction.

As shown in Table 3, ALGM (w/o event) yields inferior results, demonstrating that event-related characteristics do indeed impede the utility of MRF and confound our model. ALGM (w/o mrf) also yields poorer performance, indicating that the apparent consideration of shared features in fake news is crucial for our task. ALGM (w/o both) performs the worst of the three variants. This proves that both modules provide significant discriminability when identifying fake news, and that integrating them could improve their effectiveness.

Effectiveness of GDMRFF. The second experiment is to explicate the effectiveness of the generalized deep Markov Random Fields framework we proposed. Our framework can be adapted to any other supervised DNN-based model for detecting fake news, reducing the model's dependency on labeled data and improving its performance. Consequently, we select three supervised models to verify this, including an image model (ResNet34), a text model (BERT), and a classical multimodal fake news detection model (SpotFake [Singhal *et al.*, 2019], which concatenates textual features from the BERT model and visual features from the VGG-19 [Simonyan and Zisserman, 2014] model). Specifically, we replace the multimodal feature extraction module in our framework with each selected model. Then, we compare the performance of the original models to that of the framework-modified models. The results are plainly displayed in Fig. 3. The performance improvement of each model with varying degrees strongly proves the generalization and effectiveness of GDMRFF.

3.4 Data Ratio Analysis

In real scenarios, labeled news articles are quite sparse and limited due to the high cost of acquisition, while there is a sheer volume of unlabeled data. A pivotal issue here is how to make full use of those unlabeled data, so in this pa-

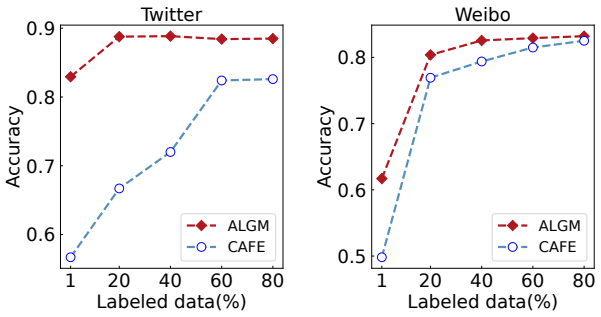


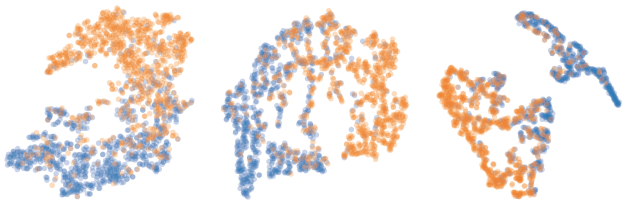
Figure 4: Results of different labeled data proportions.

per, we develop a semi-supervised framework to capture dependencies among all the samples (unlabeled data included). To explore the performance of our model with little labeled data, we vary the number of labeled samples in the training set (i.e., ranging from 1% up to 80%). The best-performing baseline CAFE is chosen for comparison. Results in Fig. 4 display that ALGM always outperforms CAFE in all proportions on both datasets. Despite the fact that all models perform poorly with less training data, it is clear that ALGM degrades more slowly, suggesting that ALGM is more useful in the real world, when labeled data is extremely sparse. We attribute this to the collaborative modeling of all labeled and unlabeled data by GDMRFF, which also takes advantage of the rich features contained in unlabeled news.

On the Twitter dataset, we can see when the proportion of labeled data decreases from 80% to 20%, yet the performance of our model has little or no variation. This is because this dataset contains a big number of tweets that are quite similar. Consequently, the relationships between news are strengthened, allowing GDMRFF to play a more significant role.

3.5 Visualization

To further evaluate our model intuitively, we visualize the news representations learned by ALGM, ALGM (w/o both) and CAFE (a best-performing baseline) with the t-SNE tool [Van der Maaten and Hinton, 2008] on the Weibo dataset. Fig. 5 shows the visualized results, from which we can observe that ALGM can notably generate more discriminative news representations than ALGM (w/o both) and CAFE. As depicted in Fig. 5(a) and Fig. 5(b), these two models are also able to separate news into two categories roughly, but the vectors still have a certain part of intersection and entanglement at the boundary.



(a) CAFE (b) ALGM(w/o both) (c) ALGM

Figure 5: The t-SNE visualization of news representations. Nodes with consistent labels have the same color.

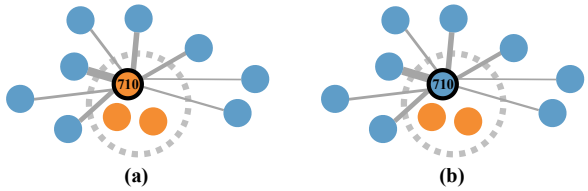


Figure 6: (a) the node with id 710 was misclassified by the DNN-based module; (b) but it was corrected by the MRF layer.

In contrast, the vector representations learned by our integrated model have a more pronounced separation boundary and larger inter-class distance shown in Fig. 5(c). We believe this benefits from the refinement of GDMRFF. During training, with the aid of the event removal module, we can get features that are only specific to news categories. After modeling the connection within each category of news through MRF, all news can have smaller intra-class distances with the larger inter-class separated areas.

3.6 Case Study

We present an example in Weibo that was misclassified by the DNN-based module in Fig. 6(a) but was corrected by our MRF module in Fig. 6(b). Nodes represent news articles, the thickness of the edges represents the similarity between articles, and the color of each node represents its predicted label. Nodes within the gray dotted circle are related to the same event. For the node with id 710 in Fig. 6(a), it was erroneously assigned to "real", since the other two nodes in the same event (these three nodes are all about shampoo) are labeled as "real" and the DNN-based module tends to classify nodes with similar semantic information into the same category. In comparison, the MRF module correctly assigned this node to "fake" as shown in Fig. 6(b). Because the MRF layer removes event-specific information and refines the initial results by using information from its all neighbors in the global similarity graph that models the unique characteristics shared among all fake news.

4 Conclusion

In this work, we focus on content-based fake news detection. A disadvantage of existing DNN-based methods is their inability to model the explicitly shared features in fake news. Thus, we propose a Generalized Deep Markov Random Fields Framework (GDMRFF) based on graph theory, which leverages the superiorities of DNN in capturing high-level features, and of MRF in integrating the explicit correlations among all labeled and unlabeled data. Moreover, based on the proposed GDMRFF, we also develop a concrete ambiguity learning guided MRF model, named ALGM. Experiments on two public datasets demonstrate the effectiveness of GDMRFF. In future work, we plan to extend our framework to be unsupervised to further diminish the reliance on labeled news and make it more applicable to real-world scenarios.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No. 62276187, 62272340, 62172056).

References

[Ajao *et al.*, 2019] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Sentiment aware fake news detection on online social networks. In *Proceedings of the 14th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2507–2511, 2019.

[Allcott and Gentzkow, 2017] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[Bian *et al.*, 2020] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 549–556, 2020.

[Boididou *et al.*, 2018] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[Chen *et al.*, 2022] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the 14th ACM Web Conference*, pages 2897–2905, 2022.

[Cheng *et al.*, 2020] Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of the 12th ACM Web Conference*, pages 2892–2898, 2020.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32th International Conference on Machine Learning*, pages 1180–1189, 2015.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 795–816, 2017.

[Jin *et al.*, 2019] Di Jin, Ziyang Liu, Weihao Li, Dongxiao He, and Weixiong Zhang. Graph convolutional networks meet markov random fields: Semi-supervised community detection in attribute networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 152–159, 2019.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[Khattar *et al.*, 2019] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *Proceedings of the 11th ACM Web Conference*, pages 2915–2921, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Singhal *et al.*, 2019] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the 5th IEEE International Conference on Multimedia Big Data*, pages 39–47, 2019.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[Wang *et al.*, 2018] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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*, pages 849–857, 2018.

[Zhang *et al.*, 2021] Wenjia Zhang, Lin Gui, and Yulan He. Supervised contrastive learning for multimodal unreliable news detection in covid-19 pandemic. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 3637–3641, 2021.

[Zhou *et al.*, 2020] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe:similarity-aware multi-modal fake news detection. In *Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367, 2020.