Multimodal Emergent Fake News Detection via Meta Neural Process Networks

Yaqing Wang[§], Fenglong Ma[◇], Haoyu Wang[§], Kishlay Jha[†] and Jing Gao[§]
[§]Purdue University, West Lafayette, Indiana, USA
[◇]Pennsylvania State University, Pennsylvania, USA
[†]University of Virginia, Charlottesville, Virginia, USA
[§]{wang5075, jinggao, wang5346}@purdue.edu, [◇]fenglong@psu.edu, [†]kishlay@email.virginia.edu

ABSTRACT

Fake news travels at unprecedented speeds, reaches global audiences and puts users and communities at great risk via social media platforms. Deep learning based models show good performance when trained on large amounts of labeled data on events of interest, whereas the performance of models tends to degrade on other events due to domain shift. Therefore, significant challenges are posed for existing detection approaches to detect fake news on emergent events, where large-scale labeled datasets are difficult to obtain. Moreover, adding the knowledge from newly emergent events requires to build a new model from scratch or continue to fine-tune the model, which can be challenging, expensive, and unrealistic for real-world settings. In order to address those challenges, we propose an end-to-end fake news detection framework named MetaFEND, which is able to learn quickly to detect fake news on emergent events with a few verified posts. Specifically, the proposed model integrates meta-learning and neural process methods together to enjoy the benefits of these approaches. In particular, a label embedding module and a hard attention mechanism are proposed to enhance the effectiveness by handling categorical information and trimming irrelevant posts. Extensive experiments are conducted on multimedia datasets collected from Twitter and Weibo. The experimental results show our proposed MetaFEND model can detect fake news on never-seen events effectively and outperform the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Web applications**.

KEYWORDS

meta-learning; fake news detection; natural language processing

ACM Reference Format:
Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, Jing Gao. 2021. Multimodal Emergent Fake News Detection via Meta Neural Process Networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467153

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3467153

1 INTRODUCTION

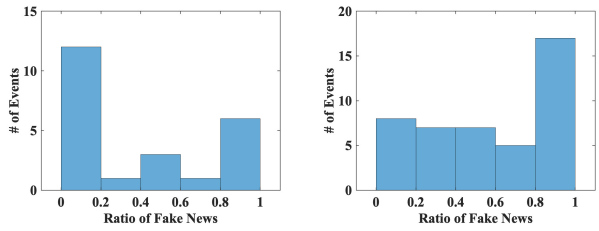
The recent proliferation of social media has significantly changed the way in which people acquire information. According to the 2018 Pew Research Center survey, about two-thirds of American adults (68%) get news on social media at least occasionally. The fake news on social media usually take advantage of multimedia content which contain misrepresented or even forged images, to mislead the readers and get rapid dissemination. The dissemination of fake news may cause large-scale negative effects, and sometimes can affect or even manipulate important public events. Recent years have witnessed a number of high-impact fake news spread regarding terrorist plots and attacks, presidential election and various natural disasters. Therefore, there is an urgent need for the development of automatic detection algorithms, which can detect fake news as early as possible to stop the spread of fake news and mitigate its serious negative effects.

Figure 1: Fake news examples on an emergent event Boston Bombing from Twitter.

Task Challenges. Thus far, various fake news detection methods, including both traditional learning [5, 32] and deep learning based models [21–23, 26, 28, 35] have been exploited to identify fake news. Despite the success of deep learning models with large amounts of labeled datasets, the algorithms still suffer in the cases where fake news detection is needed on emergent events. Due to the domain shift in the news events [38], the model trained on past events may not achieve satisfactory performance and thus the new knowledge from emergent events are needed to add into fake news detection models. However, adding the knowledge from newly emergent events requires to build a new model from scratch or continue to fine-tune the model on newly collected labeled data, which can be challenging, expensive, and unrealistic for real-world settings.

Moreover, fake news usually emerged on newly arrived events where we hardly obtain sufficient posts in a timely manner. In the early stage of emergent events, we usually only have a handful of related verified posts (An example is shown in the Fig. 1). How to leverage *a small set of verified posts* to make the model learn quickly to detect fake news on the newly-arrived events is a crucial challenge.

Limitations of Current Techniques. To overcome the challenge above, the few-shot learning, which aims to leverage a small set of data instances for quick learning, is a possible solution. One promising research line of few-shot learning is **meta-learning** [6, 19], whose basic idea is to leverage the global knowledge from previous tasks to facilitate the learning on new task. However, the success of existing meta-learning methods is highly associated with an important assumption: the tasks are from a similar distribution and the shared global knowledge applies to different tasks. This assumption usually does not hold in the fake news detection problem as the writing style, content, vocabularies and even class distributions of news on different events usually tends to differ. As it



(a) Twitter (b) Weibo

Figure 2: The number of events with respect to different percentages of fake news.

can be observed from Figure 2, the ratios of fake news on events are significantly different. The significant difference across events posts serious challenges on **event heterogeneity**, which cannot be simply handled by globally sharing knowledge [40]. Another research line of few-shot learning is **neural processes** [7, 8, 16], which conduct inference using a small set of data instances as conditioning. Even though neural processes show better generalizablity , they are based on a fixed set of parameters and usually suffer from the limitations like **underfitting** [16], thereby leading to unsatisfactory performance. These two research lines of models are complementary to each other: the parameter adaptation mechanism in meta-learning can provide more parameter flexibility to *alleviate unfitting issues* of the neural process. Correspondingly, the neural processes can help handle *the heterogeneity challenge* for MAML by using a small set of data instances as conditioning instead of encoding all the information into parameter set. Although it is promising to integrate two popular few-shot approaches together, the incompatible operations on the given small set of data instances is the main obstacle for developing the model based on these two.

Our Approach. To address the aforementioned challenges, in this paper, we propose a novel meta neural process network (namely MetaFEND) for emergent fake news detection. MetaFEND unifies the incompatible operations from meta-learning and neural process via a simple yet novel simulated learning task, whose goal is to adapt the parameters to better take advantage of given support data points as conditioning. Toward this end, we propose to conduct leave-one-out prediction as shown in the Fig. 3, i.e., we repeatedly use one of given data as target data and the rest are used as context set for conditioning on all the data in support set. Therefore, the proposed model can handle heterogeneous events via event adaption parameters and conditioning on event-specific data instances simultaneously. Furthermore, we incorporate two novel components - *label embedding* and *hard attention* - to handle categorical characteristics of label information and extract the most informative instance as conditioning despite imbalanced class distributions of news events. Experimental results on two large real-world datasets show that the proposed model effectively detect fake news on new events with a handful of posts and outperforms the state-of-the-art approaches.

Our Contributions. The main contributions of this paper can be summarized as follows:

- We recognize the challenges of fake news detection on emergent events and formulate the problem into a few-shot learning setting. Towards this end, we propose an effective meta neural process framework to detect fake news on emergent events with a handful of data instances.
- The proposed MetaFEND method fuses the meta-learning method and neural process models together via a simulated learning task design. We also propose two components *label embedding* and *hard attention* to handle categorical information and select the formative instance respectively. The effects of two components are investigated in the experiments.
- We empirically show that the proposed method MetaFEND can effectively identify fake news on various events and largely outperform the state-of-the-art models on two real-world datasets.

2 BACKGROUND

We define our problem and introduce preliminary works in this section.

2.1 Problem Formulation

There are many tasks related to fake news detection, such as rumor detection [14] and spam detection [29]. Following the previous work [28, 30], we specify the definition of fake news as news which is intentionally fabricated and can be verified as false. In this paper, we tackle fake news detection on emergent events and make a practical assumption that a few labeled examples are available per event. Our goal is to leverage the knowledge learned from past events to conduct effective fake news detection on newly arrived events with a few examples. More formally, we define the fake news detection following the few-shot problem.

Few-shot Fake News Detection Let $\mathcal{E}$ denote a set of news events. In each news event $e \sim \mathcal{E}$, we have a few labeled posts on the event e. The core idea of few-shot learning is to use episodic classification paradigm to simulate few-shot settings during model training. In each episode during the training stage, the labeled posts are partitioned into two independent sets, support set and query set. Let $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\} = \{x_{e,i}^s, y_{e,i}^s\}_{i=1}^K$ represent the support set, and $\{\mathbf{X}_e^q, \mathbf{Y}_e^q\} =$

$\{x^q_{e,i}, y^q_{e,i}\}_{i=K+1}^{N}$ be the query set. The model is trained to learn to conduct fake news detection on the query set $\{\mathbf{X}^q_e, \mathbf{Y}^q_e\}$ given the support set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\}$. During the inference stage, K labeled posts are provided per event. For each event e, the model leverages its corresponding K labeled posts as support set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\} = \{x^s_{e,i}, y^s_{e,i}\}_{i=1}^{K}$ to conduct fake news detection on given event e.

2.2 Preliminary Work

MAML. We first give an overview of MAML method [6], a representative algorithm of gradient-based meta-learning approaches, and take few-shot fake news detection as an example. The meta-learning procedure is split into two stages: meta-training and meta-testing.

During the *meta-training* stage, the baseline learner f_θ is adapted to specific event e as f_{θ_e} with the help of the support set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\}$. Such an event specific learner f_{θ_e} is evaluated on the corresponding query set $\{\mathbf{X}^q_e, \mathbf{Y}^q_e\}$. The loss $\mathcal{L}(f_{\theta_e}, \{\mathbf{X}^q_e, \mathbf{Y}^q_e\})$ on $\{\mathbf{X}^q_e, \mathbf{Y}^q_e\}$ is used to update the parameters of baseline learner θ. During the meta-testing stage, the baseline learner f_θ is adapted to the testing event e' using the procedure in meta-training stage to obtain event specific parameters $\theta_{e'}$, which is employed to make predictions on the query set $\{\mathbf{X}^q_{e'}, \mathbf{Y}^q_{e'}\}$ of event e'.

MAML update parameter vector θ using one or more gradient descent updates on event e. For example, when using one gradient update:

$$\theta_e = M(f_\theta, \{\mathbf{X}^s_e, \mathbf{Y}^s_e\}) = \theta - \alpha \bigtriangledown_\theta \mathcal{L}(f_\theta, \{\mathbf{X}^s_e, \mathbf{Y}^s_e\}).$$

The model parameters are trained by optimizing for the performance of f_{θ_e} with respect to θ across events sampled from $p(\mathcal{E})$. More concretely, the meta-objective is as follows:

$$\min_\theta \sum_{e\sim\mathcal{E}} \mathcal{L}(f_{\theta_i}) = \sum_{e\sim\mathcal{E}} \mathcal{L}(f_{\theta-\alpha\bigtriangledown_\theta \mathcal{L}(f_\theta, \{\mathbf{X}^s_e, \mathbf{Y}^s_e\})}, \{\mathbf{X}^q_e, \mathbf{Y}^q_e\}).$$

Limitations of MAML. The MAML can capture task uncertainty via one or several gradient updates. However, in fake news detection problem, when events are heterogeneous, the event uncertainty is difficult to encode into parameters via one or several gradient steps. Moreover, even if given support data and query data of interest are from the same event, there is no guarantee that they are all highly related to each other. In such a case, the parameter adaption on fake news detection loss on support set may be misleading for some posts.

Conditional Neural Process (CNP). The CNP includes four components: encoder, feature extractor, aggregator and decoder. The basic idea of conditional neural process is to make predictions with the help of support set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\} = \{x^s_{e,i}, y^s_{e,i}\}_{i=1}^{K}$ as context. The dependence of a CNP on the support set is parametrized by a neural network encoder, denoted as $g(\cdot)$. The encoder $g(\cdot)$ embeds each observation in the support set into feature vector, and the aggregator $\text{agg}(\cdot)$ maps these feature vectors into an embedding of fixed dimension. In CNP, the aggregation procedure is a permutation-invariant operator like averaging or summation. The query data of interest $x^q_{e,i}$ is fed into feature extractor $h(\cdot)$ to get the feature vector. Then the decoder $f(\cdot)$ takes the concatenation of aggregated embedding and given target data $x^q_{e,i}$ as input and output the corresponding prediction as follows:

$$p(y^q_{e,i}|\{\mathbf{X}^s_e, \mathbf{Y}^s_e\}, x^q_{e,i}) = f\big(\text{agg}(g(\{\mathbf{X}^s_e, \mathbf{Y}^s_e\})) \oplus h(x^q_{e,i})\big).$$

where $\oplus$ is concatenation operator.

Limitations of CNP. One widely recognized limitation of CNP is underfitting [16]. For different context data points, their importance is usually different in the prediction. However, the aggregator of CNP treats all the support data equally and cannot achieve query-dependent context information. Moreover, the CNP simply concatenates the input features and numerical label values of posts together as input, ignoring the categorical characteristics of labels.

3 METHODOLOGY

In this paper, we study how to develop an effective model which can identify fake news on emergent events with a small set of labeled data. To this end, we propose a meta neural process framework which can fuse meta-learning and neural process methods together via a simulated task. To tackle the challenges brought by heterogeneous news events, we further propose a label embedding component to handle categorical labels and a hard attention component, which can select the most informative information from the support set with imbalanced class distributions. In the next subsection, we introduce our overall design and architecture.

3.1 Meta-learning Neural Process Design

As shown in Figure 3, our proposed framework includes two stages: event adaptation and detection. The event adaptation stage is to adapt the model parameters to specific event with the help of the support set. The detection stage is to detect fake news on the given event with the help of the support and the adapted parameter set.

Event adaption. We take the i-th support data $\{x^s_{e,i}, y^s_{e,i}\}$ as an example, in the event adaption stage, the $\{x^s_{e,i}, y^s_{e,i}\}$ is used as target data and the rest of support set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\} \setminus \{x^s_{e,i}, y^s_{e,i}\}$ are used as context set accordingly. The context set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\} \setminus \{x^s_{e,i}, y^s_{e,i}\}$ and target data $x^s_{e,i}$ are fed into the proposed model to output the prediction. The loss can be calculated between the prediction $\hat{y}^s_{e,i}$ and the corresponding label $y^s_{e,i}$. For simplicity, we use θ to represent all the parameters included in the proposed model. Then, our event adaption objective function on the support set can be represented as follows:

$$\mathcal{L}^s_e = \sum_i \log p_\theta(y^s_{e,i}|\{\mathbf{X}^s_e, \mathbf{Y}^s_e\} \setminus \{x^s_{e,i}, y^s_{e,i}\}, x^s_{e,i}). \tag{1}$$

We then update parameters θ one or more gradient descent updates on $\mathcal{L}^s_e$ for event e. For example, when using one gradient update:

$$\theta_e = \theta - \alpha \bigtriangledown_\theta \mathcal{L}^s_e. \tag{2}$$

Detection stage. The proposed model with event-specific parameter set θ_e takes query set $\mathbf{X}^q_e$ and entire support set $\{\mathbf{X}^s_e, \mathbf{Y}^s_e\}$ as input and outputs predictions $\hat{\mathbf{Y}}^q_e$ for query set $\mathbf{X}^q_e$. The corresponding loss function in the detection stage can be represented as follows:

$$\mathcal{L}^q_e = \log p_{\theta_e}(Y^q_e|X^s_e, Y^s_e, X^q_e). \tag{3}$$

Through this meta neural process, we can learn an initialization parameter set which can rapidly learn to use given context input-outputs as conditioning to detect fake news on newly arrived events.

Neural Network Architecture. From Figure 3, we can observe that the network structures used in these two stages are the same,

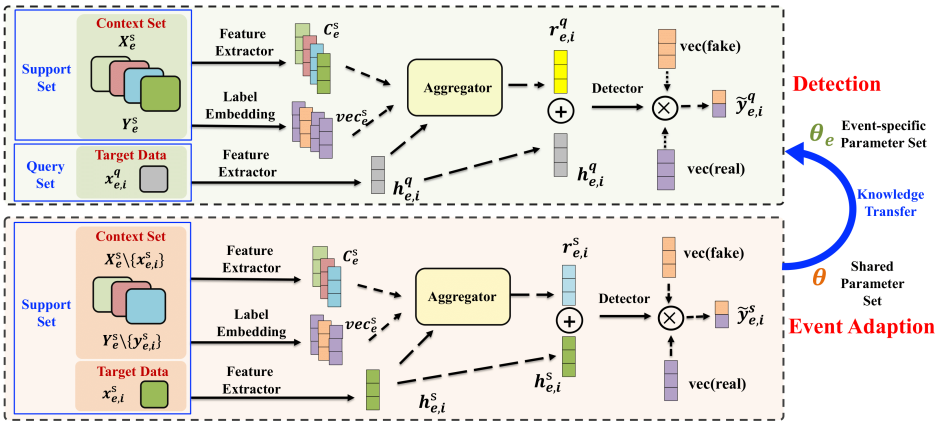


Figure 3: The proposed framework MetaFEND. The proposed framework has two stages: event adaption and detection. During the event adaption stage, the model parameter set θ is updated to event-specific parameter set θ_e. During the detection stage, the event-specific parameter set θ_e is used to detect fake news on event e. $\oplus$ denotes concatenation operation and $\otimes$ means element-wise product.

including feature extractor, label embedding, aggregator and detector. The feature extractor is a basic module which can take posts as input and output corresponding feature vectors. Label embedding component is to capture semantic meanings of labels. Then we use an aggregator to aggregate these information into a fixed dimensional vector, namely context embedding, which is used as reference for fake news detection. Thereafter both the context embedding and target feature vector are fed into detector to output a vector. The final prediction is based on the similarities between this output vector and label embeddings. In the following subsections, we use event adaption to introduce the details of each component in our proposed model. For simplicity, we omitted superscript s and q in the illustrations about components.

3.2 Feature Extractor

From Figure 3, we can observe that feature extractor is a basic module to process raw input. Following the prior works [35, 38], our feature extractor consists of two parts: textual feature extractor and visual feature extractor. For a minor note, the feature extractor is a plug-in component which can be easily replaced by other state-of-the-art models.

Textual feature extractor. We adopt convolutional neural network [17], which is proven effective in the fake news detection [35, 38], as textual feature extractor. The input of the textual feature extractor is unstructured news content, which can be represented as a sequential list of words. For the t-th word in the sentence, we represent it by the word embedding vector which is the input to the convolutional neural network. After the convolutions neural network, we feed the output into a fully connected layer to adjust the dimension to d_f dimensional textual feature vector.

Visual feature extractor. The attached images of the posts are inputs to the visual feature extractor. In order to efficiently extract visual features, we employ the pretrained VGG19 [31] which is used in the multi-modal fake news works [13, 35]. On top of the last layer of VGG19 network, we add a fully connected layer to adjust the dimension of final visual feature representation to the same dimension of textual feature vector d_f. During the joint training process with the textual feature extractor, we freeze the parameters of pre-trained VGG19 neural network to avoid overfitting.

For a multimedia post, we feed the text and image of the example into textual and visual feature extractor respectively. The output of two feature extractors are concatenated together to form a feature vector. For the target data $x_{e,i}$, we denote its feature vector as $\mathbf{h}_{e,i}$. For the context data $x_{e,k}$ where $k \neq i$, we denote its feature vector as $\mathbf{c}_{e,k} \in \mathbf{C}_e$.

3.3 Aggregator

To construct context embedding for target data, we need to design an aggregator which satisfies two properties: permutation-invariant and target-dependent. To satisfy the two properties, we choose to adopt the attention mechanism which can compute weights of each observations in context set with respect to the target and aggregates the values according to their weights to form the new value accordingly.

Attention mechanism. In this paper, we use scaled dot-product attention mechanism [33]. This attention function can be described as mapping a query and a set of key-value pairs to an output, where the query $\mathbf{Q}$, keys $\mathbf{K}$, values $\mathbf{V}$, and output are all vectors. In our problem, for the target data $x_{e,i}$ and the context set $\mathbf{X}_e \setminus \{x_{e,i}\} = \{x_{e,k}\}_{k=1,k\neq i}^{K}$ on event e. We use the the target feature vector $\mathbf{h}_{e,i} \in \mathbb{R}^{1\times d}$ after linear transformation as query vector $\mathbf{Q}_i$, the context feature vector $\mathbf{C}_e = [c_{e,1}, ..., c_{e,K}] \in \mathbb{R}^{K\times d}$ after linear transformation as the Key vector $\mathbf{K}$. For the context set, we represent its label information $\mathbf{Y}_e \setminus \{y_{e,i}\} = \{y_{e,k}\}_{k=1,k\neq i}^{K}$ by semantic embeddings as $\mathbf{vec}_e = \{\mathbf{vec}_{e,k}\}_{k=1,k\neq i}^{K}$. The details of label embedding are introduced in the next subsection. Then we concatenate context feature vector and label embedding as $\mathbf{C_e} \oplus \mathbf{vec}_e = [\mathbf{c}_{e,1} \oplus \mathbf{vec}_{e,1}, ..., \mathbf{c}_{e,K} \oplus \mathbf{vec}_{e,K}] \in \mathbb{R}^{(K-1)\times 2d}$. The concatenated embedding after linear transformation is used as value vector $\mathbf{V}$. We represent $\mathbf{Q}_i, \mathbf{V}, \mathbf{K}$ as follows:

$$\mathbf{Q}_i = \mathbf{W}_q \mathbf{h}_{e,i},$$

$$\mathbf{K} = \mathbf{W}_k \mathbf{C}_e,$$

$$\mathbf{V} = \mathbf{W}_v(\mathbf{C}_e \oplus \mathbf{vec}_e),$$

where $\mathbf{W}_q \in \mathbb{R}^{d\times d}$, $\mathbf{W}_k \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_v \in \mathbb{R}^{2d\times d}$.

The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by dot-product function of the query with the corresponding key. More specifically, attention function can be represented as follows:

$$\mathbf{a}_i = \text{softmax}(\frac{\mathbf{Q}_i\mathbf{K}^T}{\sqrt{d}}) \tag{4}$$

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}) := a_i\mathbf{V}. \tag{5}$$

Limitation of Soft-Attention. The attention mechanism with soft weight values is categorized into soft-attention. However, soft-attention cannot effectively trim irrelevant data especially when we have a context set with an imbalanced class distribution shown in Fig. 2. Moreover, we show a case study in the experimental section for a better illustration.

Hard-Attention. To overcome the limitation of soft-attention, we propose to select the most related context data point instead of using weighted average. To enable argmax operation to be differentiable, we use Straight-Through (ST) Gumbel SoftMax [12] for discretely sampling the context information given target data. We introduce the sampling and arg max approximations of ST Gumbel SoftMax procedure next.

The Gumbel-Max trick [9] provides a simple and efficient way to draw samples z from a categorical distribution with class probabilities. In our problem, for the i-th target data point $x_{e,i}$ with context set $\mathbf{X}_e \setminus \{x_{e,i}\} = \{x_{e,k}\}_{k=1,k\neq i}^{K}$, the class probabilities can be obtained from the weight vector $\mathbf{a}_i = [a_{i,1}, ..., a_{i,K}]$ from dot-product attention mechanism according to Eq. 4. Because arg max operation is not differentiable, we use the softmax function as a continuous, differentiable approximation to arg max, and generate K-dimensional sample vectors $\mathbf{P}_i = [p_{i,1}, p_{i,2}.., p_{i,K}]$ as follows:

$$p_{i,k} = \frac{\exp((\log(a_{i,k}) + g)/\tau)}{\sum_{k,k\neq i}^{K}\exp((\log(a_{i,k}) + g)/\tau)} \tag{6}$$

where τ is a temperature parameter, $g = -\log(-\log(\mu))$ is the Gumbel noise and μ is generated by a certain noise distribution (e.g., $u \sim \mathcal{N}(0, 1)$). As the softmax temperature τ approaches 0, the Gumbel-Softmax distribution becomes identical to the categorical distribution. Moreover, Straight-Through (ST) gumbel-Softmax takes different paths in the forward and backward propagation, so as to maintain sparsity yet support stochastic gradient descent. Through gumbel-softmax, the hard-attention mechanism is able to draw the most informative sample based on weight vectors from $\mathbf{P}_i$ for given target sample $x_{e,i}$.

The hard-attention can trim the irrelevant data points and select the most related data point, denoted as $\mathbf{c}_{e,k} \oplus \mathbf{v}_{e,k} \in \mathbb{R}^{2d}$. Besides the hard-attention mechanism, the aggregator includes an additional fully connected layer on top of hard-attention to adjust the dimension. The $\mathbf{c}_{e,k} \oplus \mathbf{v}_{e,k}$ is fed into this fully connected layer to output context embedding $\mathbf{r}_{e,i} \in \mathbb{R}^d$.

3.4 Detector based on Label Embedding

Categorical characteristic of label information. The context information includes posts and their corresponding labels. The existing works like CNP [7] and ANP [16] usually simply concatenate the input features and numerical label values together as input to learn a context embedding via a neural network. Such operation discards the fact that label variables are categorical. Moreover, this operation tends to underestimate the importance of labels as the dimension of input features is usually significantly larger than that of single dimensional numerical value. To handle categorical characteristic, we propose to embed labels into fixed dimension vectors inspired by word embedding [24]. We define two embeddings $\mathbf{vec}(\mathbf{fake})$ and $\mathbf{vec}(\mathbf{real})$ for the labels of fake news and real news respectively. For example, given the k-th post $x_{e,k}$ on event e, the corresponding label is fake and its label embedding vector is $\mathbf{vec}(\mathbf{fake})$, and we denote the label embedding of $x_{e,k}$ as $\mathbf{vec}_{e,k}$. To ensure that the label embedding can capture the semantic meanings of corresponding labels, we propose to use embeddings $\mathbf{vec}(\mathbf{fake})$ and $\mathbf{vec}(\mathbf{real})$ in the detector as metrics and output predictions are determined based on metric matching.

The detector is a fully-connected layer which takes target feature vector and context embedding as inputs and outputs a vector that has the same dimensionality as that of the label embedding. More specifically, for i-th target data, the context embedding $\mathbf{r}_{e,i}$ and target feature vector $\mathbf{h}_{e,i}$ are concatenated. Then the detector takes $\mathbf{r}_{e,i} \oplus \mathbf{h}_{e,i} \in \mathbb{R}^{2d}$ as input and produces a output vector $\mathbf{o}_{e,i} \in \mathbb{R}^d$. The similarities between output $\mathbf{o}_{e,i}$ from our model and label embeddings $\mathbf{vec}(\mathbf{fake})$ and $\mathbf{vec}(\mathbf{real})$ are calculated as follows:

$$\text{similarity}(\mathbf{o}_{e,i}, \mathbf{vec}(\mathbf{fake})) = \left\|\mathbf{o}_{e,i} \circ \mathbf{vec}(\mathbf{fake})\right\|, \tag{7}$$

$$\text{similarity}(\mathbf{o}_{e,i}, \mathbf{vec}(\mathbf{real})) = \left\|\mathbf{o}_{e,i} \circ \mathbf{vec}(\mathbf{real})\right\|. \tag{8}$$

The two similarity scores are then mapped into $[0, 1]$ as probabilities via *softmax*. The trainable label embedding capture semantic meaning of labels and can generalize easily to new events with the help of adaptation step according to Eq. 2.

3.5 Algorithm Flow

After introducing the meta-learning neural process design, feature extractor, label embedding, aggregator and detector components, we present our algorithm flow.

As it can be observed from Figure 3, when tackling an event e, our proposed framework MetaFEND has two stages: event adaption and detection. In more details, our proposed model adapts to the specific event according to Eq. 2 and then the event-specific parameter is used in the fake news detection on given event. The algorithm flow is same in the two stages and we use event adaption stage as an example to illustrate this procedure.

Our input includes handful instances as context set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\} \setminus \{x_{e,i}^s, y_{e,i}^s\}$ and $x_{e,i}^s$ as target data. We first feed $\mathbf{X}_e^s \setminus \{x_{e,i}^s\}$ into feature extractor and get context feature representations $\mathbf{C}_e^s$. The context feature representations $\mathbf{C}_e^s$ is then concatenated with label embedding $\mathbf{vec}_e^s$ of $\mathbf{Y}_e^s$. In the target side, the target data $x_{e,i}^s$ is also fed into feature extractor to get representation as $\mathbf{h}_{e,i}^s$. The aggragator component aggregates $\mathbf{h}_{e,i}^s$, $\mathbf{C}_e^s$ and $\mathbf{vec}_e^s$ as introduced in section 3.3 to output context embedding $\mathbf{r}_{e,i}^s \in \mathbb{R}^d$. Then we concatenate $\mathbf{r}_{e,i}^s$ with target feature vector $\mathbf{h}_{e,i}^s \in \mathbb{R}^d$. The concatenated feature goes through the detector which is consisted of a fully connected layer to output a vector $\mathbf{o}_{e,i}^s$. The similarity scores between $\mathbf{o}_{e,i}^s$ and $\mathbf{vec}(\mathbf{fake})$, $\mathbf{vec}(\mathbf{real})$ are calculated according to

Eq. 7 and Eq. 8 respectively. In the end, the similarity scores are mapped to probability values for fake news detection via softamax operation.

4 EXPERIMENTS

In this section, we introduce the datasets used in the experiments, present the compared fake news detection models, validate the effectiveness and explore some insights of the proposed framework.

4.1 Datasets

To fairly evaluate the performance of the proposed model, we conduct experiments on datasets collected from two real-world social media datasets, namely Twitter and Weibo. The detailed description of the datasets are given below:

Table 1: The Statistics of the Datasets.

	Twitter	Weibo
# of fake News	6,934	4,050
# of real News	5,683	3,558
# of images	514	7,606

The **Twitter dataset** is from MediaEval Verifying Multimedia Use benchmark [2], which is used in [13, 35] for detecting fake content on Twitter. The **Weibo dataset**[1] is used in [13, 27, 35] for detecting multi-modal fake news. The news events are included in the Twitter dataset and we follow the previous works [13, 27, 35] to obtain events on Weibo via a single-pass clustering method [14]. In the two datasets above, we only keep the events which are associated with more than 20 posts and randomly split the posts on same event into support and query data. To validate performance of the models on newly emergent events, we ensure that the training and testing sets do not contain any common event. We adopt Accuracy and F1 Score as evaluation metrics. These two datasets cover diverse news events and thus can be used as good test-grounds for evaluation of fake news detection on heterogeneous events.

4.2 Baselines

To validate the effectiveness of the proposed model, we choose baselines from multi-modal models and the few-shot learning models. For the multi-modal models, we fine-tune them on support set from events in the testing data for a fair comparison. In the experiments, we have the 5-shot and 10-shot settings. In our problem, 5-shot setting refers to that 5 labeled posts are provided as support set.

Fine-tune models. All the multi-modal approaches take the information from multiple modalities into account, including VQA [1], att-RNN [13] and EANN [35]. In the fine-tune setting, the training data including labeled support data and labeled query data is used to train the baselines. In the testing stage, the trained models are first fine-tuned on the labeled support data of given event, and then make predictions for testing query data. (1) **VQA** [1]. Visual Question Answering (**VQA**) model aims to answer the questions based on the given images and is used as a baseline for multimodal fake news in [13]. (2) **att-RNN** [13]. att-RNN is the state-of-the-art model for multi-modal fake news detection. It uses attention mechanism to fuse the textual, visual and social context features. In our experiments, we remove the part dealing with social context information, but the remaining parts are the same. (3) **EANN** [35]. EANN is one of the state-of-the-art models for fake news detection. It consists of three components: feature extractor, event discriminator and fake news detector. It captures shared features across different events of news to improve generlziation ability.

Few-shot learning models. We use CNP [7], ANP [16], MAML [6] and Meta-SGD [19] as few-shot learning baselines. (1) **CNP** [7]. Conditional neural process is the state-of-the-art model for few-shot learning. It combines neural network and gaussian process by using a small set of input-output pairs as context to output predication for given input of data. (2) **ANP** [16]. Attentive neural process belongs to the family of neural process which outputs prediction based on concatenation of learned distribution of context, context features and given input. (3) **MAML** [6]. Model-aganostic Meta-learning is a representative optimization-based meta-learning model. The mechanism of MAML is to learn a set of shared model parameters across different tasks which can rapidly learn novel task with a small set of labeled data. (4) **Meta-SGD** [19]. Meta-SGD is one of the state-of-the-art meta learning method for few-shot learning setting. Besides a shared global initialized parameters as with MAML, it also learns step sizes and update direction during the training procedure.

The proposed model share the same feature extractor backbone with EANN, CNP, ANP, MAML, Meta-SGD to study the effects of other designs in addition to benefits of the feature extractor backbone.

Implementations In the proposed model, the 300 dimensional FastText pre-trained word-embedding weights [3] are used to initialize the parameters of the embedding layer. The window size of filters varies from 1 to 5 for textual CNN extractor. The hidden size d_f of the fully connected layer in textual and visual extractor and dimension d are set as 16 which is searched from options {8, 16, 32, 64}. τ decays from 1 to 0.5 as the suggested way in [12]. The gradient update step is set to 1 an inner learning rate β is set to 0.1 for fine-tune models: MAML, Meta-SGD and our proposed framework MetaFEND. We implement all the deep learning baselines and the proposed framework with PyTorch 1.2 using NVIDIA Titan Xp GPU. For training models, we use Adam [18] in the default setting. The learning rate α is 0.001. We use mini-batch size of 10 and training epochs of 400.

4.3 Performance Comparison

Table 2 shows the performance of different approaches on the Twitter and Weibo datasets. We can observe that the proposed framework MetaFEND achieves the best results in terms of most of the evaluation metrics in both 5-shot and 10-shot settings.

Twitter. On the Twitter dataset in 5-shot setting, compared with CNP, ANP incorporates the attention mechanism and hence can achieve more informative context information. Due to the heterogeneity of events, it is not easy for Meta-SGD to learn a shareable learning directions and step size across all events. Thus, Meta-SGD's performance is lower than MAML's in terms of accuracy. Compared with all the baselines, MetaFEND achieves the best performance in terms of most the metrics. Our proposed model inherits the advantages of MAML to learn a set of parameters which can rapidly learn to detect fake news with a small support set. Moreover, MetaFEND

[1] https://github.com/yaqingwang/EANN-KDD18

Table 2: The performance comparison of models for fake news detection on the Twitter and Weibo datasets under 5-shot and 10-shot settings. Accuracy and F1 score of models are followed by standard deviation. The percentage improvement (↑) of MetaFEND over the best baseline per setting is in the last row. EANN, CNP, ANP, MAML, Meta-SGD and MetaFEND share the same feature extractor as the backbone.

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND	**86.45 ± 1.83**	**86.21 ± 1.32**	**88.79 ± 1.27**	**88.66 ± 1.09**	**81.28 ± 0.75**	**80.19 ± 1.27**	**82.92 ± 0.13**	**82.37 ± 0.28**
(Improvement)	**(↑5.12%)**	**(↑3.91%)**	**(↑4.19%)**	**(↑4.33%)**	**(↑4.41%)**	**(↑4.13%)**	**(↑5.22%)**	**(↑5.51%)**

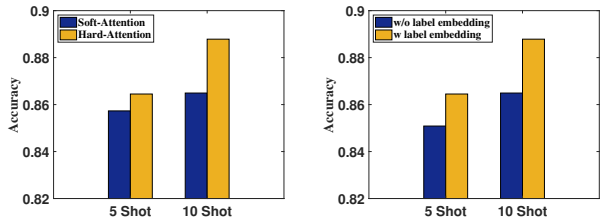
can use the support data as conditioning set explicitly to better capture the uncertainty of events and thus it is able to achieve more than 5% improvement compared with MAML in terms of accuracy. In the 10-shot setting, as the size of give support data increases, the soft attention mechanism of ANP unavoidably incorporates the irrelevant data points. In contrast, the proposed model MetaFEND employs the hard-attention mechanism to trim irrelevant data points from context set and significantly outperforms all the baselines in terms of all the metrics.

Weibo. Compared with the Twitter data, the Weibo dataset has different characteristics. On the Weibo dataset, most of the posts are associated with different images. Thus, we can evaluate the performance of models under the circumstance where support datasets do not include direct clues with query set. As EANN tends to ignore event-specific features, it achieves the lowest accuracy among finetune models in 10-shot setting. For the few-shot models, ANP and CNP achieves better performance compared with gradient-based meta-learning methods MAML and Meta-SGD. This is because the parameter adaptation may not be effective when support data set and query set do not share the same patterns. Compared with ANP in 5-shot setting, our proposed method MetaFEND achieves 4.39% improvement in terms of accuracy and 5.51% improvement in terms of F1 score. The reason is that our MetaFEND can learn a base parameter which can rapidly learn to use a few examples as reference information for fake news detection. Thus, our proposed model enjoys the benefits of neural process and meta-learning model families.

4.4 Ablation Study

We show ablation study to analyze the role of Hard-Attention and label embedding components.

Soft-Attention v.s. Hard-Attention. To intuitively illustrate the role of hard-attention mechanism in the proposed model, we show ablation study by replacing hard-attention with soft-attention. Then we repeatedly run the new designed model on the Twitter dataset five times in 5-shot and 10-shot settings respectively and report the average of accuracy values. The results are show in the Figure 4. From Figure 4a, we can observe that accuracy scores of "Hard-Attention" in 5-shot and 10-shot settings are greater than those of "Soft-Attention" respectively. As the number of support set increases, hard-attention mechanism does not have the limitation of soft-attention mechanism which unavoidably incorporates unrelated data points and significantly outperforms the soft-attention in terms of accuracy score. Thus, we can conclude that hard-attention mechanism can take effectively advantage of support set, and the superiority is more significant as we enlarge size of support set.



(a) Attention **(b) Label Embedding**

Figure 4: The ablation study about (a) Soft-Attention and Hard-Attention and (b) Label Embedding.

w/o Label Embedding v.s. w/ Label Embedding. To analyze the role of label embedding in the proposed model, we design MetaFEND's corresponding reduced model by replacing label embedding with label value 0 or 1. Accordingly, we change the multiplication between output with label embedding to a binary-class fully connected layer to directly output the probabilities. Figure 4b shows the results in terms of accuracy score. In Figure 4b, "w/o label embedding" denotes that we remove the label embedding, and "w label embedding" denotes the original approach. We can observe that the accuracy score of "w label embedding" is greater than "w/o label embedding" in 5-shot and 10-shot settings, demonstrating the effectiveness of label embedding

4.5 Case Study

In order to illustrate the challenges of emergent fake news detection and how our model handles challenges, we show one example in 5-shot learning setting as case study in Fig. 5. As it can be observed, the four of five news examples in the support set are real news. Due to imbalanced class condition in the support set, it is difficult for Soft-Attention to provide correct prediction for news of interest in the query set. More specifically, Fig. 5 shows the attention score values (red color) between examples in support set and query set based on multi-modal features. Although the first example with

largest attention score value is most similar to news example in the query set, the majority of context information is from the other four examples due to imbalanced class distribution. Such an imbalanced class distribution leads to incorrect prediction for Soft-Attention. The Hard-Attention mechanism can achieve correct result by focusing on the most similar sample in the support set. Through this example, we can also observe the necessity of event adaption stage. The posts and images for the same event are very similar and difficult to distinguish. Without event adaption stage, the model cannot capture informative clues to make correct predictions.

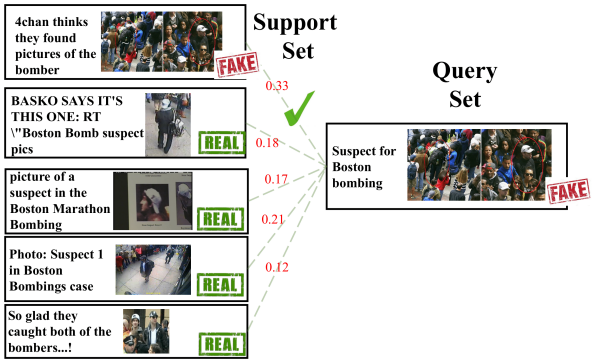


Figure 5: Fake news examples missed by Soft-Attention but spotted by Hard-Attention

5 RELATED WORK

In this section, we briefly review the work related to the proposed model from fake news detection and few-shot learning.

5.1 Fake News Detection

Many fake news detection algorithms try to distinguish news according to their features, which can be extracted from social context and news content. (1) *Social context features* represent the user engagements of news on social media [30] such as the number of followers, hash-tag (#), propagation patterns [39] and retweets. However, social context features are very noisy, unstructured and labor intensive to collect. Especially, it cannot provide sufficient information for newly emerged events. (2) *Textual features* are statistical or semantic features extracted from text content of posts, which have been explored in many literatures of fake news detection [4, 10, 30]. Unfortunately, linguistic patterns are not yet well understood, since they are highly dependent on specific events and corresponding domain knowledge [28]. To overcome this limitation, approaches like [20–23, 26] propose to use deep learning models to identify fake news and have shown the significant improvements. (3) *Visual features* have been shown to be an important indicator for fake news detection [15, 30]. The basic features of attached images in the posts are explored in the work [11, 15, 25].

In this paper, we consider multi-modal features when identifying fake news on social media. To tackle *multi-modal fake news detection*, in [13], the authors propose a deep learning based fake news detection model, which extracts the multi-modal and social context features and fuses them by attention mechanism. To detect fake news on never-seen events, Wang et al. [35] propose an event-adversarial neural network (EANN) which can capture event-invariant features for fake news detection. However, EANN cannot take advantage of a small set of labeled data to further capture event specification and thus is not suited for our task.

5.2 Few-Shot Learning

Meta-learning has long been proposed as a form of learning that would allow systems to systematically build up and re-use knowledge across different but related tasks [34, 36, 37]. MAML [6] is to learn model initialization parameters that are used to rapidly learn novel tasks with a small set of labeled data. Following this direction, besides initialization parameters, Meta-SGD [19] learns step sizes and updates directions automatically in the training procedure. As tasks usually are different in the real setting, to handle task heterogeneity, HSML [40] customizes the global shared initialization to each cluster using a hierarchical clustering structure. The event heterogeneity is widely observed for fake news detection, where nonexistence of hierarchical relationship in news events makes this task more challenging.

Neural process approaches [7, 8, 16] combine stochastic process and neural network to handling task heterogeneity by conditioning on a context set. Conditional Neural Process (CNP) [7] and Neural Process (NP) [8] use neural networks to take input-output pairs of support set as conditioning for inference, incorporating task specific information. However, these two works aggregate the context set by average or sum, ignoring different importance among context data samples and thereby leading to unsatisfactory performance. Attentive Neural Process (ANP) [16] incorporates attention mechanism into Neural Process to alleviate such a issue. However, ANP still suffers from underfitting issue due to fixing parameters for different tasks. Additionally, ANP directly concatenates the label numeric values with feature representation, discarding the categorical characteristics of label information.

Different from existing works, our proposed framework maintains the parameter flexibility following the principle of meta-learning and inherits generalization ability to handle event heterogeneity from neural processes. Moreover, we incorporate label embedding component to handle categorical characteristics of label information and utilize hard attention to extract most informative context information. Thus, our proposed model enjoys the benefits of two model families without suffering their limitations.

6 CONCLUSIONS

In this work, we study the problem of fake news detection on emergent events. The major challenge of fake news detection stems from newly emerged events on which existing approaches only showed unsatisfactory performance. In order to address this issue, we propose a novel fake news detection framework, namely MetaFEND, which can rapidly learn to detect fake news for emergent events with a few labeled examples. The proposed framework can enjoy the benefits of meta-learning and neural process model families without suffering their own limitations. Extensive experiments on two large scale datasets collected from popular social media platforms show that our proposed model MetaFEND outperforms the state-of-the-art models.

7 IMPACT STATEMENT

Fake news can manipulate important public events and becomes a global concern. If the fake news detection algorithm can function as intended, it is beneficial to prevent the spread of fake news in the early stage and correspondingly many negative public events caused by fake news may be avoided. However, we are also aware that automatic detection may suppress the public discussion. The failure modes may lie in the negation cases: if someone tries to spot the fake news by citing false information contents, the automatic algorithm may not understand the logic behind the post and incorrectly identify it as fake news. The bias may be unavoidable included in the dataset especially when the events are controversial or lacking a clear standard for annotation. Our proposed model explicitly uses the labeled sample as reference information and thus it is possible to replace the incorrect annotated support set by correct ones to correct the bias. To reduce harm brought by the automatic algorithm, both technology and human review are needed and an effective user appeal system should be employed in case the incorrect detection happened.

ACKNOWLEDGMENT

The authors thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grants NSF IIS-1553411 and IIS-1956017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[2] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying Multimedia Use at MediaEval 2015.. In *MediaEval*.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

[5] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1126–1135.

[7] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. 2018. Conditional Neural Processes. In *International Conference on Machine Learning*. 1704–1713.

[8] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. 2018. Neural processes. *arXiv preprint arXiv:1807.01622* (2018).

[9] Emil Julius Gumbel. 1948. *Statistical theory of extreme values and some practical applications: a series of lectures*. Vol. 33. US Government Printing Office.

[10] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.

[11] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 153–164.

[12] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[13] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 795–816.

[14] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining*. IEEE, 230–239.

[15] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19, 3 (2017), 598–608.

[16] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. 2019. Attentive neural processes. *arXiv preprint arXiv:1901.05761* (2019).

[17] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

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

[19] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835* (2017).

[20] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. *arXiv preprint arXiv:2004.11648* (2020).

[21] 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.. In *IJCAI*. 3818–3824.

[22] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the The Web Conference 2018*. 585–593.

[23] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*. 3049–3055.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[25] Dong ping Tian et al. 2013. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering* 8, 4 (2013), 385–396.

[26] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* (2018).

[27] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. *arXiv preprint arXiv:1908.04472* (2019).

[28] 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*. ACM, 797–806.

[29] Hua Shen, Fenglong Ma, Xianchao Zhang, Linlin Zong, Xinyue Liu, and Wenxin Liang. 2017. Discovering social spammers from multiple views. *Neurocomputing* 225 (2017), 49–57.

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

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

[32] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[34] Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review* 18, 2 (2002), 77–95.

[35] 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*. 849–857.

[36] Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive Self-training for Few-shot Neural Sequence Labeling. *arXiv preprint arXiv:2010.03680* (2020).

[37] Yaqing Wang, Yifan Ethan Xu, Xian Li, Xin Luna Dong, and Jing Gao. 2020. Automatic Validation of Textual Attribute Values in E-commerce Catalog by Learning with Limited Labeled Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[38] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[39] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 651–662.

[40] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. 2019. Hierarchically structured meta-learning. *arXiv preprint arXiv:1905.05301* (2019).