IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Multi-modal Meta Multi-Task Learning for Social Media Rumor Detection

Jelinasri M

Computer Science Engineering,Madha Engineering College,Tamilnadu ,India.

Abstract—With the rapid development of social media platforms and the increasing scale of the social media data, the rumor detection task has become vitally important since the authenticity of posts cannot be guaranteed. To date, many approaches have been proposed to facilitate the rumor detection process by utilizing the multi-task learning mechanism, which aims to improve the performance of rumor detection task by leveraging the useful information contained in stance detection task. However, most of the existing approaches suffer from three limitations: (1) only focus on the textual content and ignore the multi-modal information which is key component contained in social media data; (2) ignore the difference of feature space between the stance detection task and rumor detection task, resulting in the unsatisfactory usage of stance information; (3) largely neglect the semantic information hidden in the finegrained stance labels. Therefore, in this paper, we design a Multi-modal Meta Multi-Task Learning (MM-MTL) framework for social media rumor detection task. To make use of multiple modalities, we design a multi-modal post embedding layer which considers both textual and visual content. To overcome the feature sharing problem of the stance detection task and rumor detection task, we propose a meta knowledge-sharing scheme to share some higher meta network layers and capture the metaknowledge behind the multi-modal post. To better utilize the semantic information hidden in the fine-grained stance labels, we employ the attention mechanism to estimate the weight of each reply. Extensive experiments on two Twitter benchmark datasets demonstrate that our proposed method achieves state-of-the-art performance.

Index Terms—Social Media, Rumor Detection, Meta Learning, Multi-task Learning, Multi-Modal

1. Introduction

Online social media platforms have become the most important medium for people to share, coordinate, and spread information. Unlike the traditional media, where news is published by reputable organizations, online news on social media platforms is released and shared by hundreds of millions of users spontaneously. However, few users carefully check the authenticity of the information they share, which means large volumes of rumors may emerge and spread. Without an accurate and systematic effort to verify the posts, the dissemination of social media rumors may cause large-scale negative effects and sometimes may affect or even manipulate critical public events. Therefore, how to effectively detect misinformation and minimize its negative impact has become a significant challenge faced by social media platforms. To minimize the harmful effects of rumors, many efforts have been made. The early efforts come from news websites, such as snopes.com and politifact.com, which try to expose or confirm rumors by expert analysis and crowdsourcing. However, manually collecting and investigating rumors is quite time-consuming and has obvious limitations on efficiency. Thus, automatically mining and detecting rumors has drawn much attention in the research community. Basically, existing studies on automatic rumor detection can be summarized into two categories: (1) The first category is to extract or construct comprehensive and complex features with manual ways. For example, Castillo et al. design plenty of handcrafted features from the media content of posts and the social context of users, then use these features to train a support vector machine. (2) The second category is to automatically capture deep features based on neural networks. For example, Ma et al. introduce a recurrent neural network to learn the hidden representations from the text content of relevant posts. Yu et al. use a convolutional neural network to obtain key features and their high-level interactions from IEEE Transaction on Multimedia,Issue Date:Jan.2022 2 the text content of the claims. Although these algorithms show promising performance in rumor detection, most of these methods only focus on the text content. In fact, the content of the post in social media platforms may consist of multiple modalities (e.g., text, images), and these multiple modality information can complement each other. Moreover, the tweets in rumor detection tasks are all posted by users, and the user's stance can play important roles for rumor detection. Therefore, it is critical and important to exploit multimedia content and the user's stance for rumor detection. Recently, a novel multi-task learning method is proposed to introduce the stance information of users' replies to the rumor detection task. Ma et al. propose a novel shared-private multi-task learning method to model information sharing and representation reinforcement between the stance detection task and the rumor detection task, which can expand valuable features for their respective tasks. It is observed that doubtful and opposing voices against rumors always arise along with its propagation, which can serve as helpful indicators that signal the truthfulness of the information. Although it seems reasonable to improve the performance of social media rumor detection by identifying rumors and jointly analyzing various stances, these existing methods still cannot effectively address the following three challenges. 1) Feature level challenge: The data of the rumor detection task and stance detection task are multi-modal. The content of the post in social media platforms may consist of multiple modalities (e.g., text, images), and these multiple modality information can complement each other. For example, in the root post of Table I, images in the post also convey important information. However, most approaches only focus on the textual content of the post. 2) Meta level challenge: Unlike the generic multi-task learning methods which consist of tasks with only one type, the rumor detection task and the stance detection task belong to different categories. As shown in Table I, the rumor detection task is a classification problem, while the stance detection task is a sequence labeling problem. Their output structures vary widely between the two tasks. Therefore, the rumor detection task and stance detection task lie differently in the feature space. However, existing methods mechanically apply the generic feature-sharing multitask learning method to rumor detection. The shared features in the shared layer are equally sent to their respective tasks, which causes some useless and even adverse features being mixed in different tasks. 3) Task level challenge: Existing approaches typically introduce the stance information in the stance labels into the rumor detection task through the backpropagation. However, the semantic information hidden in the finegrained stance labels is largely ignored. As indicated in Table I, the data in the stance detection task are annotated at the post level. The fine-grained tweet level labels can effectively promote the performance of the rumor detection task. For example, replies with strong support or denial stance have a greater impact on the veracity of the rumor prediction. Thus, posts with different stance labels should have different weights to obtain a comprehensive representation of the sequence for rumor detection. In this paper, we aim to tackle the above issues by introducing a novel multi-task learning method: Multi-modal Meta Multi-Task Learning (MM-MTL). The advantages of MM-MTL are three-folds: (1) To make use of multiple modalities, we design a multi-modal post embedding layer which considers both textual and visual content. (2) To overcome the problem faced by feature-sharing multi-task learning methods, we propose a knowledge-sharing scheme for multi-task learning. Instead of sharing some lower layers to extract common features across tasks, the proposed meta multi-task learning shares some higher layers: meta network, which learns mutual meta-knowledge from various tasks. The meta-knowledge is then used to dynamically generate the parameters of the task specific models. Thus, the stance information of user's replies is effectively introduced into the rumor detection. (3) To better exploit the semantic information hidden in the fine grained stance labels, we employ an attention mechanism to estimate the weight of each reply and explicitly include the hidden states from the stance layer in the attention calculation. Therefore, the rumor detection performance is further boosted. Extensive experiments on two Twitter benchmark datasets demonstrate that our rumor detection model outperforms the state-of-the-art methods. The main contributions of our paper are summarized as follows: • We propose a multi-modal meta multi-task learning method for social media rumor detection. Different from the generic multi-task learning methods which share lower common features, the proposed method shares the higher meta-knowledge from rumor detection and stance tasks. With the guide of meta-knowledge, the task-specific model can obtain a precise multi-modal representation of every post. • We apply an attention mechanism to the multi-modal meta multi-task learning to fully utilize the stance information of user replies. With the weighted user responses, the performance of the proposed multi-modal meta multitask learning can be further boosted. • We experimentally demonstrate that our model is more robust and effective than state-of-the-art based on two public benchmark datasets for rumor detection tasks on Twitter.

2. RELATED WORKS

A. Social Media Rumor Detection

Online social media platforms have become the most important medium for people to share information. Hundreds of millions of users on social platforms create a huge scale of social media data. These social media data have great research values and draw much attention in the research community. Various studies are proposed to explore social media, such as IEEE Transaction on Multimedia, social media analysis, social events understanding, the cyberbullying phenomena understanding, multimedia summarization on microblogs, election prediction , visual concept learning , opinion mining and multimodal data learning. However, many models are suffering the error caused by the misinformation in social media data. To debunk rumors and minimize their harmful effects, many efforts have been made. Existing work regards social media rumor detection as a supervised classification problem. The main concern of the supervised classification approach is to define effective features for training rumor classifiers. Early methods design plenty of hand-crafted features to debunk rumors. For example, Castillo et al. provided a wide range of features crafted from the post contents, user profiles, and propagation patterns. Instead of defining complex feature sets, data-driven models are proposed to obtain state-of-the-art detection performance. For example, Ma et al. propose a recurrent neural network to learn the hidden representations from the text content of claims. Recently, several approaches conduct rumor detection based on the multimedia content. Jin et al. propose a recurrent neural network with an attention mechanism to fuse image and text features of the post for rumor detection. Although some good performances have been achieved, existing methods mainly focus on capturing the content features of post, while always ignore the strong connections between the veracity of claim and the stances expressed in responsive posts. However, the stance information expressed by users toward a particular rumor can be indicative of the veracity. In this work, we introduce a novel multi-modal meta multitask learning method to improve the performance of the rumor detection task by jointly training the related stance detection task.

B. Multi-task Learning

Multi-task learning aims to improve the performance of one task by using other related tasks. Most of the multi-task learning or joint learning models can be regarded as parameter sharing approaches, where models are trained jointly, and parameters or features are shared across multiple tasks. Multi-task learning has been widely used in the various tasks of natural language processing and achieved excellent results. For example, Collobert et al. propose a unified framework which uses a shared lookup table for input words, and then jointly train several NLP tasks using convolutional neural networks. Song et al. propose a multi-source multi-task learning scheme to co-regularize the source consistency and the tree-guided task relatedness for user interest inference. In most of these models, multi-task architectures use the shared private schema to share features across tasks, which divides the features of different tasks into private and shared spaces, and the task-irrelevant features in shared space are used as extra features for various tasks. Recently, several multi-task learning methods are proposed to improve the performance of rumor detection by jointly train rumor detection task and stance detection task. For example, Ma et al. introduce a GRU-based multi-task learning method that shares features across the rumor detection task and stance detection task. In this paper, we propose a multi-modal meta multi-task learning method for social media rumor detection. Different from the above multi-task learning methods which share features in various tasks, our model shares the meta-level knowledge. Specifically, a meta network is proposed to capture the meta-knowledge across tasks and control the parameters of task-specific networks.

C. Meta Learning

Meta learning, also known as “learning to learn”, intends to design models that can learn new skills or adapt to new environments rapidly with a few training examples. There are three common approaches: 1) metric-based approaches which learn an efficient distance metric; 2) model-based approaches which use recurrent network with external or internal memory to enhance the generalization; 3) optimization-based approaches which optimize the model parameters explicitly for fast learning. In this subsection, we briefly review the model-based meta learning methods, which use a small hyper meta-networks to predict the weight parameters of a large backbone network. Since the introduction of hyper meta-networks, it has found wide applications in neural architecture search, Bayesian neural networks, network pruning and also multi-task learning. In this paper, we propose a new design of a hyper meta network based multi-task learning for rumor detection, which makes the meta knowledge sharing possible for rumor related tasks, and enhance both tasks by the dynamic weight generation mechanism. Different from the pervious meta multi-task learning methods in this work, we aim to deal with heterogeneous multi-modal rumor related tasks where input and output structures vary widely among the tasks. Furthermore, we extend the meta multi-task learning with the attention mechanism, which can further utilize the multigranularity label information of the heterogeneous rumor related tasks. To the best of our knowledge, no prior work studies our problem from a hyper meta-network perspective.

3. PROBLEM STATEMENT

Our goal is to formulate a multi-task model that jointly learns the rumor detection and stance detection models. We model the Twitter data as a set of claims C = {c1, c2, · · · , cn}, where each claim c = {p0, p1, · · · , pT } is a sequence composed of a set of relevant tweets. T is the sequence length. Specifically, p0 is the root post, and {p1, · · · , pT } is the set of user replies. The post pt consists of text content xt and visual content vt.

a) Rumor Detection: Following the previous work [7], [12], [15], we define the rumor detection as a binary classification problem, which aims to determine the veracity value of claims in social media. Formally, rumor detection aims to learn a projection:

Fr(c) → Y

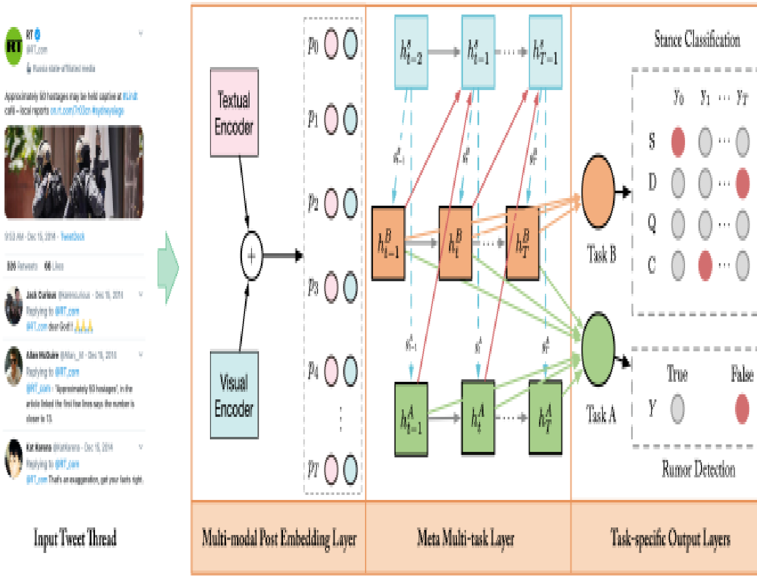


Fig. 1. Illustration of the Multi-modal Meta Multi-task Learning (MM-MTL) framework. We design a multi-modal post embedding layer to understand the multi-modal information in Tweet Thread. To effectively conduct multi-task learning on the rumor related tasks, we propose a meta multi-task learning method. By sharing the hyper meta-network, we capture the mutual meta-knowledge from various tasks. The meta-knowledge is then used to generate the parameters of the task-specific models. In the task-specific output layer, we employ an attention mechanism to estimate the weight of each reply, which can further utilize the semantic information hidden in the fine-grained stance labels.

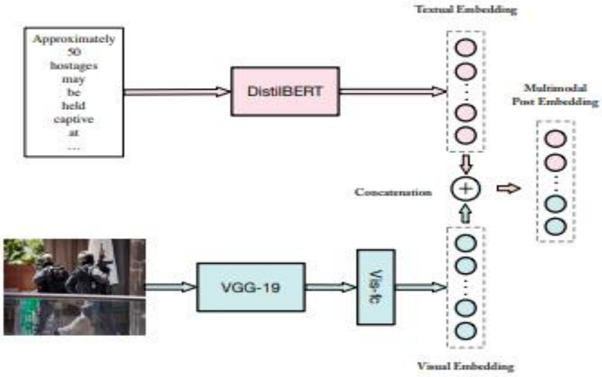

Textual Embedding
Approximately 50 hostages may be held captive at ...
DistilBERT
Multimodal Post Embedding
Concatenation
VGG-19
Vis-fc
Visual Embedding

Fig. 2. Multi-modal Feature Embedding of Post.

where Y ∈ {0, 1} represent the false rumor and true rumor, respectively.

b) Stance Detection: The stance detection task aims to determine the type of opinion (i.e., support, deny) that each reply post {p1, · · · , pT } reflects the veracity of the root post p0. We model this task as a sequence labeling problem, that is:

$$Fs(c) \rightarrow y0,\ y1,\ \cdots,\ yT$$

where each yt is the label that takes one of Support, Deny, Question or Commenting (SDQC). The input and output structures are shown in Table I. It is obvious that the rumor detection task and the stance detection task belong to different categories.

4. METHOD

A. Overall framework

The overall architecture is illustrated in Figure 1. We introduce a novel multi-modal meta multi-task learning method to improve the performance of rumor detection task by leveraging the meta knowledge of rumor detection and stance detection tasks. Our model consists of the following components:

• Multi-modal Post Embedding Layer: For each post in the claim, we model its multi-modal content as embedding vectors. Specifically, we employ BERT [44] to generate the embedding vector for the text content and use VGG19 to obtain the visual embedding for the visual content.

• Meta Multi-task Layers: We propose a meta multi-task learning method for social media rumor detection, in which a shared meta layer is used to learn the metaknowledge from rumor detection and stance detection tasks. With the guide of the shared meta network, the task-specific layer can obtain a precise representation of each post.

• Task-specific Output Layer: We apply an attention mechanism to the task-specific output layer to fully utilize the stance information of the user replies. In particular, the replies with the strong support or denial stance have a greater impact on the veracity of the rumor prediction.

Given a sequence of posts {p0, p1, · · ·, pT}, The MM-MTL first uses the multi-modal post embedding layer to encode the posts into vectors. Then, the post embedding is fed into the task-specific layer to extract the task-dependent features. The shared meta layer takes the post embedding and task-dependent features as input to learn the meta-knowledge h s ∗ . The meta-knowledge h s ∗ is then used to generate the parameters of the next time step GRU g A ∗ or g B ∗ for the task specific layers. After the task-specific output layers, the MM-MTL produces the prediction for every task.

B. Multi-modal Post Embedding Layer

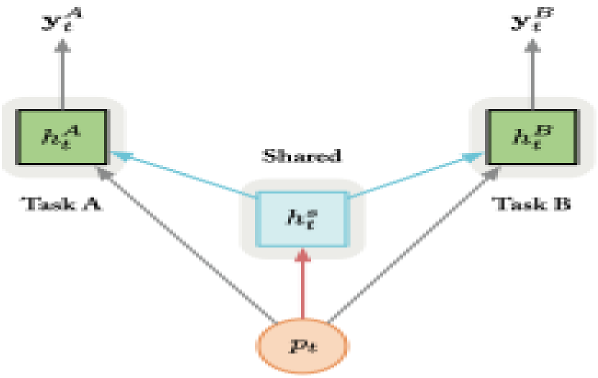
As mentioned in the problem statement, the input of rumor detection task and stance detection task is a claim c = {p0, p1, · · · , pT}, which consists of a set of relevant tweets. Specifically, pt is a tweet posted at time t, p0 is the root post, the {p1, · · · , pT} is the set of user replies, and the post pt consists of textual content (text sentence) xt and visual content (attached image) vt. Given a sequence of posts {p0, · · · , pT}, the multi-modal post embedding layer aims to convert them into a series of embedding vectors, where each vector represents its corresponding post. The structure of the multi-modal post embedding layer is shown in Figure 2. The sequential list of the words xt = {xt1, xt2, · · · , xtn} in the post pt is fed into the textual language model to get the embeddings. In order to precisely model both the semantic of the word and the linguistic contexts, we employ Bidirectional Encoder Representations from Transformers (BERT) as the core module of our textual language model. BERT has been proven to be effective in many fields such as question answering, translation, reading comprehension and text classification. As can be seen in Figure 2, we incorporate a modified pre-train BERT model, namely DistilBERT, in our textual representation learning. The DistilBERT model is composed of 6 transformer encoder layers with 768 hidden units and 66 million parameters. Extensive experiments demonstrate that DistilBERT can reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities. The text representation of post pt is calculated by pre-trained BistilBERT pxt = DistilBERT(xt) where pxt ∈ R dx is the last layer hidden-state of the classification token in BERT. Follow the previous work [44], we use pxt as the semantic summary of the input sentence. The dx is the dimension of the text embedding. The attached image vt of the post pt is fed into the visual feature extractor. In order to efficiently extract visual features, we employ the pre-trained VGG19 [45]. 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 dv. The operation of the last layer in the visual feature extractor can be represented as:

pvt = σ(Wv · VGG(vt))

where Wv is the weight matrix of the fully connected layer in the visual feature extractor. The textual feature representation pxt and visual feature representation pvt will be concatenated to form the multimodal feature representation denoted as pt ∈ Rd , which is the output of the post embedding layer, and d = dx + dv. Notice that, the parameters of pre-trained BistilBERT and VGG19 are kept static to avoid overfitting during training.

C. Meta Multi-task Layers

In this paper, we take a very different multi-task architecture from the meta-learning perspective. Figure 3 shows the difference between the generic multi-task learning and the proposed multi-modal meta multi-task learning. The generic multi-task learning shares some lower layers, which extract common features. After the shared layers, the remaining higher layers are parallel and independent respectively to each specific task. The proposed multi-modal meta multi-task learning shares some higher layers, which learn mutual meta knowledge from various tasks. Then the meta-knowledge is used to generate parameters for each specific task. The whole neural architecture of multi-modal meta multi task learning is shown in Figure 1. For each task, a task specific network is used for task-specific prediction, whose parameters are controlled by a shared meta network across all the tasks. Specifically, we use two recurrent neural networks with Gated Recurrent Unit (GRU) as meta and task network, respectively. The Meta-GRU is shared for all the tasks. The parameters of the Task-GRU are generated based on the meta vector g produced by the Meta-GRU. The Task-GRU unit at each time step t consists of a collection of vectors in Rh : an update gate zt, a reset gate rt, a memory cell h˜t and a hidden state ht:



a) Generic multi-task learning unit
(b) Meta multi-task learning unit

Fig. 3. The difference between the generic multi-task learning unit and the proposed meta multi-task learning unit. The generic multi-task learning shares some lower layers that extract common features from different tasks. The meta multi-task learning shares some higher layers, which learn mutual meta knowledge from various tasks.

zt = σ (Wz(gt)it + Uz(gt)ht−1)

rt = σ (Wr(gt)it + Ur(gt)ht−1)

h˜ t = tanh (Wh(gt)it + Uh(gt) (rt ht−1))

ht = zt ht−1 + (1 − zt) h˜ t

where it ∈ R d is the input at the current time step, p, W∗ (gt) ∈ R h×d , U∗ (gt) ∈ R h×h are the weight connections inside GRU, σ denotes the logistic sigmoid function and denotes element-wise multiplication. For the sake of simplicity, we concat W∗ (gt) and U∗ (gt) to θ∗ (gt) ∈ R h×(d+h) , for ∗ ∈ {z, r, h}.

The parameters θ∗ (gt) of GRU are now dynamically controlled by the meta vector gt. We define θ∗ (gt) with a low-rank factorized representation of the weights to reduce the output space of the dynamic parameters:

$$\begin{bmatrix} \theta_z(\mathbf{g}_t) \\ \theta_r(\mathbf{g}_t) \\ \theta_h(\mathbf{g}_t) \end{bmatrix} = \begin{bmatrix} G_z\mathbf{D}(\mathbf{g}_t)Q_z \\ G_r\mathbf{D}(\mathbf{g}_t)Q_r \\ G_h\mathbf{D}(\mathbf{g}_t)Q_h \end{bmatrix}$$

where G∗ ∈ R h×g , Q∗ ∈ R g×(d+h) are parameters for ∗ ∈ {z, r, h}, D(gt) is the diagonal matrix of gt. The Meta-GRU is a smaller network, which depends on the input it and the previous hidden state h k t−1 of the k-th task GRU. The Meta-GRU cell is given by:

$$\begin{aligned} \mathbf{z}_t^s &= \sigma\left(\mathbf{W}_z^s\mathbf{i}_t + \mathbf{W}_z^k\mathbf{h}_{t-1}^k + \mathbf{U}_z^s\mathbf{h}_{t-1}^s\right) \\ \mathbf{r}_t^s &= \sigma\left(\mathbf{W}_r^s\mathbf{i}_t + \mathbf{W}_r^k\mathbf{h}_{t-1}^k + \mathbf{U}_r^s\mathbf{h}_{t-1}^s\right) \\ \tilde{\mathbf{h}}_t^s &= \tanh\left(\mathbf{W}_h^s\mathbf{i}_t + \mathbf{W}_h^k\mathbf{h}_{t-1}^k + \mathbf{U}_h^s\left(\mathbf{r}_t^s \odot \mathbf{h}_{t-1}^s\right)\right) \\ \mathbf{h}_t^s &= \mathbf{z}_t^s \odot \mathbf{h}_{t-1}^s + (1 - \mathbf{z}_t^s) \odot \tilde{\mathbf{h}}_t \\ \mathbf{g}_t &= \mathbf{W}_g\mathbf{h}_t^s \end{aligned}$$

where Wg ∈ R g×m is a transformation matrix.

For multi-task learning, we can assign a task network to each task while sharing a meta network among tasks. The meta network captures the meta-knowledge of different tasks and predicts parameters for the task-specific network. For task k, the hidden states of the shared meta layer and the task-specific layer are:

$$[\mathbf{h}^s_t, \mathbf{g}_t] = \text{Meta-GRU}\ (\mathbf{p}_t, \mathbf{h}^k_{t-1}, \mathbf{h}^s_{t-$$

$$\mathbf{h}^k_t = \text{Task-GRU}\ (\mathbf{p}_t, \mathbf{h}^k_{t-1}; \mathbf{g}_t, \theta$$

where h 8 t and h k t are the hidden states of the shared meta GRU and the k-th task-specific GRU respectively; θ s and θ k denote their parameters.

D. Task-specific Output Layers

The task-specific representations h k , which is emitted by the multi-task architecture, are ultimately fed into different task specific output layers. We mark the rumor detection task and the stance detection task as Task A and Task B, respectively.

The rumor detection task is a sequence classification problem. As shown in the Equation 1, given a sequence of post p0, p1, · · · , pT , the classifier aims to determine its veracity value Y . Previous work only uses the last hidden state of the sequence hT as the representation of sequence, and feeds them to a linear classifier to predict the label. However, different posts may have different impacts on the rumor veracity in a conversation branch. For example, the tweets with strong support or deny stance should have more impacts for predicting rumor veracity. In order to better exploit the stance information, we explicitly include the hidden states from the stance layer in the attention calculation.

At each step t, the hidden state from the rumor detection layer h A t and the hidden state from stance detection layer h B t are concatenated and fed in an attention layer to generate the attention weight of current reply αt:

αt = attention(h A t ⊕h B t)

where ⊕ is the concat operator. The final representation of the claim c is calculated as:

c = X T t=0 αt(h A t ⊕h B t)

Then, the final classification decision for the claim is formulated probabilistically as:

Yˆ A = softmax(WAc + bA)

In the stance detection task, we aim to determine whether different replies support the root post or not. Equation 2 shows the target function of stance detection. Given a sequence of post p0, p1, · · · , pT , we classify every single post and give them a label, resulting in y0, y1, · · · , yT . The low-dimensional task-specific representations of the posts are {h B t } T t=0. We then feed them into a fully connected layer with softmax activation functions to generate the prediction for each post:

yˆ B t = softmax

where yˆ B t is the predicted probabilities over different stance classes at time t, VB 0 and VB t respectively denote the weights of the output layer for the posts, and b B is a bias term. All of them are task-specific parameters for stance detection. The parameters of the network are trained to minimise the cross-entropy of the predicted and true distributions for rumor task in Equation 14 and stance task in Equation 15.

$$\mathcal{L}_A(\Theta) = -\sum_{i=1} Y_i \log(\hat{Y}^A) + \lambda||\Theta||_2^2$$

$$\mathcal{L}_B(\Theta) = -\sum_{i=1}^{n}\sum_{t=0}^{T} \mathbf{y}_{it} \log\left(\hat{\mathbf{y}}^B_{it}\right) + \lambda||\Theta||_2^2$$

where Θ is the model parameter, and the Yi and yit are the ground truth of rumor detection task and stance detection task, respectively. We add the L2 regularizer to trade off the error and the scale of the model, where the λ is the trade-off coefficient. The training procedure is described in Algorithm 1. To balance the two tasks, we set a hyper-parameter β. In each iteration, a random number δ ∈ [0, 1) is first generated, then we pick the random training samples from task A if δ < β, or task B if δ ≥ β, then the model is updated according to the task-specific objective.

5. EXPERIMENTS AND RESULTS

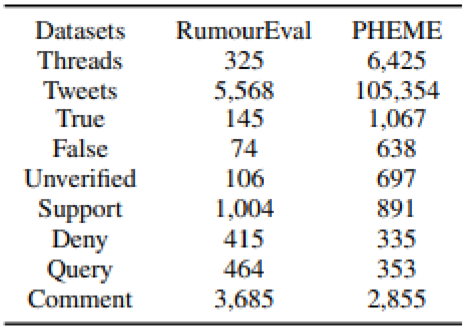
A. Experimental Setup

1) Datasets: Two benchmark datasets RumourEval and Pheme are used to validate the effectiveness of the proposed MM-MTL on detecting social media rumors. Both datasets are constituted by Twitter conversation threads. As shown in Table I, a conversation thread consists of a tweet making a true or false claim, and people’s replies expressing their opinion about it. Table II shows the statistics of the two

datasets. The RumourEval dataset has been developed for the SemEval2017 Task 8 competition1 , which aims to identify rumors and the reactions to rumors. The Pheme dataset is constructed to help understand how users treat online rumor before and after the news is detected to be true or false. Both datasets have the same labels on rumor detection and stance detection. Claims are labeled as true, false, and unverified. Since we aim to only distinguish true and false claims, we filter out the unverified tweets. The stance of tweets is annotated as Support, Deny, Query, and Comment (SDQC). We hold out 10% of the instances in each dataset for model tuning, and the rest of the instances are performed 5-fold cross-validation throughout all experiments. 2) Evaluation Metrics: The rumor detection task is a binary classification task, whose evaluation measure is commonly the Accuracy metric. However, when datasets suffer from class imbalance, it goes less reliable. Therefore, in addition to the Accuracy metric, we add Precision, Recall, and F1 score as complementary evaluation metrics for tasks. 3) Implementation Details: We use stochastic gradient descent to train our model by looping over the tasks. State-of-theart techniques have been employed to optimize the objective function: Dropout is applied to improve neural networks training, L2-norm regularization is imposed on the weights of the neural networks, and stepwise

exponential learning rate decay is adopted to anneal the variations of convergence. We use pre-trained BistilBERT, which is fine-tuned in sequence classification, and the dimension of text embedding dx is 768. The dimension of visual embedding dv is set to 256, hence the post embedding size d is 1024. For posts without image attached, we pad pvt with zeros. The dimension of the hidden layer of task-specific GRUs h is 256, the dimension of the meta GRU g is set to 64, the rate of dropout is 0.3 and the minibatch size is set to 32. The threshold β is set to 0.5.

TABLE II
STATISTICS OF THE TWO DATASETS.



Datasets	RumourEval	PHEME
Threads	325	6,425
Tweets	5,568	105,354
True	145	1,067
False	74	638
Unverified	106	697
Support	1,004	891
Deny	415	335
Query	464	353
Comment	3,685	2,855

B. Baselines

We test our multi-modal meta multi-task learning model against eight state-of-the-art models.

- SVM: A Support Vector Machine model detects misinformation relying on manually extracted features.
- CNN: A Convolutional Neural Network model employs pre-trained word embeddings based on Word2Vec as input embeddings to capture features similar to ngrams.
- TE: Tensor Embeddings model leverages tensor decomposition to derive concise claim embeddings, which are used to create a claim-by-claim graph for label propagation.
- DeClarE: Evidence-Aware Deep Learning uses claims as queries to retrieve evidence from replies. Then both claims and retrieved replies are input into a deep neural network with the attention mechanism.
- LSTM-MTL: LSTM-MTL trains the rumor detection and stance detection jointly with a LSTM-based multi-task. This model uses the hard parameter sharing scheme, where different tasks share the same LSTM layers to extract features.
- GRU-MTL: A GRU-based multi-task learning model, which jointly learns the rumor detection task and stance detection task. Unlike LSTM-MTL, this model utilizes the shared-private sharing scheme for multi-task learning. The model shares some lower layers to determine common features, while keep some taskspecific layers to extract task-specific features.
- Bayesian-DL: The Bayesian Deep Learning adopts Bayesian to represent both the prediction and uncertainty of claims and then encodes replies based on LSTM to update and generate a posterior representations.
- Trans-MTL: Trans-MTL jointly learns the rumor detection and stance detection tasks by a transformer based multi-task learning method and a selected sharing layer. The selected sharing layer adopts the gate mechanism and the attention mechanism to select shared feature flows between task.

C. Results and Analysis

Table III shows the performance of all the compared models based on the two datasets.

From Table III, we can draw the following observations:

• Most deep learning based models, such as Trans-MTL and Bayesian-DL, outperform feature engineering-based methods, like SVM. This demonstrates that deep learning methods learn a better hidden representation of claims and replies.

• The methods exploring relationships between a claim and its replies, such as Trans-MTL and Bayesian-DL, achieve better performance than claim content-based methods like TE and CNN. This demonstrates the significance of utilizing people's replies in the rumor detection task.

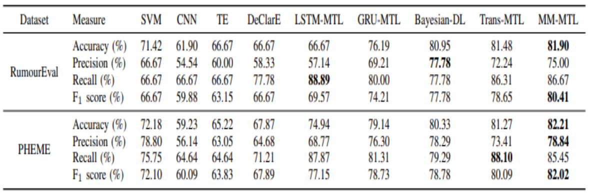
• The LSTM-MTL performs the worst among all multitask learning methods. This is because it uses the hard parameter sharing scheme, where different tasks share the same layers to extract the features. The task-specific layers take only common features as input but need to make predictions for different tasks.

• The GRU-MTL outperforms the LSTM-MTL because it adopts the shared-private sharing scheme. Specifically, the model shares some lower layers to determine common features across various tasks while keeping some taskspecific layers to extract task-depend features.

• Although the Trans-MTL model also adopts the shared private sharing scheme, the Trans-MTL is superior to GRU-MTL. This is because the Trans-MTL uses the transformer as the basic feature extractor and applies the selected sharing layer to filter and select shared feature flows between tasks.

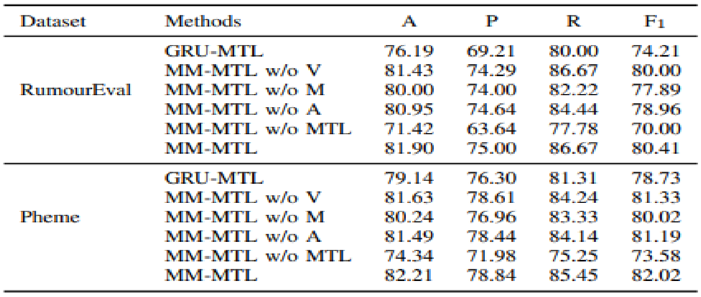
• Compared with all the baselines, the MM-MTL achieves the best performance and outperforms other rumor detection methods in most cases. We attribute the superiority of MM-MTL to its three properties: 1) The MM-MTL takes advantage of the multiple modalities of posts. 2) The proposed meta multi-task learning method can more effectively introduce the stance information of users's replies into rumor detection. 3) We employ the attention mechanism to estimate the weight of each reply.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED MULTI-MODAL META MULTI-TASK LEARNING METHOD AGAINST THE BASELINES.



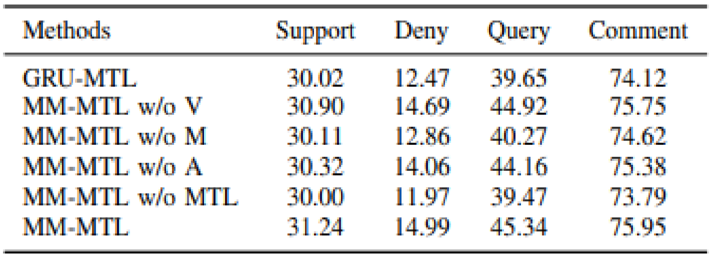
Dataset	Measure	SVM	CNN	TE	DeClarE	LSTM-MTL	GRU-MTL	Bayesian-DL	Trans-MTL	MM-MTL
RumourEval	Accuracy (%)	71.42	61.90	66.67	66.67	66.67	76.19	80.95	81.48	**81.90**
	Precision (%)	66.67	54.54	60.00	58.33	57.14	69.21	**77.78**	72.24	75.00
	Recall (%)	66.67	66.67	66.67	77.78	**88.89**	80.00	77.78	86.31	86.67
	F_1 score (%)	66.67	59.88	63.15	66.67	69.57	74.21	77.78	78.65	**80.41**
PHEME	Accuracy (%)	72.18	59.23	65.22	67.87	74.94	79.14	80.33	81.27	**82.21**
	Precision (%)	78.80	56.14	63.05	64.68	68.77	76.30	78.29	73.41	**78.84**
	Recall (%)	75.75	64.64	64.64	71.21	87.87	81.31	79.29	**88.10**	85.45
	F_1 score (%)	72.10	60.09	63.83	67.89	77.15	78.73	78.78	80.09	**82.02**

TABLE IV
COMPARISON AMONG MM-MTL VARIANTS (A: ACCURACY; P: PRECISION; R: RECALL; F_1: F_1 SCORE).



Dataset	Methods	A	P	R	F_1
RumourEval	GRU-MTL	76.19	69.21	80.00	74.21
	MM-MTL w/o V	81.43	74.29	86.67	80.00
	MM-MTL w/o M	80.00	74.00	82.22	77.89
	MM-MTL w/o A	80.95	74.64	84.44	78.96
	MM-MTL w/o MTL	71.42	63.64	77.78	70.00
	MM-MTL	81.90	75.00	86.67	80.41
Pheme	GRU-MTL	79.14	76.30	81.31	78.73
	MM-MTL w/o V	81.63	78.61	84.24	81.33
	MM-MTL w/o M	80.24	76.96	83.33	80.02
	MM-MTL w/o A	81.49	78.44	84.14	81.19
	MM-MTL w/o MTL	74.34	71.98	75.25	73.58
	MM-MTL	82.21	78.84	85.45	82.02

TABLE V
STANCE DETECTION PERFORMANCE (F_1) COMPARISON AMONG MM-MTL VARIANTS ON PHEME DATASET.



Methods	Support	Deny	Query	Comment
GRU-MTL	30.02	12.47	39.65	74.12
MM-MTL w/o V	30.90	14.69	44.92	75.75
MM-MTL w/o M	30.11	12.86	40.27	74.62
MM-MTL w/o A	30.32	14.06	44.16	75.38
MM-MTL w/o MTL	30.00	11.97	39.47	73.79
MM-MTL	31.24	14.99	45.34	75.95

D. Model Ablation

The proposed MM-MTL consists of three components: the multi-modal post embedding layer, the meta multi-task layer, and the task-specific output layer. To evaluate the effectiveness of different components in our method, we ablate our method into several simplified models and compare their performance against related methods. The details of these methods are described as follows:

MM-MTL w/o V: A variant of MM-MTL, in which we remove the visual encoder from the multi-modal post embedding layer. In this model, only the textual modality feature is used for rumor detection.

MM-MTL w/o M: A variant of MM-MTL, which removes the shared meta layer. Without the proposed meta network, this model can be viewed as the GRU-MTL model which equipped with a multi-modal post embedding layer and attention-based output layer.

MM-MTL w/o A: A variant of MM-MTL with the attention mechanism being removed.

MM-MTL w/o MTL: A variant of MM-MTL with the multi-task learning mechanism being removed. In this model, only one of the rumor detection task and the stance detection task is considered.

As shown in Table IV,we compare the rumor detection performance of the MM-MTL variants on RumourEval and PHEME dataset.

From these tables, we can conclude that:

- The MM-MTL outperforms MM-MTL w/o V on both datasets. The results prove that the multi-modal features learned by multi-modal post embedding layer can improve the performance of rumor detection.
- Compared with MM-MTL w/o M, MM-MTL obtains more excellent performance on both datasets, which indicates that the meta network learns and shares valuable knowledge in rumor detection.
- MM-MTL achieves better results than MM-MTL w/o A model in both datasets, which shows that it is effective to focus on the useful parts of the input sequence.
- The MM-MTL outperforms MM-MTL w/o MTL in a large gap. This demonstrates that the stance information introduced by multi-task learning can boost the rumor detection task's performance. The MM-MTL w/o MTL performs the best in the non-MTL methods (SVM, CNN, TE, DeClarE), which indicates the effectiveness of the proposed multi-modal post embedding layer and the meta GRU module.

In addition to the comparison of the primary rumor detection task, we also verify the effect of meta-network on the auxiliary stance detection task. In Table V, we show the stance detection results of the MM-MTL variants on the Pheme dataset. From the Table V, we can find that all the modules proposed in our paper can benefit the stance detection task. In particular, the MM-MTL method is significantly better than the MM-MTL w/o M, which shows that the shared meta layer is able to learn the meta knowledge across tasks, and benefit for both tasks.

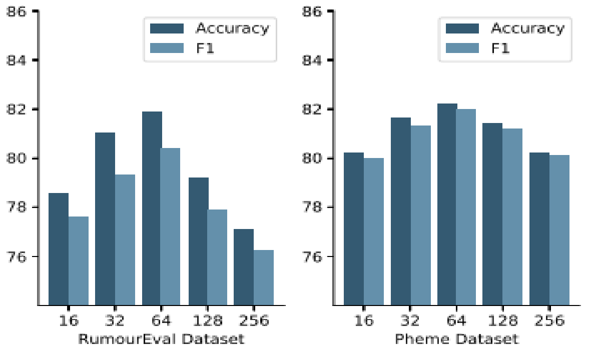
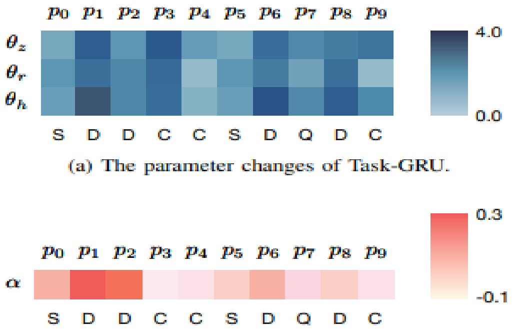


Fig. 4. The effect of the meta vector dimension on model performance.

TABLE VI
THE RUMOR CLAIM SAMPLE FROM PHEME DATASET.

Index	Input	Stance Label
p_0	sydney siege gunman forces hostages to hold up isis flag in window	Support
p_1	these not the same 1st shahadah flag 2nd is specifically claimed by is	Deny
p_2	took me sec to google isis flag and see this was not it your research budget that bad	Deny
p_3	guess sydneys off our list	Comment
p_4	if they only knew that isis was created by the american government	Comment
p_5	sydney siege gunman forces hostages to hold up isis flag in window	Support
p_6	stop point scoring...ok so it's not the official is flag...these ppl still held at gunpoint	Deny
p_7	are you sure they have guns australia has banned a lot of them	Query
p_8	isis and islamicstate work for cia claims former alqaeda commander	Deny
p_9	oh dear pray for these innocent people in sydney who are being held up in a cafe ...	Comment



(a) The parameter changes of Task-GRU.

(b) The attention weight of posts.

Fig. 5. The upper figure presents parameter changes of the Task-GRU. The θz, θr, θh are the parameter changes of update gate, reset gate and new hidden state respectively. The lower figure presents the attention weight α of hidden state of each time step in rumor veracity prediction.

E. Sensitivity Analysis In this subsection we evaluate the effect of the dimension of the meta vector g. We first set a dimension for g, then optimize the rest of the hyper parameters on the validation subset. In Figure 4, we illustrates the effect on performance of the dimension of g on both datasets. We observe that the results are similar for both evaluation measures, Accuracy and F1. Varying the dimension from 16 to 64 brings a larger performance improvement to the model. When the dimension is 64, the model obtains the highest accuracy of 81.90% and the highest F1 of 80.41%, on the RumourEval test subset and the highest accuracy of 82.21% and F1 of 82.02%, on the Pheme test subset. These results show that maintaining an increase in the meta vector dimension may not necessarily lead to an improvement in performance. The reason could be found in the limited size of the datasets, which might cause overfitting when the model is too complex.

F. Case Study

To obtain deeper insights and detailed interpretability about the effectiveness of the shared meta layer and the attention mechanism, we design an experiment to try to disclose the behaviors of them. We randomly sample a claim from the development set of Pheme dataset, the claim can be found in Table VI. Then, we visualize the changes of matrices generated by Meta-GRU in Figure 5(a). In Figure 5(b), we visualize which posts are concerned in the attention mechanism. From Figure 5, we obtain the following observations: • The matrices change obviously when facing the stance gap like Support to Deny in p0 and p1, and slowly change when facing similar input or same stance. This shows that the parameters of the Task-GRU vary from position to position, and dynamic matrices generated by Meta-GRU make the layer more flexible. • The strong denying posts and supporting posts are given more attention, while the commenting posts which do not add anything to the veracity of a claim are getting little attention. This shows that the attention mechanism can effectively capture key replies and boost the performance of rumor detection.

6. CONCLUSION

In this paper, we design a novel Multi-modal Meta MultiTask Learning (MM-MTL) framework to detect social media rumors. Specifically, we aims to improve the performance of rumor detection task by leveraging the stance information of users' replies in stance detection. Different from the generic multi-task learning methods which share lower layers to extract common features, MM-MTL shares higher meta network layers, which capture the meta-knowledge behind the multi-modal posts. The shared meta-knowledge then benefits each task by dynamically generating the parameters of the task-specific models. Besides, we employ the attention mechanism to estimate the weight of each reply, which can better exploit the semantic information hidden in the finegrained stance labels. Extensive experiments on two Twitter benchmark datasets demonstrate that our proposed method achieves state-of-the-art performance. In the future, we will introduce more rumor related tasks into the multi-modal meta multi-task learning framework, such as user's trustworthiness evaluation task.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–36, May 2017.

[2] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," Plos One, vol. 12, no. 1, pp. e0 168 344–, 2017.

[3] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in 2013 IEEE 13th International Conference on Data Mining, Dec. 2013, pp. 1103–1108.

[4] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ser. CIKM '15. New York, NY, USA: ACM, 2015, pp. 1867–1870.

[5] J. Ma, W. Gao, Z. Wei, Y. Lu, and K. Wong, "Detect rumors using time series of social context information on microblogging websites," in Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015. ACM, 2015, pp. 1751–1754.

[6] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011. ACM, 2011, pp. 675–684.

[7] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. IJCAI/AAAI Press, 2016, pp. 3818–3824.

[8] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation

identification," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. ijcai.org, 2017, pp. 3901– 3907.

[9] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017. ACM, 2017, pp. 795– 816.

[10] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019. ACM, 2019, pp. 1942–1951.

[11] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," in Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. Association for Computational Linguistics, 2018, pp. 3402–3413.

[12] J. Ma, W. Gao, and K. Wong, "Detect rumor and stance jointly by neural multi-task learning," in Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018. ACM, 2018, pp. 585–593.

[13] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, "Fake news stance detection using stacked ensemble of classifiers," in Proceedings of the 2017 Workshop: Natural Language Processing meets Journalism, NLPmJ@EMNLP, Copenhagen, Denmark, September 7, 2017. Association for Computational Linguistics, 2017, pp. 80–83.

[14] S. Dungs, A. Aker, N. Fuhr, and K. Bontcheva, "Can rumour stance alone predict veracity?" in Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. Association for Computational Linguistics, 2018, pp. 3360–3370.

[15] L. Wu, Y. Rao, H. Jin, A. Nazir, and L. Sun, "Different absorption from the same sharing: Sifted multi-task learning for fake news detection," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, 2019, pp. 4643–4652.