Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection

Abstract

Recently, neural networks based on multi-task learning have achieved promising performance on fake news detection, which focus on learning shared features among tasks as complementary features to serve different tasks. However, in most of the existing approaches, the shared features are completely assigned to different tasks without selection, which may lead to some useless and even adverse features integrated into specific tasks. In this paper, we design a sifted multi-task learning method with a selected sharing layer for fake news detection. The selected sharing layer adopts gate mechanism and attention mechanism to filter and select shared feature flows between tasks. Experiments on two public and widely used competition datasets, i.e. RumourEval and PHEME, demonstrate that our proposed method achieves the state-of-the-art performance and boosts the F1-score by more than 0.87%, 1.31%, respectively.

Introduction

In recent years, the proliferation of fake news with various content, high-speed spreading, and extensive influence has become an increasingly alarming issue. A concrete instance was cited by Time Magazine in 2013 when a false announcement of Barack Obama's injury in a White House explosion “wiped off 130 Billion US Dollars in stock value in a matter of seconds". Other examples, an analysis of the US Presidential Election in 2016 BIBREF0 revealed that fake news was widely shared during the three months prior to the election with 30 million total Facebook shares of 115 known pro-Trump fake stories and 7.6 million of 41 known pro-Clinton fake stories. Therefore, automatically detecting fake news has attracted significant research attention in both industries and academia.

Most existing methods devise deep neural networks to capture credibility features for fake news detection. Some methods provide in-depth analysis of text features, e.g., linguistic BIBREF1, semantic BIBREF2, emotional BIBREF3, stylistic BIBREF4, etc. On this basis, some work additionally extracts social context features (a.k.a. meta-data features) as credibility features, including source-based BIBREF5, user-centered BIBREF6, post-based BIBREF7 and network-based BIBREF8, etc. These methods have attained a certain level of success. Additionally, recent researches BIBREF9, BIBREF10 find that doubtful and opposing voices against fake news are always triggered along with its propagation. Fake news tends to provoke controversies compared to real news BIBREF11, BIBREF12. Therefore, stance analysis of these controversies can serve as valuable credibility features for fake news detection.

There is an effective and novel way to improve the performance of fake news detection combined with stance analysis, which is to build multi-task learning models to jointly train both tasks BIBREF13, BIBREF14, BIBREF15. These approaches model information sharing and representation reinforcement between the two tasks, which expands valuable features for their respective tasks. However, prominent drawback to these methods and even typical multi-task learning methods, like the shared-private model, is that the shared features in the shared layer are equally sent to their respective tasks without filtering, which causes that some useless and even adverse features are mixed in different tasks, as shown in Figure FIGREF2(a). By that the network would be confused by these features, interfering effective sharing, and even mislead the predictions.

To address the above problems, we design a sifted multi-task learning model with filtering mechanism (Figure FIGREF2(b)) to detect fake news by joining stance detection task. Specifically, we introduce a selected sharing layer into each task after the shared layer of the model for filtering shared features. The selected sharing layer composes of two cells: gated sharing cell for discarding useless features and attention sharing cell for focusing on features that are conducive to their respective tasks. Besides, to better capture long-range dependencies and improve the parallelism of the model, we apply transformer

encoder module BIBREF16 to our model for encoding input representations of both tasks. Experimental results reveal that the proposed model outperforms the compared methods and gains new benchmarks.

In summary, the contributions of this paper are as follows:

We explore a selected sharing layer relying on gate mechanism and attention mechanism, which can selectively capture valuable shared features between tasks of fake news detection and stance detection for respective tasks.

The transformer encoder is introduced into our model for encoding inputs of both tasks, which enhances the performance of our method by taking advantages of its long-range dependencies and parallelism.

Experiments on two public, widely used fake news datasets demonstrate that our method significantly outperforms previous state-of-the-art methods.

Related Work

Fake News Detection Exist studies for fake news detection can be roughly summarized into two categories. The first category is to extract or construct comprehensive and complex features with manual ways BIBREF5, BIBREF8, BIBREF17. The second category is to automatically capture deep features based on neural networks. There are two ways in this category. One is to capture linguistic features from text content, such as semantic BIBREF7, BIBREF18, writing styles BIBREF4, and textual entailments BIBREF19. The other is to focus on gaining effective features from the organic integration of text and user interactions BIBREF20, BIBREF21. User interactions include users' behaviours, profiles, and networks between users. In this work, following the second way, we automatically learn representations of text and stance information from response and forwarding (users' behaviour) based on multi-task learning for fake

news detection.

Stance Detection The researches BIBREF22, BIBREF23 demonstrate that the stance detected from fake news can serve as an effective credibility indicator to improve the performance of fake news detection. The common way of stance detection in rumors is to catch deep semantics from text content based on neural networksBIBREF24. For instance, Kochkina et al.BIBREF25 project branch-nested LSTM model to encode text of each tweet considering the features and labels of the predicted tweets for stance detection, which reflects the best performance in RumourEval dataset. In this work, we utilize transformer encoder to acquire semantics from responses and forwarding of fake news for stance detection.

Multi-task Learning A collection of improved models BIBREF26, BIBREF27, BIBREF28 are developed based on multi-task learning. Especially, shared-private model, as a popular multi-task learning model, divides the features of different tasks into private and shared spaces, where shared features, i.e., task-irrelevant features in shared space, as supplementary features are used for different tasks. Nevertheless, the shared space usually mixes some task-relevant features, which makes the learning of different tasks introduce noise. To address this issue, Liu et al. BIBREF29 explore an adversarial shared-private model to alleviate the shared and private latent feature spaces from interfering with each other. However, these models transmit all shared features in the shared layer to related tasks without distillation, which disturb specific tasks due to some useless and even harmful shared features. How to solve this drawback is the main challenge of this work.

Method

We propose a novel sifted multi-task learning method on the ground of shared-private model to jointly train the tasks of stance detection and fake news detection, filter original outputs of shared layer by a selected sharing layer. Our model consists of a 4-level hierarchical structure, as shown in Figure

FIGREF6. Next, we will describe each level of our proposed model in detail.

Method ::: Input Embeddings

In our notation, a sentence of length l tokens is indicated as ${\rm \textbf {X}}=\lbrace x_1, x_2, ... ,x_l\rbrace $. Each token is concatenated by word embeddings and position embeddings. Word embeddings w_i of token x_i are a d_w-dimensional vector obtained by pre-trained Word2Vec model BIBREF30, i.e., $w_i \in \mathbb {R}^{d_w}$. Position embeddings refer to vectorization representations of position information of words in a sentence. We employ one-hot encoding to represent position embeddings p_i of token x_i, where $p_i \in \mathbb {R}^{d_p}$, d_p is the positional embedding dimension. Therefore, the embeddings of a sentence are represented as $ {\rm \textbf {E}}=\lbrace [w_1;p_1], [w_2;p_2], ..., [w_l;p_l]\rbrace , {\rm \textbf {E}}\in \mathbb {R}^{l \times (d_p+d_w)}$. In particular, we adopt one-hot encoding to embed positions of tokens, rather than sinusoidal position encoding recommended in BERT model BIBREF31. The reason is that our experiments show that compared with one-hot encoding, sinusoidal position encoding not only increases the complexity of models but also performs poorly on relatively small datasets.

Method ::: Shared-private Feature Extractor

Shared-private feature extractor is mainly used for extracting shared features and private features among different tasks. In this paper, we apply the encoder module of transformer BIBREF16 (henceforth, transformer encoder) to the shared-private extractor of our model. Specially, we employ two transformer encoders to encode the input embeddings of the two tasks as their respective private features. A transformer encoder is used to encode simultaneously the input embeddings of the two tasks as shared features of both tasks. This process is illustrated by the shared-private layer of Figure FIGREF6. The red box in the middle denotes the extraction of shared features and the left and right boxes represent the

extraction of private features of two tasks. Next, we take the extraction of the private feature of fake news detection as an example to elaborate on the process of transformer encoder.

The kernel of transformer encoder is the scaled dot-product attention, which is a special case of attention mechanism. It can be precisely described as follows:

where ${\rm \textbf {Q}} \in \mathbb {R}^{l \times (d_p+d_w)}$, ${\rm \textbf {K}} \in \mathbb {R}^{l \times (d_p+d_w)}$, and ${\rm \textbf {V}} \in \mathbb {R}^{l \times (d_p+d_w)}$ are query matrix, key matrix, and value matrix, respectively. In our setting, the query ${\rm \textbf {Q}}$ stems from the inputs itself, i.e., ${\rm \textbf {Q}}={\rm \textbf {K}}={\rm \textbf {V}}={\rm \textbf {E}}$.

To explore the high parallelizability of attention, transformer encoder designs a multi-head attention mechanism based on the scaled dot-product attention. More concretely, multi-head attention first linearly projects the queries, keys and values h times by using different linear projections. Then h projections perform the scaled dot-product attention in parallel. Finally, these results of attention are concatenated and once again projected to get the new representation. Formally, the multi-head attention can be formulated as follows:

where ${\rm \textbf {W}}_i^Q \in \mathbb {R}^{(d_p+d_w) \times d_k}$, ${\rm \textbf {W}}_i^K \in \mathbb {R}^{(d_p+d_w) \times d_k}$, ${\rm \textbf {W}}_i^V \in \mathbb {R}^{(d_p+d_w) \times d_k}$ are trainable projection parameters. d_k is $(d_p+d_w)/h$, h is the number of heads. In Eq.(DISPLAY_FORM11), ${\rm \textbf {W}}^o \in \mathbb {R}^{(d_p+d_w) \times (d_p+d_w)}$ is also trainable parameter.

Method ::: Selected Sharing Layer

In order to select valuable and appropriate shared features for different tasks, we design a selected

sharing layer following the shared layer. The selected sharing layer consists of two cells: gated sharing cell for filtering useless features and attention sharing cell for focusing on valuable shared features for specific tasks. The description of this layer is depicted in Figure FIGREF6 and Figure FIGREF15. In the following, we introduce two cells in details.

Gated Sharing Cell Inspired by forgotten gate mechanism of LSTM BIBREF32 and GRU BIBREF33, we design a single gated cell to filter useless shared features from shared layer. There are two reasons why we adopt single-gate mechanism. One is that transformer encoder in shared layer can efficiently capture the features of long-range dependencies. The features do not need to capture repeatedly by multiple complex gate mechanisms of LSTM and GRU. The other is that single-gate mechanism is more convenient for training BIBREF34. Formally, the gated sharing cell can be expressed as follows:

where ${\rm \textbf {H}}_{shared}\! \in \! \mathbb {R}^{1 \times l(d_p+d_w)}$ denotes the outputs of shared layer upstream, ${\rm \textbf {W}}_{fake} \in \mathbb {R}^{l(d_p+d_w) \times l(d_p+d_w)}$ and ${\rm \textbf {b}}_{fake} \in \mathbb {R}^{1 \times l(d_p+d_w)}$ are trainable parameters. $\sigma $ is a non-linear activation - sigmoid, which makes final choices for retaining and discarding features in shared layer.

Then the shared features after filtering via gated sharing cell ${\rm \textbf {g}}_{fake}$ for the task of fake news detection are represented as:

where $\odot $ denotes element-wise multiplication.

Similarly, for the auxiliary task - the task of stance detection, filtering process in the gated sharing cell is the same as the task of fake news detection, so we do not reiterate them here.

Attention Sharing Cell To focus on helpful shared features that are beneficial to specific tasks from upstream shared layer, we devise an attention sharing cell based on attention mechanism. Specifically, this cell utilizes input embeddings of the specific task to weight shared features for paying more attention to helpful features. The inputs of this cell include two matrixes: the input embeddings of the specific task and the shared features of both tasks. The basic attention architecture of this cell, the same as shared-private feature extractor, also adopts transformer encoder (the details in subsection SECREF8). However, in this architecture, query matrix and key matrix are not projections of the same matrix, i.e., query matrix ${\rm \textbf {E}}_{fake}$ is the input embeddings of fake news detection task, and key matrix ${\rm \textbf {K}}_{shared}$ and value matrix ${\rm \textbf {V}}_{shared}$ are the projections of shared features ${\rm \textbf {H}}_{shared}$. Formally, the attention sharing cell can be formalized as follows:

where the dimensions of ${\rm \textbf {E}}_{fake}$, ${\rm \textbf {K}}_{shared}$, and ${\rm \textbf {V}}_{shared}$ are all $\mathbb {R}^{l\times (d_p+d_w)}$. The dimensions of remaining parameters in Eqs.(DISPLAY_FORM16, DISPLAY_FORM17) are the same as in Eqs.(DISPLAY_FORM10, DISPLAY_FORM11). Moreover, in order to guarantee the diversity of focused shared features, the number of heads h should not be set too large. Experiments show that our method performs the best performance when h is equal to 2.

Integration of the Two Cells We first convert the output of the two cells to vectors ${\rm \textbf {G}}$ and ${\rm \textbf {A}}$, respectively, and then integrate the vectors in full by the absolute difference and element-wise product BIBREF35.

where $\odot $ denotes element-wise multiplication and $;$ denotes concatenation.

Method ::: The Output Layer

As the last layer, softmax functions are applied to achieve the classification of different tasks, which emits the prediction of probability distribution for the specific task i.

where $\hat{{\rm \textbf {y}}}_i$ is the predictive result, ${\rm \textbf {F}}_i$ is the concatenation of private features ${\rm \textbf {H}}_i$ of task i and the outputs ${\rm \textbf {SSL}}_i$ of selected sharing layer for task i. ${\rm \textbf {W}}_i$ and ${\rm \textbf {b}}_i$ are trainable parameters.

Given the prediction of all tasks, a global loss function forces the model to minimize the cross-entropy of prediction and true distribution for all the tasks:

where $\lambda _i$ is the weight for the task i, and N is the number of tasks. In this paper, $N=2$, and we give more weight $\lambda $ to the task of fake news detection.

Experiments ::: Datasets and Evaluation Metrics

We use two public datasets for fake news detection and stance detection, i.e., RumourEval BIBREF36 and PHEME BIBREF12. We introduce both the datasets in details from three aspects: content, labels, and distribution.

Content. Both datasets contain Twitter conversation threads associated with different newsworthy events including the Ferguson unrest, the shooting at Charlie Hebdo, etc. A conversation thread consists of a tweet making a true and false claim, and a series of replies. Labels. Both datasets have the same labels on fake news detection and stance detection. Fake news is labeled as true, false, and unverified. Because we focus on classifying true and false tweets, we filter the unverified tweets. Stance of tweets is annotated as support, deny, query, and comment. Distribution. RumourEval contains 325 Twitter threads discussing rumours and PHEME includes 6,425 Twitter threads. Threads, tweets, and class distribution of

the two datasets are shown in Table TABREF24.

In consideration of the imbalance label distributions, in addition to accuracy (A) metric, we add Precision (P), Recall (R) and F1-score (F1) as complementary evaluation metrics for tasks. 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.

Experiments ::: Settings

Pre-processing - Processing useless and inappropriate information in text: (1) removing nonalphabetic characters; (2) removing website links of text content; (3) converting all words to lower case and tokenize texts.

Parameters - hyper-parameters configurations of our model: for each task, we strictly turn all the hyper-parameters on the validation dataset, and we achieve the best performance via a small grid search. The sizes of word embeddings and position embeddings are set to 200 and 100. In transformer encoder, attention heads and blocks are set to 6 and 2 respectively, and the dropout of multi-head attention is set to 0.7. Moreover, the minibatch size is 64; the initial learning rate is set to 0.001, the dropout rate to 0.3, and $\lambda $ to 0.6 for fake news detection.

Experiments ::: Performance Evaluation ::: Baselines

SVM A Support Vector Machines model in BIBREF36 detects misinformation relying on manually extracted features.

CNN A Convolutional Neural Network model BIBREF37 employs pre-trained word embeddings based on

Word2Vec as input embeddings to capture features similar to n-grams.

TE Tensor Embeddings BIBREF38 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 BIBREF39 encodes claims and articles by Bi-LSTM and focuses on each other based on attention mechanism, and then concatenates claim source and article source information.

MTL-LSTM A multi-task learning model based on LSTM networks BIBREF14 trains jointly the tasks of veracity classification, rumor detection, and stance detection.

TRNN Tree-structured RNN BIBREF40 is a bottom-up and a top-down tree-structured model based on recursive neural networks.

Bayesian-DL Bayesian Deep Learning model BIBREF41 first adopts Bayesian to represent both the prediction and uncertainty of claim and then encodes replies based on LSTM to update and generate a posterior representations.

Experiments ::: Performance Evaluation ::: Compared with State-of-the-art Methods

We perform experiments on RumourEval and PHEME datasets to evaluate the performance of our method and the baselines. The experimental results are shown in Table TABREF27. We gain the following observations:

On the whole, most well-designed deep learning methods, such as ours, Bayesian-DL, and TRNN,

outperform feature engineering-based methods, like SVM. This illustrates that deep learning methods can represent better intrinsic semantics of claims and replies.

In terms of recall (R), our method and MTL-LSTM, both based on multi-task learning, achieve more competitive performances than other baselines, which presents that sufficient features are shared for each other among multiple tasks. Furthermore, our method reflects a more noticeable performance boost than MTL-LSTM on both datasets, which extrapolates that our method earns more valuable shared features.

Although our method shows relatively low performance in terms of precision (P) and recall (R) compared with some specific models, our method achieves the state-of-the-art performance in terms of accuracy (A) and F1-score (F1) on both datasets. Taking into account the tradeoff among different performance measures, this reveals the effectiveness of our method in the task of fake news detection.

Experiments ::: Discussions ::: Model Ablation

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:

Single-task Single-task is a model with transformer encoder as the encoder layer of the model for fake news detection.

MT-lstm The tasks of fake news detection and stance detection are integrated into a shared-private model and the encoder of the model is achieved by LSTM.

MT-trans The only difference between MT-trans and MT-lstm is that encoder of MT-trans is composed of transformer encoder.

MT-trans-G On the basis of MT-trans, MT-trans-G adds gated sharing cell behind the shared layer of MT-trans to filter shared features.

MT-trans-A Unlike MT-trans-G, MT-trans-A replaces gated sharing cell with attention sharing cell for selecting shared features.

MT-trans-G-A Gated sharing cell and attention sharing cell are organically combined as selected sharing layer behind the shared layer of MT-trans, called MT-trans-G-A.

Table TABREF30 provides the experimental results of these methods on RumourEval and PHEME datasets. We have the following observations:

Effectiveness of multi-task learning. MT-trans boosts about 9% and 15% performance improvements in accuracy on both datasets compared with Single-task, which indicates that the multi-task learning method is effective to detect fake news.

Effectiveness of transformer encoder. Compared with MT-lstm, MT-trans obtains more excellent performance, which explains that transformer encoder has better encoding ability than LSTM for news text on social media.

Effectiveness of the selected sharing layer. Analysis of the results of the comparison with MT-trans, MT-trans-G, MT-Trans-A, and MT-trans-G-A shows that MT-trans-G-A ensures optimal performance with the help of the selected sharing layer of the model, which confirms the reasonability of selectively sharing

different features for different tasks.

Experiments ::: Discussions ::: Error Analysis

Although the sifted multi-task learning method outperforms previous state-of-the-art methods on two datasets (From Table TABREF27), we observe that the proposed method achieves more remarkable performance boosts on PHEME than on RumourEval. There are two reasons for our analysis according to Table TABREF24 and Table TABREF27. One is that the number of training examples in RumourEval (including 5,568 tweets) is relatively limited as compared with PHEME (including 105,354 tweets), which is not enough to train deep neural networks. Another is that PHEME includes more threads (6,425 threads) than RumourEval (325 threads) so that PHEME can offer more rich credibility features to our proposed method.

Experiments ::: Case Study

In order to obtain deeper insights and detailed interpretability about the effectiveness of the selected shared layer of the sifted multi-task learning method, we devise experiments to explore some ideas in depth: 1) Aiming at different tasks, what effective features can the selected sharing layer in our method obtain? 2) In the selected sharing layer, what features are learned from different cells?

Experiments ::: Case Study ::: The Visualization of Shared Features Learned from Two Tasks

We visualize shared features learned from the tasks of fake news detection and stance detection. Specifically, we first look up these elements with the largest values from the outputs of the shared layer and the selected shared layer respectively. Then, these elements are mapped into the corresponding values in input embeddings so that we can find out specific tokens. The experimental results are shown in

Figure FIGREF35. We draw the following observations:

Comparing PL-FND and PL-SD, private features in private layer from different tasks are different. From PL-FND, PL-SD, and SLT, the combination of the private features and shared features from shared layer increase the diversity of features and help to promote the performance of both fake news detection and stance detection.

By compared SL, SSL-FND, and SSL-SD, selected sharing layers from different tasks can not only filter tokens from shared layer (for instance, `what', `scary', and `fact' present in SL but not in SSL-SD), but also capture helpful tokens for its own task (like `false' and `real' in SSL-FND, and `confirm' and `misleading' in SSL-SD).

Experiments ::: Case Study ::: The Visualization of Different Features Learned from Different Cells

To answer the second question, we examine the neuron behaviours of gated sharing cell and attention sharing cell in the selected sharing layer, respectively. More concretely, taking the task of fake news detection as an example, we visualize feature weights of ${\rm \textbf {H}}_{shared}$ in the shared layer and show the weight values ${\rm \textbf {g}}_{fake}$ in gated sharing cell. By that we can find what kinds of features are discarded as interference, as shown in Figure FIGREF42(a). In addition, for attention sharing cell, we visualize which tokens are concerned in attention sharing cell, as shown in Figure FIGREF42(b). From Figure FIGREF42(a) and FIGREF42(b), we obtain the following observations:

In Figure FIGREF42(a), only the tokens “gunmen, hostages, Sydney, ISIS" give more attention compared with vanilla shared-private model (SP-M). In more details, `gunmen' and `ISIS' obtain the highest weights. These illustrate that gated sharing cell can effectively capture key tokens.

In Figure FIGREF42(b), “live coverage", as a prominent credibility indicator, wins more concerns in the task of fake news detection than other tokens. By contrast, when the sentence of Figure FIGREF42(b) is applied to the task of stance detection, the tokens “shut down" obtain the maximum weight, instead of “live coverage". These may reveal that attention sharing cell focuses on different helpful features from the shared layer for different tasks.

Conclusion

In this paper, we explored a sifted multi-task learning method with a novel selected sharing structure for fake news detection. The selected sharing structure fused single gate mechanism for filtering useless shared features and attention mechanism for paying close attention to features that were helpful to target tasks. We demonstrated the effectiveness of the proposed method on two public, challenging datasets and further illustrated by visualization experiments. There are several important directions remain for future research: (1) the fusion mechanism of private and shared features; (2) How to represent meta-data of fake news better to integrate into inputs.

Acknowledgments

The research work is supported by “the World-Class Universities(Disciplines) and the Characteristic Development Guidance Funds for the Central Universities"(PY3A022), Shenzhen Science and Technology Project(JCYJ20180306170836595), the National Natural Science Fund of China (No.F020807), Ministry of Education Fund Project “Cloud Number Integration Science and Education Innovation" (No.2017B00030), Basic Scientific Research Operating Expenses of Central Universities (No.ZDYF2017006).