TRANSFAKE: Multi-task Transformer for Multimodal Enhanced Fake News Detection

Quanliang Jing[1,2], Di Yao[1], Xinxin Fan[1], Baoli Wang[1,2], Haining Tan[1,2], Jingping Bi[1]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[1]
University of Chinese Academy of Sciences, China[2]
{jingquanliang, yaodi, fanxinxin, wangbaoli, tanhaining, bjp}@ict.ac.cn

***Abstract*—Social media has became a critical manner for people to acquire information in daily life. Despite the great convenience, fake news can be widely spread through social networks, causing various adverse effects on people's lives. Detecting these fake news or misinformations has proved to be a critical task and draws attentions from both governments and individuals. Recently, many methods have been proposed to solve this problem, but most of them rely on the body content of the news, ignoring the social context information such as the comments. We argue that the comments of a specific news contain common judgements of the whole society and could be extremely useful for detecting fake news. In this paper, we propose a new method TRANSFAKE which jointly models the body content and comments of news systemically, and detects fake news with multi-task learning framework. TRANSFAKE model is a Transformer-based model. It takes different modalities as input and employs multiple tasks, *i.e.* rumor score prediction and event classification, as intermediate tasks for extracting useful hidden relationships across various modalities. These intermediate tasks promote each other and encourage TRANSFAKE making the right decision. Extensive experiments on two standard real-life datasets demonstrate that TRANSFAKE outperforms state-of-the-art methods. It improves the detection accuracy by margins as large as ∼2.6% and F1 scores as large as ∼5%.**

I. INTRODUCTION

Nowadays, social media plays an important role in daily life and changes the way people getting information. Users are convenient to create, access and share news which can be widely speared and affect other users through the social network. According to the Pew Research Center's survey, about 68% of adults obtain information from social media.[1]. Meanwhile, social media is also flooded with all kinds of fake news and misinformation, *i.e.* news stories with intentionally false information [1], [17], which misleads people's views and even results in wrong decisions. For example, during the 2016 U.S. presidential election, fake news of the two nominees was shared more than 37 million times on Facebook [1], [8] and affected the election result. Therefore, detecting fake news in social media has became an important task.

However, fake news detection is not trivial due to its multimodality and the huge volume. Fake news are delicately wrapped from truth facts and mixed in the real ones. The detection model should consider all the information containing in the news to make the right decision. Moreover, social media news are usually multimodal. It not only contains the textural messages but also attach some multimedia information, such as images and user comments. Figure1 shows an example of fake news from social media. Obviously, even though the content is difficult to distinguish whether the news is fake or not, it can be easily judged basing on the image. Moreover, the comments reflects the viewpoints from other users, which is apparently useful for fake news detection. The challenges in handling the multimodality lie in two aspects: (1) How to integrate the information from different modalities to detect fake news? (2) How to extract the viewpoints in user comments to assist the decision making?

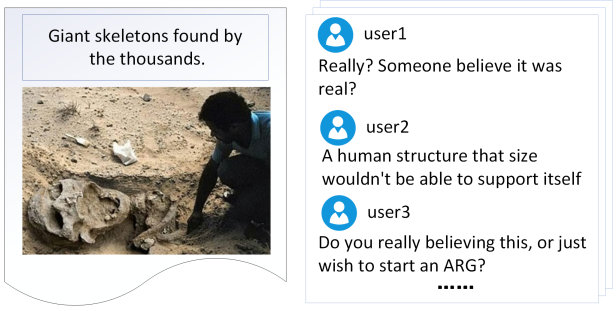

Fig. 1. An illustration of a piece of fake news and related user comments, which can be used for extracting useful information to assist fake news detection. Users often express disapproval for fake news.

To address these challenges, various methods have been proposed for detecting fake news. These methods can be categorized in two groups: *i.e.*, single-modal based methods [3], [4], [13], [15], [19] and multimodal based methods [7], [12], [21]. Methods in the first group only extract features from the textual or vision information for fake news detection. Ma et al. [15] and Jin et al. [13] use temporal-linguistic features and image features respectively to detect fake news. However, single modal information is usually insufficient for detecting,which leads the poor performance of this group methods. For multimodal-based methods, visual and textual information are integrated for detecting fake news. Jin et al. [12] predict fake news by combining visual, text and social context features using an attention mechanism. Khattar et al. [7] try to learn shared representations for text and visual with a auto-encoders model. Although these methods consider multimodal information, they model each modal of informa-

Jingping Bi (bjp@ict.ac.cn) is the corresponding author
[1]https://bit.ly/39zPnMd

tion separately and can not discover the correlations across different modalities. Meanwhile, some of the above work also use user comments, but only use statistical information, such as the credibility of user commenters, and seldom consider the semantic information of user comments.

To overcome the limitations, we propose a novel method, namely TRANSFAKE (short for multi-task Transformer for fake news detection), to fuse multimodal information and extract the viewpoints of social users to detect fake news. In general, TRANSFAKE is a multi-task learning framework which combines the supervision of fake news labels and user comments. We assume that fake news trend to having larger variance in the comments sentiment and utilize this information as an auxiliary supervision to learn the detection model. For the multimodal information, TRANSFAKE first transforms the related image into ROIs(Region of Interest) and feed them into a Transformer encoder along with the textural news content. Then, The self-attention mechanism in the encoder can automatically fuse the information from both textural and image data to support the detection.

Compared with the existing methods, TRANSFAKE has two attractive characteristics. On the one hand, the transformer encoder can automatically fuse the information of different data modalities, rather than simply splicing the features of them. It can effectively learn the cross-modalities relationship between textural and image to generate deep fusion features for fake news detection. On the other hand, we propose to use the sentiment variance of user comments to assist the model training, which makes TRANSFAKE suitable for label insufficient situation. Moreover, TRANSFAKE is elastic to other data, which can input any useful information to make the model learn better features for fake news detection.

To summarize, the contributions of our work are as follows:

- To the best of our knowledge, we are the first to try to use multimodal information, *i.e.*, text, visual and comments, to detect fake news. Towards this end, we apply a novel transformer-based method for the fake news detection task, which can learn the joint embedding of multimodal information (textural, visual and comments) to detect fake news effectively.
- We enhance the proposed method by adding explicit supervision of user comments, taking into account the uncertainty present in the detection result of the model.
- TRANSFAKE is general for any other information. Taking the embedding of other information as the input, TRANSFAKE is able to learn the relationships of them automatically for fake news detection.
- Through extensive experiments on real-world datasets, we illustrate the effectiveness of the model in fake news detection. It outperforms the start-of-the-art multimodal-based methods.

II. RELATED WORK

Methods for detecting fake news can divide them into two categories according to the used data modality, *i.e.* single modality based detection and multiple modality-based fake news detection. In this section, we summarize the related works respectively.

Single Modality-based Fake News Detection. The single modality-based methods mainly utilize one type of data, such as news contents [3], [9], visual [13] and social contexts [21], to detect fake news. Previously, Castillo et al. [3] detect fake news exploiting a set of features from the news content, *i.e.* characters and sensational emotions. Due to the requirement of corresponding domain knowledge, manually extracting features from huge volume of news is infeasible. Ma et al. [15] employ deep neural networks to detect fake news by capturing temporal-linguistic features in news content. Recently, visual features are proved to be an important indicator for fake news detection [13]. However, the work [10], [13], [22] that explored the visual features are still hand-crafted and hardly represent complex distributions of visual contents. Moreover, citeauthor wang2018eann [21] utilize social context features to detect news, such as the number of followers, hash-tag(#), and propagation patterns. However, methods in this category are only based on one modality. Omitting the information from other data modalities would harm the accuracy of fake news detection.

Multiple modality-based Fake News Detection. Methods in this category fuse data from different modalities to detect fake news [7], [12], [14], [21]. Jin et al. [12] fuse the visual, textual and social context features by attention mechanism. But the way they use social context is just to use statistical information. Wang et al. [21] learn the representation of text and image using an adversarial network. It helps to detect fake news because of learning the common characteristics shared between all events. Khattar et al. [7] tries to learn shared representations for text and visual with a fake news detector to detect fake news. Nevertheless, all these methods do not consider the relationship between the multi-modality information and no model directly extracts features from the user's comment to assist the detection of fake news.

III. METHODOLOGY

A. Overview of TRANSFAKE

TRANSFAKE aims to learn a multi-modal joint feature representation for fake news detection. As shown in Figure2, the proposed model integrates three main components: multi-modal feature representation, multi-modal information fusion, and multi-task learning. First, the multimodal feature representation layer is designed to obtain basic text and visual features. Each image is represented as a sequence of ROI features extracted from a Faster-RCNN [2] model, and each textual is represented as a sequence of words. After obtaining the features of text and images, we connect them into a series, and feed them into vision-language transfomer model to learn the joint representation of multi-modal features. To optimize the model parameters, a multi-task training loss is proposed. It not only contains fake news labels but also includes the review sentiment score on top of the joint representation to guide the model learning. The review sentiment loss minimize

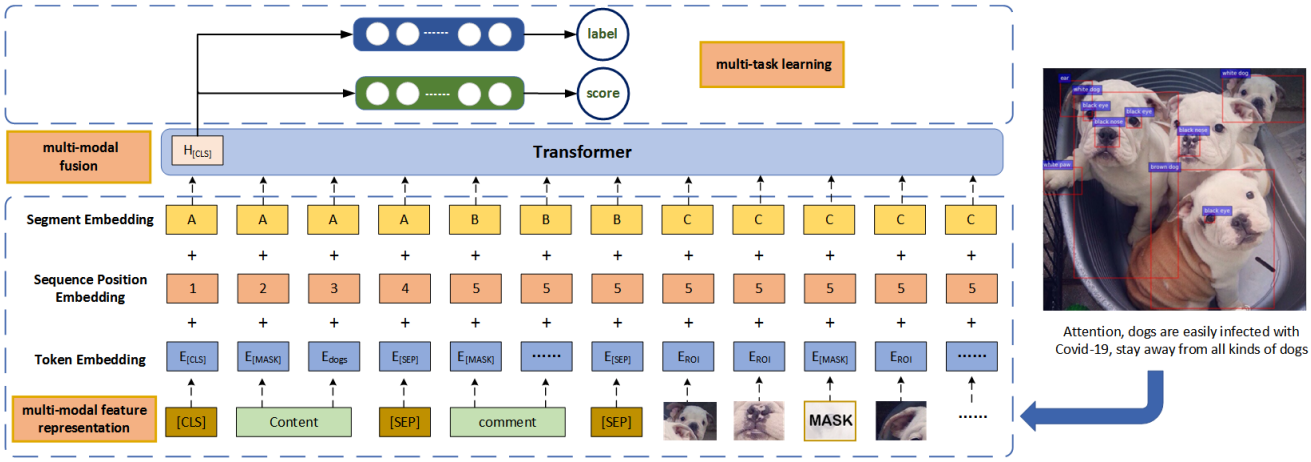

Fig. 2. Architecture of the TRANSFAKE.

the variance of sentiment score in real news to extract useful information for detecting fake news. Moreover, the fake news labels are also employed in TRANSFAKE as the direct supervision to learn task-specific representations to predict whether the post is fake or real. Next, we will describe the components of the model in detail.

B. Multi-modal Feature Representation

We first specify how we generate the basic representation for multimodal information through embedding layers. The embedding layers convert the inputs(*i.e.*, an image and linguistic) into two sequences of features: word-level embeddings and object-level image embeddings. Specifically, the linguistic includes tweet text and user comments.

Language Processing We first adopt the similar the word pre-processing method as BERT for tweet text and user comment. The input is split into n sub-word tokens w_1,w_2,...,w_N. Special tokens such as [CLS] and [SEP] are also added to the split tokens. To be specific, [CLS] is placed as the first token of the input, and [SEP] is added between the text and the comment. At the same time, there is also [SEP] among the comments. Similarly, [SEP] is also added between the image and the comment. Next, the word w_i is projected to vectors by embedding layer:

$$\hat{\mathbf{w}}_\mathbf{i} = WordEmbed(w_i) \tag{1}$$

We adopt the same policy as BERT that the final embedding of each token is generated by combining its original word embedding, segment embedding and position embedding.

$$\hat{\mathbf{p}}_\mathbf{i} = WordPositionEmbed(i) \tag{2}$$

$$\hat{\mathbf{w}}^\mathbf{s}_\mathbf{i} = WordSegmentEmbed(w^s_i) \tag{3}$$

$$\hat{\mathbf{h}}_\mathbf{i} = LayerNorm(\hat{\mathbf{w}}_\mathbf{i} + \hat{\mathbf{p}}_\mathbf{i} + \hat{\mathbf{w}}^\mathbf{s}_\mathbf{i}) \tag{4}$$

Language sequence segment and position embedding The sequence position of each token is embedded to indicate the order of input tokens. For the language part, we use ascending order to represent the order of the words in the tweet text. One thing to note here is that because the order of each comment has no effect on the detection of fake news, so we just use the same position index for all comments. That is, the index value is a continuation of the maximum index value of Twitter text. Besides, segment embedding is added to each input token to distinguish the different patterns. For the language part, this article defines two types of segments A and B to separate input elements from different sources, that is, to represent the content of tweets and user comments respectively. For example, for the input format of <content, comment1, comment2>, the segment should be <A, B, B>, A represents the content of the tweet, and B represents the user comment. The segment embedding is added to each input element so that it can learn which segment belongs to.

Vision Processing On the vision side, instead of using only one visual feature from vgg19 or other models. A Faster-RCNN model is used to extract object level visual features from the input images following [2]. This forces the model to reason at the object-level instead of at the pixel-level or global level. Due to the limitation of our computer resources, we just extract $N = 10$ bounding boxes and associated 2048-dimensional region-of-interest (RoI) feature:

$$(\mathbf{F}, \mathbf{P}) = FasterRCNN(\mathbf{IMG}) \tag{5}$$

Where IMG represents the attached image of the tweet. $\mathbf{F} = \mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_N}$ and $\mathbf{P} = \mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_N}$ are the ROI features and the bounding boxes respectively. Instead of directly using the RoI feature $\mathbf{f_i}$ without considering its position $\mathbf{p_i}$, we add position embeddings to image embeddings:

$$\mathbf{q^{(i)}} = (\frac{p_i^{tx}}{W}, \frac{p_i^{ty}}{H}, \frac{p_i^{bx}}{W}, \frac{p_i^{by}}{H}) \tag{6}$$

Where (p_i^{tx}, p_i^{ty}) and (p_i^{bx}, p_i^{by}) denote top-left and bottom-rigth coordinates of the bounding box p_i. Both the ROI features and position embeddings are projected into the same dimension with language embeddings. The final embedding result of an ROI is obtained via summing up its object

embedding, segment embedding, sequence position embedding and ROI position embedding:

$$\mathbf{v^{(i)}} = ImageEmbed(\mathbf{f_i}) \tag{7}$$

$$\hat{\mathbf{v}}^{(\mathbf{i})} = LayerNorm(\mathbf{v}^{(\mathbf{i})}) \tag{8}$$

$$\mathbf{s}^{(\mathbf{i})} = SegmentEmbed(f_i^s) \tag{9}$$

$$\mathbf{p}_{\mathbf{roi}}^{(\mathbf{i})} = PositionEmbed(\mathbf{q}^{(\mathbf{i})}) \tag{10}$$

$$\hat{\mathbf{p}}_{\mathbf{roi}}^{(\mathbf{i})} = LayerNorm(\mathbf{p}_{\mathbf{roi}}^{(\mathbf{i})}) \tag{11}$$

$$\mathbf{p}_{\mathbf{seq}}^{(\mathbf{i})} = SeqPositionEmbed(k) \tag{12}$$

$$\mathbf{e}^{(\mathbf{i})} = LayerNorm((\hat{\mathbf{v}}^{(\mathbf{i})} + \hat{\mathbf{p}}_{\mathbf{roi}}^{(\mathbf{i})})/\mathbf{2} + \mathbf{s}^{(\mathbf{i})} + \mathbf{p}_{\mathbf{seq}}^{(\mathbf{i})}) \tag{13}$$

where f_i^s in Eq.9 and k in Eq.12 are, respectively the type and position. Finally, the layer normalization (LN) is performed again.

Vision sequence segment and position embedding Unlike the position encoding of the tweet content, we just use fixed positions for all visual tokens, i.e., image ROIs, because there is no order of detected RoIs and any arrangement of them in the input sequence should get the same result. At the same time, the object's coordinates have been added to the image embedding. Segment embedding is also added to each input token.

C. Vision-Language Transformer

A multi-layer bidirectional Transformer [20] encodes the vision and linguistic embedding described above. The embedding features are transformed layer-by-layer in a manner that aggregates the features of other elements with adaptive attention weights. Let $\mathbf{x^l} = \mathbf{x_1^l}, \mathbf{x_2^l}, ..., \mathbf{x_N^l}$ be the features of the l-th layer. The features is computed in $(l+1)$-th layer as follows:

$$\mathbf{h}_{\mathbf{i}}^{\hat{\mathbf{l+1}}} = \sum_{k=1}^{N} \mathbf{W_k^{l+1}} \sum_{j=1}^{M} \mathbf{A_{i,j}^k} \cdots \mathbf{V_k^{l+1}} \mathbf{x_j^l} \tag{14}$$

$$\mathbf{h_i^{l+1}} = LayerNorm(\mathbf{x_i^l} + \mathbf{h}_{\mathbf{i}}^{\hat{\mathbf{l+1}}}) \tag{15}$$

$$\mathbf{x}_{\mathbf{i}}^{\hat{\mathbf{l+1}}} = \mathbf{W_2^{l+1}} \cdot GELU(\mathbf{W_1^{l+1}} \mathbf{h_i^{l+1}} + \mathbf{b_1^{l+1}}) + \mathbf{b_2^{l+1}} \tag{16}$$

$$\mathbf{x_i^{l+1}} = LayerNorm(\mathbf{h_i^{l+1}} + \mathbf{x}_{\mathbf{i}}^{\hat{\mathbf{l+1}}}) \tag{17}$$

where k in Eq.14 represents the index over the attention heads, and $\mathbf{A_{i,j}^k}$ denotes the attention weights between elements i and j in the k-th head. $\mathbf{W_k^{l+1}}$, $\mathbf{A_{i,j}^k}$, $\mathbf{V_k^{l+1}}$, $\mathbf{W_1^{l+1}}$, $\mathbf{W_2^{l+1}}$ and $\mathbf{b_1^{l+1}}$, $\mathbf{b_2^{l+1}}$ are learnable weights and biases, respectively.

D. Multi-Task Learning

We design several tasks to model the language and visual content, as well as their interaction. Each task has a simple task-specific network, and all of these networks share the output of the transfomer model. Therefore, we learned shared parameters and a set of task-related specific layer parameters. Our goal is to learn the parameters to minimize the loss of all tasks.

Task 1: Weak Supervision of visual-linguistic Alignment In order to enable the model to learn useful information from user comments, this article uses weak supervision to guide the optimization process of the model. In this part, we specifically explain how the weak label is combined in the transformer framework.

Generating Weak Labels The weak label we use is generated from user comments. Research shows user opinions towards fake news have more diverse sentiment polarity and less likely to be neutral [5]. A widely used tool VADER [11] is used to compute the sentiment scores, and then measure the standard deviation of all the scores. if a piece of news has a standard deviation of sentiment scores greater than a threshold β, then we set the weak label is 1, otherwise 0. To determine the proper thresholds for β. we vary the threshold values from [0, 1] through binary search, and compare the resultant weak labels with the true labels from the training set of annotated clean data, and choose the value that achieves the best accuracy on the training set. We set the threshold as 0.625 in the weibo dataset. As for GossipCop, due to the data usage policy, we cannot get the original data set, we directly use the data already processed in the work [18].

How to use the weak label There are many ways to use weak labels. We can predict weak labels through the model. As shown in Figure2, Multilayer perceptron(MLP) on top of the final output of the element [CLS] is used to predict the weak label. Among them, the last layer is a softmax layer. For the entry x in the training set D with ground truth labels y$\in\{0, 1\}$, We apply the binary classification loss for optimization:

$$\mathcal{L}_w = -E_{x \in D}(y \log(h_w(\hat{\mathbf{x_o}})) + (1-y)\log(1 - h_w(\hat{\mathbf{x_o}}))) \tag{18}$$

where $\hat{\mathbf{x_o}}$ is the final output feature of the [CLS] element, and $h_w(\hat{\mathbf{x_o}})$ is the predicted output.

In addition, we can use weak labels just as part of the model input. In other words, this weak label will be input into the model along with the content of the tweet and the image information. This way of using weak labels only makes them as input, without any other actions such as predictions. We will verify which method is better through experiments latter.

Task 2: Fake news detection In order to be able to identify fake news, just like predicting weak labels, the output of [CLS] also needs to be used. We add multiple layers to predict the final label. Here, we also use cross-entropy loss function:

$$\mathcal{L}_f = -E_{x \in D}(y_f \log(h_f(\hat{\mathbf{x_o}})) + (1-y_f)\log(1 - h_f(\hat{\mathbf{x_o}}))) \tag{19}$$

where y_f is the label for x, and $h_f(\hat{\mathbf{x}_o})$ is the classifier output.

Task 3: Masked Language Modeling (MLM) This task is the same with the MLM task in BERT [6] training. The words is masked randomly with a probability of 15% and the model will predict these masked words. The masked word is replaced with a special token [MASK], a random token or remains unchanged with a probability of 80%, 10%, 10%, respectively. We use the cross-entropy (CE) loss:

$$\mathcal{L}_{mlm} = -E_{x \in D} \sum_{j=1}^{M} CE(s(\mathbf{w_o^j}), h_k(\hat{\mathbf{w_o^j}})) \tag{20}$$

The j^{th} of the M masked tokens in text is denoted as w_o^i and $s(\mathbf{w_o^j})$ is the ground truth label. The output vector corresponding to the language masked token from the Transformer is $\hat{w_o^j}$. we add a fully-connected layer to predict the correct word and the output is $h_k(\hat{\mathbf{v_o^j}})$.

Task 4: Masked ROI Regression (MRR) MRR models the visual content. The task is to understand the content of the image from a deeper level, and its purpose is to be able to infer the embedding features of the masked image from the text or image information. We use an L2 loss to regress the feature.

$$\mathcal{L}_{mrr} = -E_{x \in D} \sum_{i=1}^{N} ||(m(v_o^i) - r_v(\hat{v_o^i}))||_2^2 \tag{21}$$

$m(v_o^i)$ is the embedding features. A fully-connected layer is added on top of the transformer output to project it to the same dimension as $m(v_o^i)$, denoted as $r_v(\hat{v_o^i})$.

The full objective function for the model is defined as follows.

$$\mathcal{L} = \lambda_1 \mathcal{L}_w + \lambda_2 \mathcal{L}_f + \lambda_3 \mathcal{L}_{mlm} + \lambda_4 \mathcal{L}_{mrr} \tag{22}$$

$\lambda_1, \lambda_2, \lambda_3$ represents the weight of each loss. We set the values to 1, 1, 0.005, 0.005 respectively.

IV. Experiments

In this section, we provide an overview of the datasets and state-of-the-art fake news detection approaches used for experiments and evaluate the effectiveness of our method. We aim to answer the following questions:

•**EQ1**: What is the performance of TransFake comparing with state-of-the-art fake news detection methods.

•**EQ2**: What is the influence of the correlations between different modalities?

•**EQ3**: How the supervised information provided by the comments influence the performance of TransFake.

Experiment settings. Our model is a 12-layer Transformer with 768 hidden units, 2048 intermediate units, and 12 attention heads. We set dropout probability to 0.1, and use GELU as activation function. The maximum length of the tweet content we use is 30. The max number of comment is 10 and maximum length of each comment is 30. During training, we use a batch size of 48 and a learning rate of 2e-5 with Adam optimizer. We train the model for 400 epochs

A. Datasets

We make use of two standard datasets to evaluate our model, which are called FakeNewsNet [16] and Weibo. To the best of our knowledge, these are the only available datasets that have paired image and textual information, including comments. FakeNewsNet dataset is collected from two-checking websites: Gossip[2] and PolitiFact[3]. The FakeNewsNet dataset was collected between January 1, 2010, and June 10, 2019 from 13 News sources, including mainstream British News media (such as the BBC and Sky News) and English-language Russian News media (such as RT and Sputnik). Due to the data usage policy, we cannot get the original dataset, we just directly use the dataset already processed in the work [18]. We ignore PolitiFact dataset because there is no weak label provided in [18]. The Weibo dataset which was used in [7], [12], [21] is collected from authoritative sources of China, such as Xinhua New Agency and Weibo, a Chinese microblogging website. We first remove Weibo without pictures and user comments. And then we process the dataset using a similar step as in [12]. Removing the duplicated and low-quality images to ensure the quality of the dataset. Finally, the dataset is split into training, validation and testing sets with an approximate ratio of 7:1:2 as in the work [21]. It should be noted that since there are no images in the gossip dataset we are using, the following model containing image information as input are not applicable to the gossip dataset. We directly ignore the image information and use other information when using the model.

B. Baselines

To validate the effectiveness of our model, we choose state-of-the-art fake news detection algorithms that can be divided into three categories: single modality models, multi-modal models, and the variant of the proposed model.

Single Modality-based Models

- **CNN**: This is a naive model that only uses textual information for detection. We use CNN to extract features following a fully connected layer with *softmax* function to predict whether the news is fake or not.
- **Vis**: This model only uses images to predict the news. Associated images are fed into a pre-trained VGG-19 model and fully connected layer sequentially to make predictions.

Multiple Modality-based Models

- **EANN** [21]: It is a novel event adversarial neural network framework that can learn transferable features for unseen events. We also use a variant of the model, named **EANN-** . The variant do not include the event discriminator.
- **MVAE** [7]: It is the state-of-the-art method for fake news detection which learns a shared representation of multi-modal information and uses a variational auto-encoder to discover correlations across modalities in tweets.

[2]https://www.gossipcop.com/

[3]https://www.politifact.com/

TABLE I
THE RESULTS OF DIFFERENT METHODS ON WEIBO DATASETS

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F1	Precision	Recall	F1
CNN	0.798	0.790	0.800	0.800	0.800	0.790	0.800
Vis	0.631	0.700	0.670	0.690	0.530	0.560	0.550
MAVE	0.790	0.830	0.820	0.820	0.730	0.750	0.74
EANN-	0.812	0.900	0.770	0.830	0.720	0.870	0.790
EANN	0.829	0.930	0.780	0.850	0.730	0.910	0.810
TRANSFAKE	0.855	0.850	0.930	0.890	0.870	0.740	0.800

TABLE II
THE RESULTS OF DIFFERENT METHODS ON GOSSIP DATASETS

Methods	F1	Accuracy
CNN	0.74	0.73
MAVE	0.77	0.77
EANN	0.77	0.74
TRANSFAKE/t	0.77	0.76
TRANSFAKE/l	0.83	0.82
TRANSFAKE	0.85	0.85

- **TRANSFAKE**: The proposed model for jointly learning with text, image and user comments for fake news detection.

Ablation Models

- **TRANSFAKE/t**. In this method, we just use tweets text as input, and the output is only classification task of fake news.
- **TRANSFAKE/b**. We use tweets text and image as input, and the output is only classification task of fake news. We can see the impact of the image on the models performance.
- **TRANSFAKE/l**. We use tweets text, image and user comments as input, and output the predictions results.
- **TRANSFAKE/e**. We use event discriminator used in [21] as the weak label for comparison for Weibo dataset.

C. RESULTS AND ANALYSIS

1) Performance Comparison: To answer **EQ1**, we compare TRANSFAKE with all the baseline methods. As shown in TableI, we can observe that the performance of TRANSFAKE is better than the baselines in terms of accuracy and F1. It increases the accuracy from 82.9% to 85.5 % and shows an increase of $\sim$2% in F1 scores compared to the previous best baseline. This proves the effectiveness of TRANSFAKE in fusing multi-modal information. Compared with the single-modal method, TRANSFAKE uses image, text and user comments to extract useful information for detecting fake news. Compared with the multi-modal method, TRANSFAKE is able to model the correlations among multiple modalities and achieve better performance.

TableI shows the experimental results of the Weibo dataset. According to the results, we have three key observations. First, compared with visual content, text contains more prominent features to detect fake news. Due to the diversity of visual content released by users, it is difficult to extract common and effective features. Therefore, the accuracy of Vis is the lowest among all methods. Second, the multi-modal methods are better than the methods based on single modality. It proves that the integration of multiple modalities is necessary for the task of detecting fake news. Third, among all multi-modal methods, EANN outperforms MAVE, which shows that applying adversarial mechanisms to learn common features can help improve the performance. For the variant EANN-, it does not contain an event discriminator, so it can only capture event-specific features. This will lead to the failure of learning shared features, and the accuracy is not as high as EANN.

TableII shows the experimental results on the gossip dataset. Since no image is provided in the dataset, we only use the single-mode models or the multimodal models with the image input removed as baseline approaches. We report the F1 score and the accuracy. Our proposed model outperforms the baseline models by a huge margin and increases the accuracy from 74% to 85 % and increases the F1 scores from 77% to 85%. This is mainly due to the weak label used by our model when prediction.

2) Ablation study: To answer **EQ2** and **EQ3**, we perform ablation experiments to analyze TRANSFAKE by comparing it with four ablations. The results are shown in Figure 3.

According to Figure3(a), TRANSFAKE/b is infier to TRANSFAKE when using tweet content and visual for our model. The performance improves compared to TRANSFAKE/t that only use text as model input. We can see that the image provide effective information for the detection. Through the result of TRANSFAKE/l, we can observe that the accuracy is further improved by %3 when user comment is added. It proves user reviews contain useful information to guide the detection. The accuracy of the model increases with the increase of input information, which is consistent with the intuition.

By comparing TRANSFAKE/l and TRANSFAKE, we can see that the accuracy of TRANSFAKE is 1.1% higher than that of TRANSFAKE/l. It indicates the semantic information of user reviews will harm the performance when directly feeding

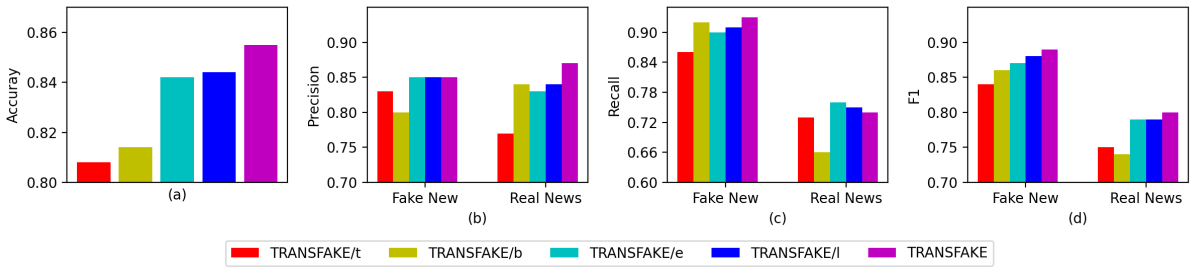

Fig. 3. The experimental results of each model on Weibo datasets. The model proposed in this paper has the highest accuracy compared with other variants. Overall, it also has better advantages in precision, recall and f1.

the reviews into the model. This is mainly because the user reviews contain too much information, and the model is not sure what kind of information should be extracted. Meanwhile, this shows the effectiveness of adding a user reviews score prediction module in TRANSFAKE, which can force the model to learn useful knowledge for fake news detection from user reviews.

In addition, as shown in Figure 3(c), the recall of TRANSFAKE is a little lower than that of TRANSFAKE/l and TRANSFAKE/e for real news. This is because our method uses a deep fusion method to fuse information from multiple modalities and is more sensitive to fake news. Therefore, some real messages are incorrectly identified as fake news. Moreover, we can observe that the metrics in real news are smaller than fake news in Figure 3(c) and Figure 3(d). The reason is that the generalization of all the model is not strong. Specifically, the training set data is limited and cannot contain all true news or a common representation of true news. However, all models tend to judge the news in the test set to be true based only on the patterns learned from the true news in the training set, and recognize the true news in the test set but not included in the training set as fake news. As a result, the recall rate and F1 indicators are slightly lower.

On the Gossip dataset, similar results can be observed as those on the Weibo dataset. From Table II, We can clearly see that the variant of the proposed model TRANSFAKE/l outperforms all the multi-modal approaches. For the proposed TRANSFAKE, it outperforms all the approaches on accuracy and F1 score. Compared with variants on the Weibo and gossip dataset, we can conclude that using the supervised information provided by the comments indeed improves the performance of fake news detection.

V. CONCLUSION AND FUTURE WORK

In this paper, we study the problem of multimodal fake news detection. The challenge of detecting fake news is how to effectively integrate various types of information, such as tweets text, visual and social context. To address this issue, we apply a transformer-based learning framework which can fuse multimodal information to learn shared representations to aid fake news detection. In addition, we utilize weak label signals to further promote multimodal fusion to improve the accuracy of detection. Meanwhile, we explore multi ways to use weak labels to acquire which way will get higher accuracy. Extensive experiments in real-world datasets show that the performance of the proposed model outperforms several state-of-the-art baseline algorithms. In the future, we want to extend other types of weak labels from social context to further improve our models. In addition, We will explore how to detect fake news early based on this model.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supposed by the National Natural Science Foundation of China (Grant No.62002343).

REFERENCES

[1] Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of economic perspectives **31**(2), 211–36 (2017)
[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
[3] Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web. pp. 675–684 (2011)
[4] Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology **52**(1), 1–4 (2015)
[5] Cui, L., Wang, S., Lee, D.: Same: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 41–48 (2019)
[6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
[7] ar Dhruv, K., Singh, G.J., Manish, G., Vasudeva, V.: Mvae: Multimodal variational autoencoder for fake news detection. In: Proceedings of the 2019 World Wide Web Conference. ACM (2019)
[8] Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., Zha, H.: Fake news mitigation via point process based intervention. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1097–1106. JMLR. org (2017)
[9] Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: Tweetcred: Real-time credibility assessment of content on twitter. In: International Conference on Social Informatics. pp. 228–243. Springer (2014)
[10] Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM International Conference on Data Mining. pp. 153–164. SIAM (2012)
[11] Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media (2014)

[12] Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 795–816 (2017)
[13] Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia **19**(3), 598–608 (2016)
[14] Liu, Y., Wu, Y.F.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
[15] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks (2016)
[16] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
[17] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017)
[18] Shu, K., Zheng, G., Li, Y., Mukherjee, S., Awadallah, A.H., Ruston, S., Liu, H.: Leveraging multi-source weak social supervision for early detection of fake news. arXiv preprint arXiv:2004.01732 (2020)
[19] Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S., de Alfaro, L.: Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506 (2017)
[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
[21] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: 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. pp. 849–857 (2018)
[22] Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on sina weibo by propagation structures. In: 2015 IEEE 31st international conference on data engineering. pp. 651–662. IEEE (2015)