ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0007

A Multimodal Fake Information Detection Method Incorporating Social Attribute Features

Lulan Zuo[1, a, *], Weiguo Wang[1, b] and Zhiyong Zhang[2, c]

[1]Information Engineering College, Henan University of Science and Technology, Luoyang, Henan 471023, China

[2]Henan International Joint Laboratory of Cyberspace Security Applications, Henan University of Science and Technology, Luoyang, Henan 471023, China

[3]Henan Representative Office, New H3C Technologies Co., Ltd. , Zhengzhou, Henan 450040, China

[a]zuolulan5079@163.com, [b]weiguowang@163.com, [c]xidianzzy@126.com

Abstract

In the fake information identification task of social media network platforms, it is difficult for unimodal detection models to perform accurate identification for fake information combined with images and texts. It is a challenge to introduce social attribute features into the disinformation detection model, not only for text, but also for information in social media which usually contains pictures and social attribute information. To this end, this paper proposes a multimodal disinformation detection method (att-XDSF) based on the xDeepFM algorithm and attention mechanism. In this model, text features and social attribute features are first fused by the attention mechanism, and then fused with image features in the same way and sent to the disinformation recognizer for classification. att-XDSF model achieves an accuracy of 81.2% on the Microblog dataset, and the F1 value, accuracy, and recall metrics are better than existing models.

Keywords

Feature extraction; Multimodal fusion; Fake information detection; Social feature.

1. INTRODUCTION

With the rapid development of social media network platforms and self-media, the number of users of social platforms has increased dramatically, and more and more people rely on social platforms, and more people publish, obtain and disseminate information on social platforms, and use them as the primary channel to obtain information[1]. But at the same time, social media platforms have also become a hotbed for the proliferation of fake information. Fake information with graphics and text is very confusing and spreads very fast. Especially in case of emergencies, such as the new epidemic in early 2020, fake information about the epidemic spread all over the Internet, causing a public outcry and threatening the safety of citizens and social stability[2]. Therefore, if fake information is not identified in a timely manner, it may cause large-scale negative impact on society and irreparable damage.

Depending on the data objects used, fake information detection methods can be divided into three categories: information content-based detection methods, user-based methods and dissemination-based methods. Content-based methods are mainly based on the content of the information itself, including text, images and videos. User-based methods mainly identify fake information by the identity attribute information of posters, such as gender, fan base and IP. The dissemination-based method mainly uses information such as comments and retweets in the process of information dissemination to identify fake information. Text and image information have been researched and proven to be effective in disinformation detection tasks, but single-text and single-image detection models usually tend to ignore the problem of text-image mismatch, and using multimodal fusion features can effectively solve the above problem. Therefore, multimodal fusion-based disinformation detection has received much attention. Zhou et al[4]. proposed a model using similarity-awareness to study the correlation between text and images before and later used for disinformation detection. However, these works ignore the connection between social attribute feature categories, for example, the number of references to a specific url in a text message, the number of #some topic, the number of @ some user, the use of special punctuation marks in the text, such as question marks? and exclamation marks ! etc.

In order to solve the above problems, this paper proposes a multimodal fusion based on xDeepFM and convolutional neural network for fake information detection. The main contributions of this paper can be summarized as follows.

(1) Taking social attribute characteristics as a mode, the fake information detection model is introduced, and the interactive information of explicit and implicit features in the social attribute feature is fused based on the xDeepFM model.

(2) The existing multimodal model fusion features are mostly splicing and direct addition, which cannot effectively combine the advantages of different modes. In this paper, attention mechanism is used for multi-feature fusion, which helps to enhance shared representation.

(3) Experiments on Weibo data sets show that the accuracy rate reaches 81.2%. Compared with the existing multimodal detection model, att-XDSF has better performance.

The contents of other chapters in this paper are as follows: The second chapter introduces the relevant literatures. The third chapter introduces the model proposed in this paper and its sub modules. The fourth chapter introduces the data set, ablation experiment of single mode, multi-mode contrast experiment, experimental results and example analysis.

2. RELATED WORK

In recent years, fake information detection has become the focus of research at home and abroad. Fake information detection is defined as a two-category problem, namely true information (T) and fake information (F). Suppose p is a data set of fake information. For any post $p_i \in P$, the fake information detection task can be described as function $f : f \to y$, where y is the label value of the post and $y \in \{0,1\}$.

In this paper, the content-based fake information detection is studied. The content-based fake information detection is mainly divided into three categories: text-based methods, image-based methods and multimodal fusion methods. Ma et al[5]. first applied depth learning technology to fake information detection, that is to obtain the hidden features of the texts through the recurrent neural network (RNN). Ma et al[6]. applied the idea of confrontation training to fake information detection for the first time, thereby proposing a model based on generative adversarial networks (GAN), that is, to expand the amount of data through confrontation training, confrontation training generator and discriminator, and improve the robustness and accuracy of the model. Cheng et al[7]. used a variational automatic encoder to self-encode text information to obtain text features and conduct multi task learning. In addition to text features, image features are effective in the field of fake information detection[8].

With the rise of deep learning, deep neural networks and pre-training models have powerful feature extraction functions, such as text feature extractor, Bert, transformer, VGG or ResNet, which are used to extract image features. In order to take advantage of the complementarities of different modal features, more and more research has been carried out to use text and image information fusion for fake information detection. Singhal et al[10]. used BERT and VGG19 to extract text and image features respectively, which will be input into the classifier after splicing to classify fake news. Khattar et al[11]. proposed a multi-mode variable automatic encoder (MVAE). In addition to extracting multimodal features, MVAE designed a news reconstruction task. The visual and textual information of the news is encoded by encoder, and the visual and textual information is reconstructed by decoder, and the multimodal information of the news is better fused by the reconstructing task. Finally, the news embedding obtained by encoder is input to the classifier to get the classification of news. Jin et al[12]. proposed a recursive neural network (att RNN) with attention mechanism o fuse multimodal features. This model combines text, image and social features, and uses the fused features for fake information detection. Wang et al[13]. proposed an event countermeasure neural network (EANN), which uses VGG to extract visual features, and Text-CNN to extract text features. The visual information and text information are spliced to get the news representation. In order to make better use of multimodal information, EANN designed an auxiliary task-event identification. The event discriminator takes the spliced multimodal news information as input and outputs the event category. Through auxiliary tasks, we can better understand multimodal information, thus helping to detect fake news. Qian et al[14]. proposed a multi-layered multi-modal context attention network (HMCAN), in which the co-attention is applied to train and enhance visual information using text information. Zhang et al. used the pre trained BERT to model text information and ResNet to model picture information. Finally, the multi header transformer is adopted to fuse the text information and image information to obtain better information representation.

To sum up, the existing multimodal fake information detection mainly includes text and images, ignoring the social attribute characteristics. To solve the above problems, this paper proposes a multimodal fake information detection model based on attention mechanism and using xDeepFM to integrate social attribute features (att-XDSF). The experimental results show that the performance of the model in Weibo data sets is improved compared with the current advanced multimodal fake information detection methods.

3. METHOD

3.1. Model Overview

Figure 1 shows the general architecture of att-XDSF model, which contains three modules: multimodal feature extraction module, multimodal feature fusion module and fake information detection module. The model uses BERT to extract text features F_t and fuse them with social attribute features F_s into the attention fusion layer to obtain feature F_{ts}; then it uses the attention mechanism to calculate the relevance of the image features F_v extracted from the vit network to form the final multimodal features $att - F_f$; finally, the fused multimodal features are sent into the disinformation detector to obtain the classification results.

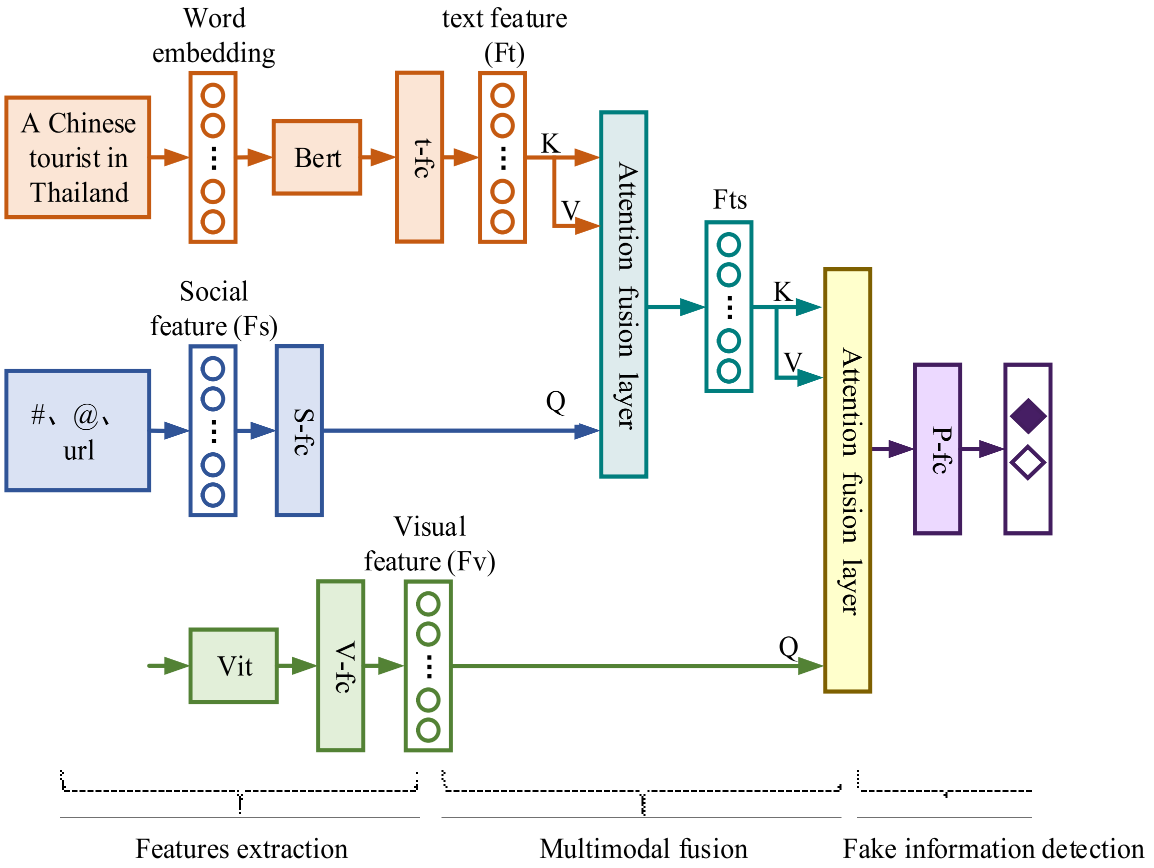

Figure 1. att-XDSF model

3.2. Text Feature Extraction Module

In order to capture the potential contextual meaning and semantic information of the text, this paper employs pre-trained BERT[16] model containing 12 encoder layers to extract text feature. Given the text content W, the input text represents as $W=\{W_0, W_1 \cdots W_m\}$, m represents words number in the text, and W_0 means embedding. Enter Wi to operate to obtain the text feature $R_w \in R^{d_w}$, where dw represents the dimension of text feature obtained from BERT model. Finally input the results into the fully connected layer whose activation function is tanh.

3.3. Image Feature Extraction Module

Vision Transformer(ViT)[17] has applied the Transformer structure to the classification of images in CV field. Compared with the current best convolutional neural network structure, ViT is still a preferred one. This paper uses ViT network to extract Extract image features from images. Thus, we can got the image features, $R_v \in R^{d_v}$, where d_v represents from dimensions obtained in vit. Finally input the results into the fully connected layer whose activation function is tanh.

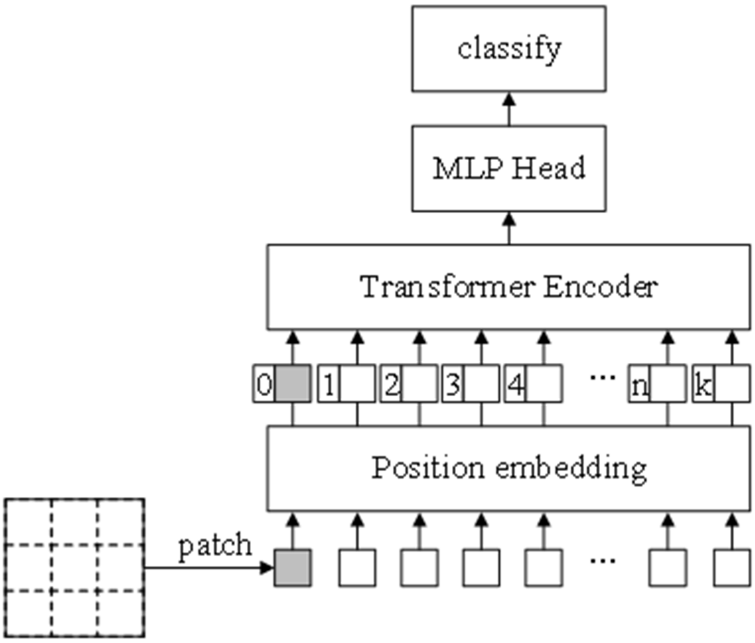

Figure 2. Vit

3.4. Social Feature Extraction Module

Posts in the social network platform contain rich social attribute information, which is effective for evaluating the credibility of posts, mainly including the following aspects:

(1) In the text message, the number of specific urls, # the number of a topic, and @ the number of a user are referenced

(2) The number of question marks ? and exclamation marks ! used in the text

(3) The emotion expressed by users when posting, for example, the number of positive words and negative words used by users

(4) Characteristics of users themselves, such as the number of fans on the microblog platform

In order to make better use of social attribute features, the text is based on the xDeepFM[15] algorithm and uses neural network to automatically and efficiently learn the high-dimensional feature interaction between implicit and explicit features, and establish the relationship between different categories of social attribute features. The overall structure of the xDeepFM model is shown in Figure 3, which is divided into DNN, linear and CIN modules. The Linear part is a simple linear regression that takes the original features without embedding processing as input. Compressed Interaction Network (CIN) is used to learn explicit high-dimensional feature intersection and make feature interaction occur at the vector level. After DNN is embedded, the CIN module is connected and connected to a simple multi-layer perceptron.

In this paper, we first use one-hot coding to represent social attribute features as vectors, and the dimension of the vectors is the size of social attribute feature categories, and then define a feature vector for each social attribute feature, and represent the correlation between feature categories by calculating the inner product between vectors of each feature, so as to effectively solve the problem that social attribute features are difficult to use.

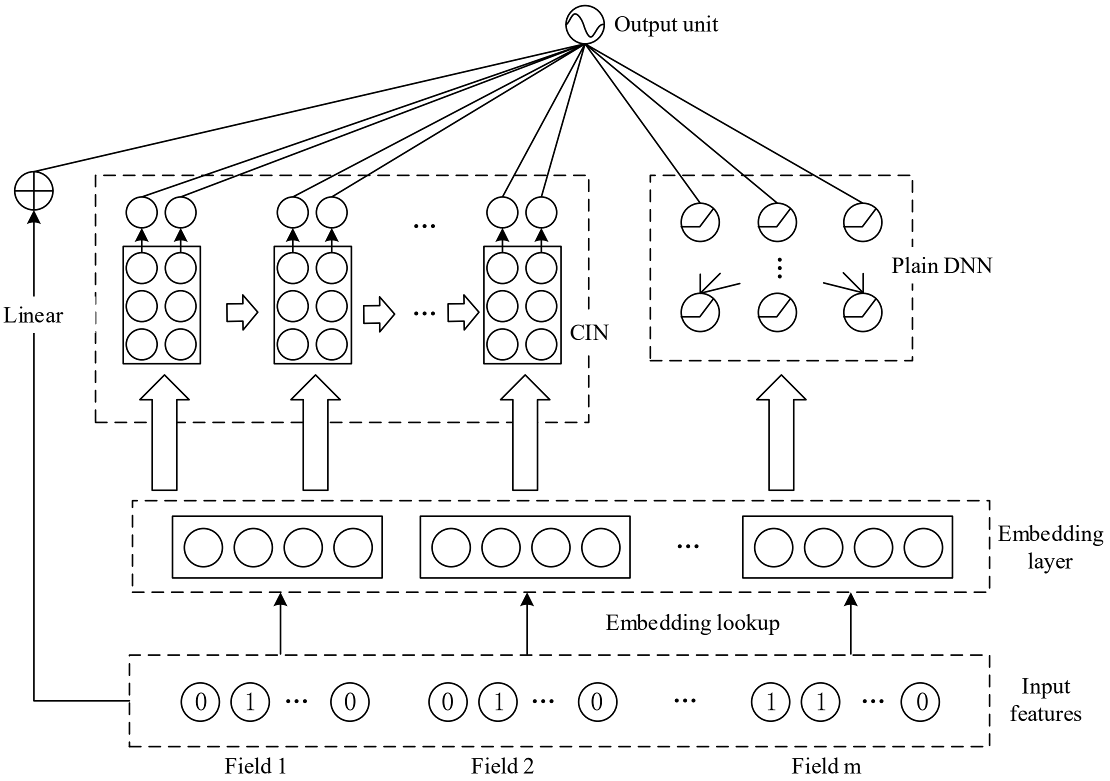

Figure 3. xDeepFM model

3.5. Multimodal Fusion Network

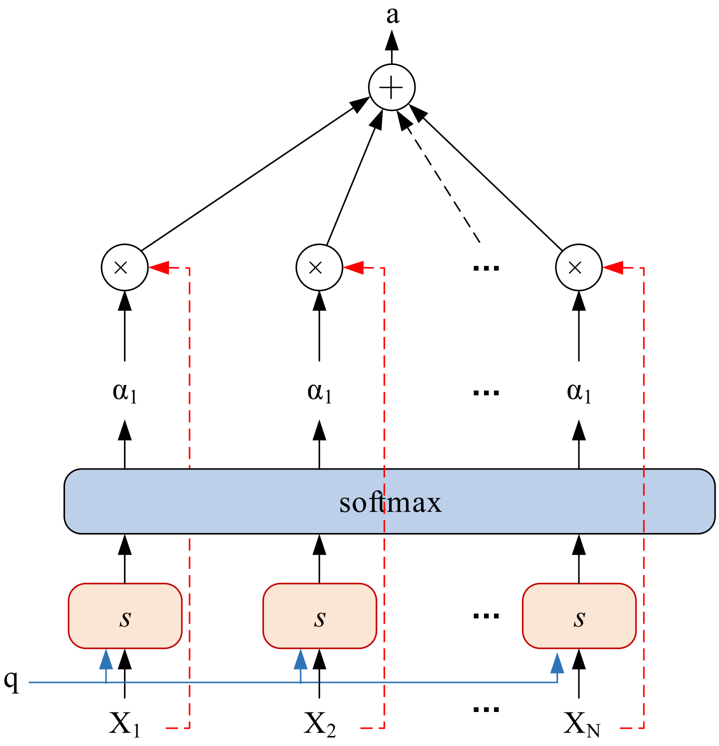

Figure 4. Multimodal Fusion Network

In this paper, two modes of modal fusion are used, namely direct splicing and attention mechanism fusion. The model of three feature stitching is directly recorded as XDSF, and the model of using attention mechanism fusion is recorded as att-XDSF. In order to test the performance of att-XDSF model, XDSF will be used as a comparative model in the experimental part of Chapter 4.

(1) Feature stitching

The text feature F_t, picture feature F_v, and social attribute feature F_s are spliced into a vector with a dimension of $3p$ and recorded as $F_f = [F_t, F_v, F_s]$.

(2) Attention mechanism fusion

In order to capture the interaction between various modes, first of all, the attention mechanism should be used to calculate the correlation between text features and social attribute features to obtain the joint representation of text and pictures, which is recorded as F_{ts}. Then, the attention mechanism is used again to calculate the correlation between F_{ts} and image feature F_v, and the final fusion feature is obtained as $att-F_f$. The formula of attention mechanism is as follows:

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

Wherein, $Attention(\cdot)$ is the operation function of attention mechanism, Q,K,V are $query$ matrix, key matrix and $value$ matrix respectively, d is the scale factor to prevent the excessive value of molecular dot product, and its value is the dimension of input feature.

As shown in Figure 3, $K = V = F_t$, $Q = F_s$ are set, and feature fusion is performed according to the following steps

(1) Calculate the similarity between Q and K to get the attention score s

(2) The attention score s is normalized by the function $soft\max$, and the probability distribution with the sum of all weight coefficients being 1 is obtained

$$P_i = soft\max(s_i) = \frac{\exp(s_i)}{\sum_{j-1}^{p}\exp(s_j)} \tag{2}$$

(3) According to the weight coefficient, V is weighted and summed to obtain the final output F_{ts}.

$K = V = F_{ts}$ and $Q = F_v$ are set again. According to the above three steps, the final fusion feature $att-F_f$ is obtained.

3.6. Fake Information Detection

The fake information detector distinguishes whether the paste is fake information through the input multimodal fusion feature. It consists of two full connection layers containing activation functions. Fake information detector is defined as D: $D(att-F_f;\theta_D)$, where att-Ff is the fused multimodal feature and θ_D is the parameter set of fake information detector.

$$\hat{y} = D(att-F_f;\theta_D) \tag{3}$$

$\hat{y}$ indicates the probability that the fake information detector outputs this message as fake. In the process of model training, the cross-entropy function is selected as the loss function, and the formula is as follows:

$$L_D(\theta_B, \theta_D) = -\sum E_{(p,y)\in(P,Y)}[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \tag{4}$$

Where P represents the input sticker, Y represents the label corresponding to the sticker, and E represents the expectation. By minimizing the classification loss, the optimal parameters are to be obtained.

$$(\theta_B, \theta_D) = \underset{\theta B, \theta D}{\arg\min}\, L_D \tag{5}$$

In the test method in this paper, Adam was used as the optimizer to add dynamic constraints to the learning rate, thus making it fluctuate within a certain range. And the specific optimization calculation process of Adam was as follows:

$$\begin{cases} m_t = \mu m_{t-1} + (1-\mu)g_t \\ n_t = v m_{t-1} + (1-v)g_t^2 \\ \hat{m}_t = \dfrac{m_t}{1-\mu^t} \\ \hat{n}_t = \dfrac{n_t}{1-v^t} \\ \theta = \theta - lr\dfrac{\hat{m}_t}{\sqrt{\hat{n}_t}+\varepsilon} \end{cases} \tag{6}$$

Where g_t is the gradient of the time step t, while m_t and n_t are the first and second moment estimators of gradient, respectively, $\hat{m}_t$ and $\hat{n}_t$ are the corrections for m_t and n_t, μ, v, ε and l_r are super parameters.

4. EXPERIMENT

This section first introduces the experimental environment, and data sets used in the experiments; then the ablation experiments and multimodal comparison tests are conducted to verify the feasibility and advancement of the att-XDSF method.

4.1. Experimental Environment and Experimental Parameters Selection

The machine configuration and the environment for this experiment are: Intel i7 2. 20GHz (processor), 64GB (memory), RTX-2080 (GPU), Python (3.7.6), Pytorch (1.4.0).

4.2. Dataset

The microblogging dataset used in this paper was proposed by Jin et al. and has been widely used in the study of multimodal disinformation detection. The dataset collects genuine information published by Chinese authorities, such as Xinhua News Agency. The dataset contains all officially confirmed disinformation posts from May 2012 to January 2016. This dataset was pre-processed by first removing duplicate and low-quality images, and then dividing it 7:3 into a training set and a test set to ensure that it does not contain any identical events. Table 1 shows the detailed statistical information of this dataset.

Table 1. Dataset statistics

Dataset	Training set	Test set
Real information	3907	996
Fake information	3896	1002

4.3. Baseline Model

To verify the validity of the model in this paper, it is compared with two types of benchmark models, i.e., the single-mode model with the model in this paper as a branch and the current state-of-the-art multimodal model.

4.3.1. Single-mode Model

(1) Text: The text is input into the pre-trained Bert model to extract text features, and then it is input into the fully connected layer with softmax activation function for classification.

(2) Image: The image is input into the pre-trained vit network to extract image features, and then it is input into the fully connected layer with softmax activation function for classification.

4.3.2. Multimodal Model

(1) VQA. The purpose of the VQA (Visual Question Answer) model is to answer questions on the basis of a given image. The original VQA model was designed for a multiclass classification task, while this paper mainly focuses on a binary classification task. Therefore, the final multiclass layer is replaced with a binary class layer in the implementation of the VQA model.

(2) att-RNN. att-RNN model uses recurrent neural network with attention mechanism to fuse textual, visual and social contextual features and output joint representation.

(3) MVAE. MVAE is used to learn shared representations between text and images by using a variational autoencoder to reconstruct the input data to obtain shared representations and combining it with a classifier for fake information detection.

4.4. Results and Analysis of Comparative Experimental

To evaluate the performance of this paper's model att-XDSF, comparison experiments were conducted on the Weibo dataset, and the results are shown in Table 2. the XDSF model is a variation of att-XDSF, which is a stitching of text, image and social attribute features for better experimental comparison.

Table 2. Experimental results of att-XDSF and baseline model comparison on Microblog dataset

Method	Accuracy	Real Information			Fake Information		
		Precision	Recall	F1	Precision	Recall	F1
VQA	0.708	0.732	0.626	0.675	0.739	0.607	0.667
att-RNN	0.788	0.785	0.799	0.792	0.790	0.716	0.751
MVAE	0.714	0.769	0.758	0.763	0.784	0.732	0.757
XDSF	0.798	0.795	0.809	0.802	0.800	0.804	0.802
att-XDSF	0.812	0.801	0.807	0.804	0.806	0.869	0.836

4.5. Results and Analysis of Ablation Experiments

To get a clearer picture of the role of each feature in the model, ablation experiments were conducted on the Weibo dataset, and the results are shown in the table below.

Table 3. Experimental results of att-XDSF and baseline model ablation experiments on Microblog dataset

Method	Accuracy	Real Information			Fake Information		
		Precision	Recall	F1	Precision	Recall	F1
Text	0.685	0.732	0.626	0.675	0.739	0.607	0.667
Visual	0.596	0.785	0.799	0.792	0.790	0.716	0.751
Social	0.529	0.769	0.758	0.763	0.784	0.732	0.757
Text+Social	0.751	0.795	0.809	0.802	0.800	0.804	0.802
Text+Visual	0.786	0.801	0.807	0.804	0.806	0.869	0.836
Visual+Social	0.720	0.732	0.801	0.765	0.739	0.815	0.775
XDSF	0.798	0.795	0.809	0.802	0.800	0.804	0.802
att-XDSF	0.812	0.801	0.807	0.804	0.806	0.869	0.836

The following conclusions were obtained based on the results of the comparison and ablation experiments.

(1) Multimodal models perform better than unimodal models

From Table 3, it can be observed that the accuracy rates of the three unimodal models based on text, image and social attribute features alone are all lower than those of the multimodal models. The accuracy of the model based on text features alone is 0.685, which is higher than the other two models. the accuracy of the XDSF model is 0.798, and after removing the text features, the accuracy of the model based on picture and social attribute features (Visual+Social above) decreases to 0.720, indicating that the text features have more influence on the disinformation detection model than the other two features. The accuracy of the model based on picture features and social attribute features alone is 0.596 and 0.529, respectively. After removing pictures and removing social attribute features, the accuracy of the model based on text and social attribute features (Text+Social in the above table) and the model based on text and picture features (Text+Visual in the above table) is 0.751 and 0.786, respectively, which can be see that although the picture and social attribute features are effective for fake information detection, the accuracy is lower than that of the multimodal model.

(2) The performance of the model based on attention mechanism for feature fusion is higher than that of the direct splicing model

The multimodal model att-RNN outperforms VQA and MVAE, indicating that the use of attention mechanism for inter-modal information fusion helps to improve the model performance. the accuracy of XDSF model is 0.798 and that of att-XDSF model is 0.812, which is supported by the fact that the use of attention mechanism makes the model accuracy improve by 1.4 percentage points.

(3) Social attribute features help improve model performance

In the Weibo dataset, the accuracy of the model based on text features and image features (Text+Visual) alone is 0.786, and after incorporating social attribute features, the accuracy improves to 0.798. att-RNN model and att-XDSF model also utilize text, image and social attribute features, and in comparison, the att-XDSF model has better results. This indicates that piece of the model. Since the pictures in the microblogging dataset contain more noise, data noise is unavoidable in the actual detection task, so a single modality or detection by text and picture features only, the accuracy often does not reach the ideal state. In this paper, we use rich social attribute features, and then combine text and picture features to obtain more comprehensive feature information of postings, which can effectively improve the performance of the model.

The att-XDSF model proposed in this paper outperforms the existing baseline model in terms of accuracy, precision, recall and F1 value, and improves the accuracy from to 0.812, which verifies that the att-XDSF model is effective in detecting fake information in social media networks.

5. CONCLUSION

In this paper, we propose a multimodal fusion of fake information detection method based on xDeepFM and convolutional neural network, which introduces social attribute features as a modality into the fake information detection model and fuses explicit and implicit feature interaction information in social attribute features based on xDeepFM model. The use of attention mechanism for multi-feature fusion helps to enhance the performance of the shared representation finally on the dataset are improved.

The existing deep learning methods are effective in fake information detection. However, a large amount of data in the early stage needs to be trained to achieve ideal results under such learning method. Furthermore, in the early detection task of fake information, a large amount of data used for training cannot be collected in a short time, which leads to the failure to achieve good results by means of the existing models. Besides, most of the existing detection methods are based on the same language, but the real social network platform contains different languages. Naturally, existing models find it difficult to understand cross language information. That is to say, the detection methods of multilingual fake news need to be developed urgently. Finally, with the development of short video platforms, fake videos are widely spread on social networks. Therefore, how to fully mine and integrate voice features, text features and image features to judge the authenticity of video will also be one of the future research directions.

REFERENCES

[1] K. Shu, A. Sliva, S.H. Wang, et al: Fake News Detection on Social Media: A Data Mining Perspective, ACM SIGKDD explorations newsletter 19.1 (2017), p.22–36.

[2] C. Guo, J. Cao, X.Y. Zhang, et al: Exploiting emotions for fake news detection on social media, CoRR, abs/1903.01728 (2019).

[3] N. Xu, W. Mao, G. Chen: Multi-interactive memory network for aspect based multimodal sentiment analysis, Proceedings of the AAAI Conference on Artificial Intelligence (2019), Vol. 33 No. 01, p. 371-378.

[4] X Zhou, J Wu, R Zafarani: Similarity-Aware Multi-modal Fake News Detection, Advances in Knowledge Discovery and Data Mining, 24th Pacific-Asia Conference(2020),p. 354-367.

[5] J. Ma, W. Gao, P. Mitra, et al: Detecting rumors from microblogs with recurrent neural networks, Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), p.3818-3824.

[6] J. Ma, W. Gao, K.F. Wong: Detect rumors on twitter by promoting information campaigns with generative adversarial learning, The world wide web conference (2019), p.3049-3055.

[7] M. Cheng, S. Nazarian, P. Bogdan: Vroc: Variational autoencoder-aided multi-task rumor classifier based on text, Proceedings of the web conference (2020), p.2892-2898.

[8] P. Qi, J. Cao, T. Yang, et al: Exploiting multi-domain visual information for fake news detection, 2019 IEEE international conference on data mining (ICDM 2019), p.518-527.

[9] P. Gao, Z. Jiang, H. You, et al: Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), p. 6639-6648.

[10] S. Singhal, R.R. Shah, T. Chakraborty, et al: Spotfake: A multi-modal framework for fake news detection, 2019 IEEE fifth international conference on multimedia big data (BigMM 2019), p.39-47.

[11] D. Khattar, J.S. Goud, M. Gupta, et al: Mvae: Multimodal variational autoencoder for fake news detection, The world wide web conference(2019), p.2915-2921.

[12] Z. Jin, J. Cao, H. Guo, et al: Multimodal fusion with recurrent neural networks for rumor detection on microblogs, Proceedings of the 25th ACM international conference on Multimedia (2017), p.795-816.

[13] Y. Wang, F. Ma, Z. Jin, et al: Eann: Event adversarial neural networks for multi-modal fake news detection, Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining (2018), p.849-857.

[14] S. Qian, J. Wang, J. Hu, et al: Hierarchical multi-modal contextual attention network for fake news detection, Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021), p.153-162.

[15] J. Lian, X. Zhou, F. Zhang, et al: xdeepfm: Combining explicit and implicit feature interactions for recommender systems, Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, p. 1754-1763.

[16] J. Devlin, M. Chang, K. Lee, et al: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[17] A. Dosovitskiy, L. Beyer and A. Kolesnikov: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv preprint arXiv:2010.11929 (2020).