SAFS: Social-Article Features-Stacking Model for Fake News Detection

Xiaojun Wu
Dept. of Information Management
Peking University
Beijing 100871, China
wuxiaojun@pku.edu.cn

Jimin Wang*
Dept. of Information Management
Peking University
Beijing 100871, China
wjm@pku.edu.cn

2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 978-1-6654-1515-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIBD51990.2021.9459078

***Abstract*—With the rapid development of social media, people can more easily and quickly produce, transmit, receive and share information, but this also provides channels for the widespread dissemination of fake news. Therefore, it is of great practical importance how to properly detect fake news and curb its continued spread. In this paper, an integrated machine learning model for fake news detection, the Social-Article Features-Stacking Model (SAFS), is proposed based on the work of our predecessors, using natural language processing and social network analysis methods. The model achieves good experimental results by multi-feature fusion of textual features and social network features of fake news. From the experimental results on the FakeNewsNet dataset, the SAFS model greatly improves the f1 value of fake news detection by 16.86% compared with the Social-Article Fusion Model (SAF), the original paper model. Finally, the paper explores the effectiveness of different features in detecting fake news.**

Keywords—component, fake news detection, natural language processing, social network analysis

I. INTRODUCTION

With the rapid development and wide popularity of the Internet, social media has become an important channel for the majority of Internet users to obtain information. According to CNNIC survey, as of December 2018, nearly half of China's Internet users obtained news information through social media.

Currently, social media are the most popular channels for news consumption today due to their ease of access, dissemination and sharing [1]. However, the rapid development of social media has also provided a channel for the widespread dissemination of fake news. Fake news generally refers to news that formally imitates the content of news media but is actually fabricated, also known as false news, misinformation or rumor. Fake news is generated and disseminated in various fields such as social security, policy and government, natural disasters, economic development, culture, sports and entertainment, transportation, health and epidemic prevention, medical education, food and drug safety, and daily life, etc.

Therefore, we need to find some effective methods to detect fake news to prevent receiving wrong information. In this paper, a state-of-the-art model of fake news detection, Social-Article Features-Stacking Model (SAFS), is proposed that combines machine learning techniques and social network analysis techniques.

II. RELATED WORK

With the development and popularity of the Internet and social networks, fake news detection has begun to receive attention from the academic community in recent years. There are several typical public datasets in the current research area of fake news detection: FakeNewsNet [1], Buzzfeed Election Dataset & Political News Dataset [2-4], LIAR [5], and Twitter16 [6]. These datasets generally provide news content of fake news itself (author, headline, body, image, video, etc.) and social contextual content (user personal portrait, user interaction behavior, etc.), which are very good corpus resources for studying fake news.

Scholars have studied and implemented many models for fake news detection, which can be classified into the following 3 types according to their input features. Firstly, Fake news detection can be identified by extracting content features. News content features are mainly Linguistic-based and Visual-based. Linguistic features include: (1) lexical features including letter-level features and word-level features, such as the number of words in a sentence, the average number of letters per word, the number of common words, and the number of unique words [7]. (2) Syntactic features, which numerically encode the whole sentence by different encoding methods, such as one-hot encoding, N-gram encoding, and word class tagging. Visual features include features such as clarity values, relevance values, diversity values, and clustering values of images [8]. Fake news can also be detected by expert systems and knowledge base inference. This approach is more suitable for fake news detection in vertical domains (e.g., biology, history, etc.), where more objective facts are gathered through the knowledge of domain experts, and it is easy to reason about the truth or falsity of news based on these objective facts [9].

Fake news detection can take into account not only the content features but also the social context features of fake news dissemination. Kai Shu et al. from Arizona State University, USA, studied the correlation between user profiles (user profiles) and fake news on social media [10]. Meanwhile, the authors themselves compiled the FakeNewsNet dataset and constructed the publisher-news-user ternary relationship model tri-FN using a singular value decomposition correlation algorithm, which

978-0-7381-3170-2/21/$31.00 ©2021 IEEE

was experimentally effective over many baseline models [11]. Another study explored the propagation behavior of fake news [12]: the propagation wandering trajectory of fake news is tracked, and then the propagation characteristics of fake news are studied in graph and evolutionary models; this method can also identify the key propagators of fake news, which is crucial to mitigate the spread range of fake news on social media.

Through the review and analysis, this paper identifies the focus of the fake news detection research: building a faster and more efficient fake news detection model. The model should ideally also verify the extent to which different features contribute to the effectiveness of fake news detection, which can help us learn more about how machine learning and deep learning models can help us detect fake news.

III. SAFS Model for Fake News Detection

In this section, a fake news detection model constructed based on text features and social network features of fake news, SAFS (Social-Article Features-Stacking Model), will be presented. Firstly, this paper will introduce the dataset and the method of feature extraction, and then the construction process of this model will be presented:

A. Dataset

In order to study the techniques related to fake news detection, this paper will conduct an experimental study on the FakeNewsNet dataset. FakeNewsNet mainly consists of two separate parts, PolitiFact and Gossip Cop (which is a dataset for entertainment news), and this paper focuses on the PolitiFact dataset. This dataset uses the fact-checking website PolitiFact[1] to obtain news content for both fake news and real news. In the PolitiFact website, journalists and domain experts review political news and provide fact-checking assessments to determine whether news articles are fake or true. We use the results of these expert checks as a basis for determining whether a data set is true news or fake news. The dataset is stored in a CSV file and has the following fields: id, url, title, text, twitter_ids which include a list of twitter user ids that share the news. The dataset contains a total of 1056 news data, including 624 true news items and 432 fake news items.

B. Feature Engineering

In general, machine learning models and deep learning models cannot directly process text data, so we need to use the text representation model in natural language processing to transform text data into numerical data that can be processed by the model, and this process is the extraction of text features. In order to better represent the news features, the representative Tfidf model, n-gram probability model, and word2vec neural network model in text processing are used in this paper to extract the title text features.

In addition to using the text features, the social network features of news dissemination help to perform the identification of fake news, the feature representation of news points can be obtained by Graph Embedding technique (GEB). In this paper, social network features are extracted by news-based collaborative filtering algorithm and node2vec algorithm, and the extraction process is as follows:

- Constructing news_id-twitter_id interaction matrix using ids and twitter_ids from the original dataset.
- Build the inverted table of twitter_id-news_id by borrowing the User-CF algorithm from the personalized recommendation domain to obtain the news co-occurrence matrix.
- Construct an undirected weighted graph using the nodes with co-occurrence frequency.
- The constructed undirected weighted graph is fed into the node2vec model [13], which consists of two parts: a weighted random walk of the graph and a sequence feature extraction neural network. After that a 512-dimensional vector representation is obtained for each news item.
- By summing the social network features (512 dimensions) and word2vec features (512 dimensions) of each news item, we name the new features as social_w2v features (social network-word vector features).

C. SAFS Model Construction

Combining the fake news features extracted above, we first construct the base model using the XGBoost [14] model and the Tfidf features, bigram features, and social_w2v features. Then, we construct a model for detecting fake news, Social-Article Features-Stacking Model (SAFS), based on the idea of stacking model [15]. The SAFS model assumes that we have three XGBoost model training base models xgb_1, xgb_2, xgb_3 and a secondary-model logistic regression (LR), the SAFS is illustrated as Fig.1 shows: the training set training process and the test set prediction process form a twin-tower structure.

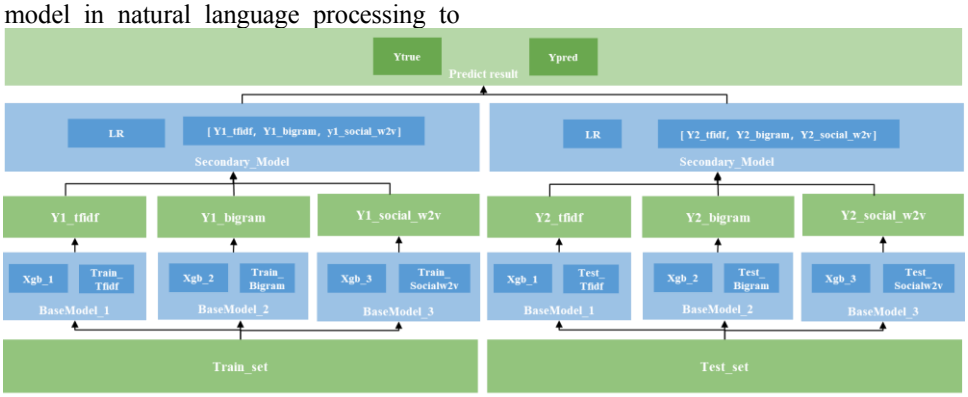

Fig. 1. SAFS Twin-tower structure model.

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

The training and prediction processes of the SAFS twin-tower structure model are described by mathematical equations as follows:

1) Equations (1) show the base model, xgb_1, is trained on the Tfidf training set and then used to predict the label columns Label for the Tfidf training set and the Tfidf test set, Y_{1_tfidf} and Y_{2_tfidf}, respectively.

$$\begin{pmatrix} \vdots \\ X_{train_{tfidf}} \\ \vdots \end{pmatrix} \overset{\text{xgb\_1 Train}}{\underset{\uparrow}{\Longrightarrow}} \begin{pmatrix} \vdots \\ Y_{True} \\ \vdots \end{pmatrix}$$

$$\begin{pmatrix} \vdots \\ X_{train\_tfidf} \\ \vdots \end{pmatrix} \overset{\text{xgb\_1 Predict}}{\underset{\uparrow}{\Longrightarrow}} \begin{pmatrix} \vdots \\ Y_{1\_tfidf} \\ \vdots \end{pmatrix} \quad (1)$$

$$\begin{pmatrix} \vdots \\ X_{test\_tfidf} \\ \vdots \end{pmatrix} \overset{\text{xgb\_1 Predict}}{\underset{\uparrow}{\Longrightarrow}} \begin{pmatrix} \vdots \\ Y_{2\_tfidf} \\ \vdots \end{pmatrix}$$

2) Similarly, xgb_2 and xgb_3 model are obtained by training on the Bigram set and the Social_w2v set and obtaining the corresponding labels: Y_{1_bigram} and Y_{2_bigram}, $Y_{1\_social\_w2v}$ and $Y_{2\_social\_w2v}$.

3) Then the secondary model LR is trained with the prediction results generated from the real training set as the training set Train_2 = $\{Y_{1\_tfidf}, Y_{1\_bigram}, Y_{1\_social\_w2v}\}$, and the prediction results generated from the test set as the test set feature set test_2 = $\{Y_{2\_tfidf}, Y_{2\_bigram}, Y_{2\_social\_w2v}\}$ for prediction, to obtain the final test set prediction results Y_{pred}, as (2) shows.

$$\overset{\text{Train\_2}}{=} \begin{pmatrix} \vdots & \vdots \\ Y_{1\_tfidf} & Y_{1\_bigram} & Y_{1\_social\_w2v} \\ \vdots & \vdots \end{pmatrix} \overset{\text{LR Train}}{\underset{\uparrow}{\Longrightarrow}} \begin{pmatrix} \vdots \\ Y_{True} \\ \vdots \end{pmatrix}$$

$$\overset{\text{Test\_2}}{=} \begin{pmatrix} \vdots & \vdots \\ Y_{2\_tfidf} & Y_{2\_bigram} & Y_{3\_social\_w2v} \\ \vdots & \vdots \end{pmatrix} \overset{\text{LR Predict}}{\underset{\uparrow}{\Longrightarrow}} \begin{pmatrix} \vdots \\ Y_{Pred} \\ \vdots \end{pmatrix} \quad (2)$$

Finally, we can use the prediction results of the test set with the real labels of the test set to evaluate the model effectiveness, and the evaluation metrics are the precision, recall, accuracy, and F1 values.

D. Model Comparison

After constructing the social network-text feature stacking model SAFS model, the fake news detection model in this paper, this section introduces two models that the SAFS model needs to be compared with: the first is the model given in the original paper on the FakeNewsNet dataset, and the second is the current natural language processing field the most cutting-edge BERT model.

The original FakeNewsNet paper provides the experimental results of seven models, from which we select four representative models, Social article fusion model (SAF) [16], as our comparison models, which the model proposed by the original authors Kai Shu et al. themselves has the best results. The SAF model proposed by the original authors uses autoencoders to learn the content features of news articles and features of user activities to classify news as fake news or true news.

BERT [17] (Bidirectional Encoder Representation Transformers) was proposed by Google in 2018, is a model built based on Transformer [18], BERT brings a breakthrough in natural language processing. The BERT model is very huge, so it poses a huge challenge to computer performance, and later people proposed compressed versions such as Distil BERT [19] and Tiny BERT [20] to reduce the number of parameters of the BERT model. In this paper, we use the pre-trained model in the version of bert_base_uncased[2] provided by the Huggingface team.

IV. EXPERIMENT

This section will then conduct experiments on the FakeNewsNet dataset using SAFS, and then compare the experimental effects with SAF and BERT Model. Also, this section will explore the effectiveness of different features for fake news detection

A. Experimental Settings

The experimental environment used in this paper (Tab.I), mainly using hardware for Intel Xeon 2GHz CPU, 16GB of RAM, NVIDIA P4 graphics card; software for the operating system is Ubuntu 18.0, deep learning acceleration environment for CUDA 10.1.

TABLE I. EXPERIMENTAL ENVIRONMENT INFORMATION

Hardware	CPU	Intel(R) Xeon(R) CPU @ 2.00GHz
	GPU	NVIDIA Tesla P4 8GB
	Memory	16GB
Software	Operation system	Ubuntu 18.04.3
	CUDA Version	10.1

Then we need to set up the dataset and the relevant hyperparameters of the model. We divide the dataset into training set: test set = 9:1 and set random number seeds to ensure reproducible experimental results. Three base models are implemented using the XGBoost toolkit [3] and the main hyperparameters are set as follows: n_estimators=120, learning_rate=0.08, max_depth=5. The secondary model LR is implemented using the logistic regression model provided by the sklearn toolkit, with main parameters set as: penalty='l2', dual=False, class_weight=None.

B. Experimental Result

We conducted experiments using Colab, a deep learning integrated environment provided by Google. The prediction results obtained using the SAFS model are shown in Tab.II and

[2] https://huggingface.co/transformers/pretrained_models.html

[3] https://pypi.org/project/xgboost/

Fig.2, and are compared with the model results of the SAF model, which is the best model in the original FakeNewsNet paper, the Xgb+Tfidf model, which is the best model in the base model, and the BERT model. Comparison result is shown in Tab.III.

TABLE II. MODEL RESULTS

Model	Accuracy	Recall	Precision	F1
Xgb_1	0.815	0.611	0.782	0.698
Xgb_2	0.479	0.944	0.552	0.636
Xgb_3	0.632	0.667	0.701	0.649
SAF	0.638	0.789	0.691	0.706
BERT	0.811	0.833	0.851	0.822
SAFS	0.750	0.917	0.839	0.825

TABLE III. MODEL COMPARISON

Model Comparison	Accuracy	Recall	Precision	F1
Ours vs. SAF	+17.55%	+16.22%	+21.42%	+16.86%
Ours vs. Xgb_1	-7.98%	+50.08%	+7.29%	+18.19%
Ours vs. BERT	-7.52%	+10.08%	-1.41%	+0.36%

According to the results in the Tab.II and Tab.III, 4 main conclusions can be seen that:

1) The prediction accuracy of the SAFS model reached 0.839, the F1 value reached 0.825. As shown in Fig.2, 33 of the 36 positive samples in the validation set were predicted correctly, with a recall rate of 0.917, which is excellent.

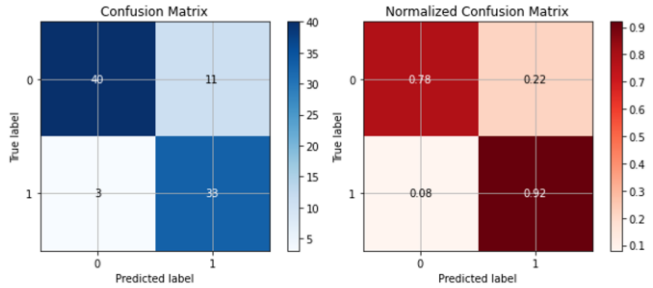

Fig. 2. Confusion Matrix of SAFS model.

2) The SAFS model substantially outperformed the SAF model of the original paper in all metrics.

3) Although the SAFS model is inferior to the Xgb+Tfidf model in terms of accuracy, the other three metrics are substantially better than the Xgb+Tfidf model.

4) SAFS is slightly inferior to the BERT model in terms of precision, accuracy and F1 value, and it is expected that the BERT model, as a model obtained by pre-training a large-scale corpus, has better generalization performance than SAFS. However, SAFS significantly outperforms the BERT model in terms of recall of fake news samples, and the SAFS model is more usable in the sense of fake news detection.

Moreover, in terms of computational speed, the training time of BERT model and SAFS model are 1h26min11s and 25min57s respectively, and the average inferring time of two models are 200ms and 30ms respectively, which shows that SAFS model is significantly faster than BERT model in terms of computational speed, taking only 30.11% of the training time and 15% of the inferring time of BERT.

C. *Result Discussion*

Stacking method is a common model integration method in machine learning, and SAFS has achieved good experimental results in the field of fake news detection based on this idea of integrated learning of multiple features and multiple base models. This is because the output of the primary learner can be regarded as a feature engineering process with good feature transformation, feature statute, and feature extraction of the original features, which is also known as the aggregated features. The good feature input helps the machine learning model to learn and output good prediction results, so it can be explained that SAFS achieves excellent results in fake news detection.

The advantage of the SAFS model is not only that the features and models can be "stacked" to get good fake news detection results, but also that the features can be "disassembled" to analyze the effectiveness of different features. This "disassembly" is actually the most frequently used control variable method in experiments, where we only change one factor in the model to study the effect of this factor on the model.

The result in Tab.IV leads to the following three conclusions.

1) Effectiveness of Tfidf features: improving the recall of fake news. SAFS-Tfidf model has some reduction in all indicators compared to SAFS model, but the most reduction is the recall rate, from 0.917 to 0.667, a reduction of 27.3%. It can be seen that the recall ability of the model for fake news samples is significantly reduced after removing the Tfidf feature, therefore, the Tfidf feature can effectively improve the recall ability of fake news in the SAFS model.

2) Ngram effectiveness: improving the prediction ability for non-fake news. The SAFS-Bigram model has the largest decrease in each index is the accuracy rate, which decreases

from 0.750 to 0.620, a decrease of 17.3%. Therefore, the Bigram feature in the SAFS model can effectively improve the prediction ability of the SAFS model for negative samples which are also non-fake news.

3) Effectiveness of social network features: greatly improving the recall of fake news. Contrary to the SAFS - Bigram model, the SAFS-Social_w2v model not only does not decrease but also increases the accuracy rate by 8.67%, indicating that the Social_w2v feature has a negative effect on the ability to predict negative samples with the SAFS model. However, the SAFS-Social_w2v model decreases from 0.917 to 0.611 in the recall rate, a decrease of 33.4%, which is the most reduced recall rate among the three demolition models. It can be seen that Social_w2v has a very important influence on the SAFS model to predict fake news samples. Predicting fake news is the main goal of this paper, so the importance of Social_w2v feature is self-evident.

TABLE IV. MODEL DISASSEMBLE

Model	Accuracy	Recall	Precision	F1
SAFS	0.750	0.917	0.839	0.825
SAFS - Tfidf	0.632	0.667	0.701	0.649
SAFS - Bigram	0.620	0.861	0.724	0.721
SAFS – Social_w2v	0.815	0.611	0.782	0.698

The main content of this section is to conduct experiments and analyze the experimental effects. The SAFS model has the advantages of good feature verifiability, scalability, and low computational overhead, and has both good experimental results and performance.

V. CONCLUSIONS AND DISCUSSION

The Internet allows all Internet users to create and share information quickly and easily, but it also provides a way for the spread of fake news. The spread of fake news will bring panic and harm to society. Therefore, how to detect fake news in advance and stop the spread of fake news at the source is of great importance to maintain a healthy, clean and true Internet environment. In this paper, we analyze and summarize the research progress and find that the current fake news detection methods can be divided into three main categories according to the classification of features: content feature-based methods, knowledge base based methods and social network analysis-based methods.

Based on these previous works, this paper builds a social network-text feature stacking model SAFS by drawing on the idea of Stacking in machine learning integration model, which makes a good integration of text features and social network features of fake news; SAFS achieves significantly better experimental results than SAF and BERT model, which has strong practicality. The main contribution of the paper is the design of a proven fake news detection model, which has obvious advantages in terms of experimental effects and computational overhead. Its innovations are mainly reflected in the following two points.

1) This paper draws on the user-based collaborative filtering recommendation algorithm, extracts the news embedding features using graph neural network, and finally fuses them with word2vec features to generate the final social network-text features. The experimental results show that the features can improve the recall rate of fake news detection very effectively.

2) The SAFS model achieves good experimental results through the stacking of base model and base features, and the effectiveness of the features is analyzed by the control variable method.

Of course, this paper also has problems such as small dataset and over-fitting of the model, which need further research and improvement.

Fake news detection is a relatively new cross-research field, which involves many fields such as natural language processing, image visual processing, social network analysis, etc. Multi-feature fusion and multi-model fusion should have a small computational overhead in addition to good experimental results, so that the model has good practicality and application prospects. Fake news detection is of great importance for maintaining network and social harmony, therefore further research is necessary.

ACKNOWLEDGMENT

The authors wish to sincerely thank the Deputy Editor-in-Chief, associate Editor and anonymous reviewers for their constructive and valuable comments which lead to the better presentation of this paper. The authors are partially supported by the key project of the National Social Science Foundation of China (20ATQ007).

REFERENCES

[1] Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36.

[2] Potthast M, Kiesel J, Reinartz K, et al. A stylometric inquiry into hyperpartisan and fake news[J]. arXiv preprint arXiv:1702.05638, 2017.

[3] Ruchansky N, Seo S, Liu Y. Csi: A hybrid deep model for fake news detection[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017: 797-806.

[4] Khattar D, Goud J S, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference. 2019: 2915-2921.

[5] Wang WY. " liar, liar pants on fire": A new benchmark dataset for fake news detection[J]. arXiv preprint arXiv:1705.00648, 2017.

[6] Qi P, Cao J, Yang T, et al. Exploiting Multi-domain Visual Information for Fake News Detection[J]. arXiv preprint arXiv:1908.04472, 2019.

[7] Shu K, Mahudeswaran D, Wang S, et al. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media[J]. arXiv preprint arXiv:1809.01286, 2018.

[8] Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE transactions on multimedia, 2016, 19(3): 598-608.

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

[10] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu. "Fake News Detection on Social Media: A Data Mining Perspective". SIGKDD 2017.

[11] Kai Shu, Suhang Wang, Huan Liu. "Beyond News Contents: The Role of Social Context for Fake News Detection". WSDM 2019.

[12] Shu, Kai, H. Russell Bernard, and Huan Liu. "Studying fake news via network analysis: detection and mitigation." Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. Springer, Cham, 2019. 43-65.

[13] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 855-864.

[14] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794.

[15] Džeroski S, Ženko B. Is combining classifiers with stacking better than selecting the best one?[J]. Machine learning, 2004, 54(3): 255-273.

[16] Shu K, Mahudeswaran D, Liu H. FakeNewsTracker: a tool for fake news collection, detection, and visualization[J]. Computational and Mathematical Organization Theory, 2019, 25(1): 60-71.

[17] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[19] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv preprint arXiv:1910.01108, 2019.

[20] Jiao X, Yin Y, Shang L, et al. Tinybert: Distilling bert for natural language understanding[J]. arXiv preprint arXiv:1909.10351, 2019.