Vietnamese Fake News Detection Based on Hybrid Transfer Learning Model and TF-IDF

Ngoc-Dong Pham, Thi-Hanh Le, Thanh-Dat Do, Thanh-Toan Vuong, Thi-Hong Vuong and Quang-Thuy Ha[†]
Vietnam National University, Hanoi (VNU),
VNU-University of Engineering and Technology (UET),
{17020687,18020457,20020045,18021279,hongvt57,thuyhq}@vnu.edu.vn

***Abstract*—There are a lot of studies about fake news detection on English social networks. However, Vietnamese fake news detection on social networks still limit. In this paper, we propose a new approach for Vietnamese Fake News Detection on Social Network Sites using a pre-train language model PhoBERT combine with Term Frequency - Inverse Document Frequency (TF-IDF) for word embedding and Convolutional Neural Network (CNN) for features extracting. Our proposed model is trained and evaluated on the dataset of Reliable Intelligence Identification on Vietnamese SNSs (ReINTEL) shared task. We process text data into two scenarios: raw data and processed data to elucidate the hypothesis of pre-processing data on social networks. In addition, we use the different extra features to improve the efficiency of model. We compare our proposed model with the baseline methods. The proposed model achieved outstanding results with 0.9538 AUC score on raw data.**

Keywords: Fake news Detection, PhoBERT, Convolutional Neural Network, TF-IDF.

I. Introduction

Fake news detection is researched in many domains, such as in scientific documents/articles, news, and social network sites (SNSs)... Fake news detection has an important role in checking the correctness of information as well as preventing the bad consequences that it causes. There are many studies on fake news detection by many groups of authors in the world [1], [11], [13], [16]. These studies mainly use traditional machine learning or transformer-based deep learning methods and have achieved high efficiency. We inherit the successes of those studies to develop a new approach for detecting fake news on Vietnamese SNS. Our approach is based on pre-trained language model PhoBERT [8] combined with TF-IDF for word representation and CNN [3] for feature extraction.

In the study [13], Rohit Kumar Kaliyar et al. combined pre-trained methods with CNN for fake news detection. In another study [11], Sunil Gundapu et al. proved the effectiveness of the model based on transformer for fake news detection. We inherit and promote two ideas above, further improve by using TF-IDF for word representation. TF-IDF helps to calculate how important a word is to a document in a collection or corpus, overcoming the limitations of the out-of vocab (OOV) problem in the pre-trained model. With powerful extraction capabilities, CNN's filters automatically learn feature vectors obtained from the pre-trained model and TF-IDF, thereby helping the model learn more effectively. We found that using extra features instead of just text features helps the model learn more efficiently. Besides, we realize that pre-processing data for news on SNSs needs to be considered because it can lose important features that distinguish true from false news. To demonstrate the superior performance of the proposed model, we compare it with three other methods: traditional machine learning with TF-IDF, deep learning with Bidirectional Long Short Term Memory (BiLSTM) [20], models based on transformer. Experimental results show that our proposed model achieved outstanding results with 0.9538 AUC score on the raw dataset from ReINTEL [10] shared task.

The rest of the paper is organized as follows, section II presents some prior works related to fake news detection. In section III, this paper describe the dataset and the selection of attributes to train the model. Methodology, and model architecture are described in detail in section IV. The experimental results and comparing the effectiveness of the methods will be presented in section V. Finally, in section VI we conclude our paper and discuss future model development.

II. Related Work

There are many different definitions of fake news in the existing literature as in [7], [15], [17]. In [4] showed that narrow definition of fake news is news articles that are intentionally and verifiably false and could mislead readers. In [17], there are two key features of this definition: authenticity and intent. First, fake news includes false information that can be verified as such. Second, fake news is created with dishonest intention to mislead consumers. In this article, we use the narrow definition of fake news as [17]. People are naturally not very good at differentiating between real and fake news. Several psychological and cognitive theories can explain this phenomenon and the influential power of fake news. Traditional fake news mainly targets consumers by exploiting their vulnerabilities. Considering the entire news consumption ecosystem, we can also describe some of the social dynamics that contribute to the proliferation of fake news. Prospect theory describes decision making as a process by which people make choices based on the relative gains and losses as compared to their current state [12].

[†]Corresponding author.

The fake news hype caused widespread disillusionment about social media, and many politicians, news publishers, IT companies, activists, and scientists concur that this is where to draw the line. In [15] identify that fake news detection contains three paradigms: fake news detection based on knowledge, context, and style. Existing fake news detection approaches utilize various linguistic features extracted from teaser messages, linked webpages, and tweet meta-information [5], [19]. Ahmed et al. [1] proposed a fake news detection model that uses n-gram analysis and machine learning techniques for the article domain. They investigated and compared two different features extraction techniques and six different machine classification techniques such as Stochastic Gradient Descent (SGD), Support Vector Machines (SVM), Linear Support Vector Machines (LSVM), K-Nearest Neighbour (KNN), and Decision Trees (DT). Experimental evaluation yields the best performance using TF-IDF as a feature extraction technique, and Linear Support Vector Machine (LSVM) as a classifier. Recent rapid technological advancements in online social networks such as Twitter and Facebook have led to a great incline in spreading false information and fake news. Gundapu et.al [11] proposed a methodology to analyze the reliability of information shared on social media about the COVID-19 pandemic. The best approach is based on an ensemble of three transformer models (BERT, ALBERT, and XLNET) to detecting fake news. This model was trained and evaluated in the context of the ConstraintAI 2021 shared task "COVID19 Fake News Detection in English" [2]. this study proposes the fake news detection models for news of social networks using advanced transformer models for the Vietnamese domain.

III. Data Description

The challenges in fake news detection are data collection and labeling. Association for Vietnamese Language and Speech Processing 2020 has shared a contest to assess the reliability of the information on Vietnamese social networks. The data includes 5172 records that are information posted on SNSs used by many Vietnamese people such as Facebook, Zalo, Lotus. Each record contains social media information and its corresponding labels, either real or fake. However, there are many error records in the data, so we handle it as the following:

- Remove null and empty records.
- For duplicate records, delete and keep only 1 record then label it 1 (positive).
- Convert numeric data to standard and unknown records to 0.

The dataset after processing remains 4837 records, of which there are 4018 records label 0 and 819 records label 1. We split the dataset into three subsets, 70% data for the training set, 10% data for the validation set (hyperparameter tuning), and 20% for the test set.

In fake news detection, the content of a post is the main attribute to determine whether the news is fake or real. We give some examples of common words in real and fake news in Figure 1. The data was collected in 2020, so the most popular news belongs to the Covid-19 epidemic topic. The words related to the epidemic that appear a lot in Figure 1.a (real news) also appear a lot in Figure 2 (fake news) is: "covid", "việt_nam", "cách_li", "virus". Besides, some words in real news such as: "bệnh_nhân", "tỉnh", "tiếp_xúc", vice versa, some words in fake news only such as: "corona", "đau", "cơ_thể". These frequent words can provide important information to distinguish between real and fake news.

Table I
Statistics of the datasets after processing.

	Dataset
Total news	4837
Users	3546
New have images	1258
Real news	4018
Fake news	819

Figure 1. Positive word cloud.

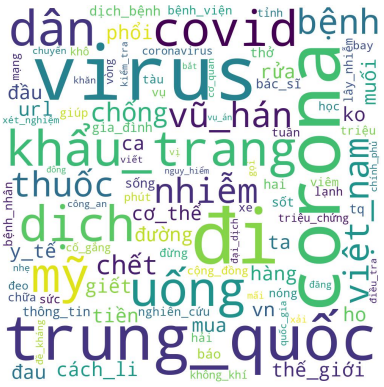

Figure 2. Negative word cloud.

From data mining, we found that in addition to text data, some other features also carry a lot of information, making an important contribution to the classification task. There is a significant difference in the number of likes, shares, and comments between real and fake news. The correlation between likes, shares, and comments in real news is quite even. On the contrary, there is a big difference between the number of shares with likes and comments in fake news. In

addition, fake news is usually posted between 10 a.m to 1 p.m and 6 p.m to 10 p.m - the period when users interact the most on social networks. While the posting time of real news is evenly distributed throughout the day. Usually, the posts are often attached with images, but the images in fake news posts are often low quality, much lower resolution than the real news. We used the extra features along with the text feature to train the model. Experimental results show that extra features help to improve the accuracy and efficiency of the model.

IV. Methodology

In this section, we describe our proposed model that combines pre-trained language model PhoBERT with TF-IDF arithmetic statistical technique and CNN. In addition, we compare our proposed model with the traditional machine learning, deep learning, and Transformer-based models with different word embeddings. Besides, we implement a system detecting fake news with extra social media features from our proposed model.

1) Data Preprocessing: For Natural Language Processing (NLP) tasks, preprocessing text data is essential, which helps to handle information noise but sometimes has the opposite effect of causing information loss. However, in the study [14] Fahim Mohammad et al. made a hypothesis that data preprocessing in social network problems is worth? Therefore, we divide the data into two experimental scenarios, which are processing data and raw data. However, words in the Vietnamese language are organized from multi-sounds, so that we use the VnCoreNLP[1] toolkit to word segments for both. The description of the preprocessing is shown in Figure 3. The preprocessing includes the following steps:

- **Convert emoji**: convert icons, emoticons, characters expressing emotions into text tokens.
  Example: :)) -> <mặt cười>
- **Convert URL**: convert the link, token <URL> to the same form.
  Example: https://www.facebook.com/ -> <url>
- **Convert the Covid-19**: virus name to the same form.
  Example: corona -> covid-19; COVID -> covid-19

2) Traditional Machine Learning Models: In this part, we experiment with traditional algorithms and ensemble learning techniques to solve the task.

- **Basic Machine learning**: With the traditional approach, we use TF-IDF for word representation to evaluate the importance of words in the text. We implement some models based on basic machine learning algorithms such as Passive Aggressive Classifier (PAC), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), Decision Tree Classification (DTC), Support Vector Machines (SVM). In which, SVM model gives the best results.

- **Ensemble Learning**: Ensemble methods are techniques that create multiple models and then combine them to obtain better predictive performance. The algorithms we use in this approach are Random Forest, XGBoost, LightGBM, and CatBoost, in which LightGBM gives the best results. With all algorithms, we also use TF-IDF for word representation tasks. In addition, we propose a model that combines SVM and LightGBM using the weighted ensemble method and achieves better performance.

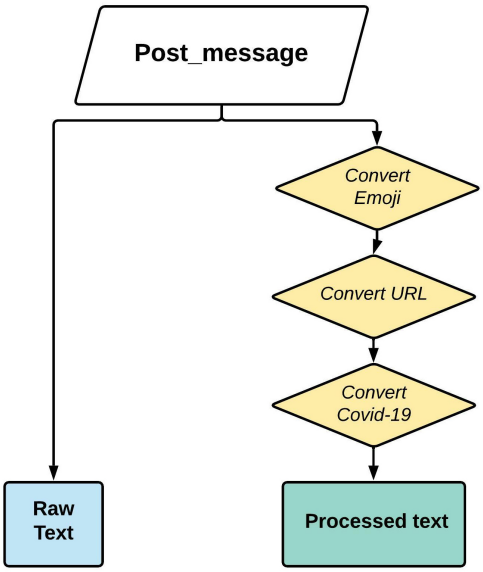

Figure 3. Text processing pipeline.

3) Deep Learning Models: With the deep learning models approach, we implement BiLSTM model then combine CNN with BiLSTM for better result.

- **BiLSTM**: Long Short Term Memory (LSTM) is a variant of the Recurrent Neural Network (RNN), it can prevent the vanishing gradient problem. In the approach, we use BiLSTM network-sequence processing model that consists of two direction LSTMs to solve the task. For word embedding, we use pre-trained Fasttext[2] (Facebook) and pre-trained GloVe[3] (Stanford) for Vietnamese by our self-training. The model using pre-trained GloVe achieves better results than using pre-trained Fasttext.

- **BiLSTM+CNN**: The idea of combining two deep learning models BiLSTM and CNN to calculate the dependence of words on words near it and retain the information that appears before in the memory. In this part, we also use pre-trained Fasttext and GloVe for word embedding. The combined deep learning model achieves better results than the conventional BiLSTM model with GloVe.

[1]https://github.com/vncorenlp/VnCoreNLP

[2]https://fasttext.cc/

[3]https://nlp.stanford.edu/projects/glove/

4) Transformer Models: In recent years, the Transformer-based model has become the state-of-the-art of many NLP tasks. Unlike RNN, the Transformer-based model allows parallel computation so it does not depend on the previous words. Besides, a self-attention mechanism helps calculate the relationship between words. In this section, we implement two transformer-based models for Vietnamese, which are PhoBERT and ViElectra.

- **Fine-tuning Pre-trained PhoBERT**: BERT (Bidirectional Encoder Representation from Transformers) language model [9] proposed by J. Devlin et al. is a two-dimensional word representation model based on Transformers technique. BERT is designed for pre-trained word embedding, unlike other word embedding techniques, it can balance contextualization in both left and right dimensions. Nowadays, BERT and its variants are very commonly used for many natural language processing tasks. PhoBERT (VinAI Research) is the most popular monolingual BERT pre-trained model in Vietnamese. From the advantages of the BERT language model, we perform a simple fine-tuning of PhoBERT by adding a classifier at the end to bring the vector to the same dimension as the output values.

- **Fine-tuning Pre-trained ViELECTRA**: The advancement of pre-trained models has made significant breakthroughs in solving NLP problems. In March 2020, Google introduced a pre-trained method called Electra [6] with an approach that uses the advantages of BERT but learns newer techniques and more computational resources. The pre-trained Electra for Vietnamese language ViElectra [18] was also just announced in October 2020 by the team of FPT AI. Our team leveraged and fine-tuned the ViElectra pre-train model to solve this problem.

5) Combination Models: Since its inception, Transformer and its variants have affirmed their position in NLP tasks. In the research [11], Sunil Gundapu et al. showed that the representation with two-dimensional context of transformer-based model is necessary for the fake news detection problem because fake news tends to deceive the reader. So we built the backbone-based model from the fine-tuning pre-trained language model Phobert model and incorporated some more advanced techniques. After 12 layers of encoder in PhoBERT base architecture, we implement CNN layers with different kernel sizes to feature extraction of context representation vectors. In particular, instead of using the output of the last 12th encoder layer, we took the average output of the three last encoder layers. This change slightly improves the efficiency of the model and avoids overfitting at the last layer. After each CNN layers are Maxpoling layers reduce the dimensionality. Then we combined them by concatenation to pass through the classification layer. In the classification layer, we also alternately stack the Linear and Drop-out layers to bring the vector dimension to the same number of dimensions as the output. This idea was proposed by Rohit Kumar in research [13], the author also said that a CNN network with more depth also makes the model avoid overfitting and is useful for the fake news detection task.

However, the disadvantage of pre-trained models is the problem of OOV (words that do not appear in the vocab will become unknown token), especially with the problem of social networks when the information content is abbreviations, teencode, misspellings,... For that reason, we have proposed incorporating TF-IDF at the embedding layer so that when traversing the network layers the unknown token information can be kept. In addition, it also takes advantage of the strength of TF-IDF, which can calculate important words in the text. We combined TF-IDF by concatenating with PhoBERT at embedding Layer, however, to facilitate the calculation we used SVD to cut the token length to fixed. Details of the model structure are shown in Figure 4.

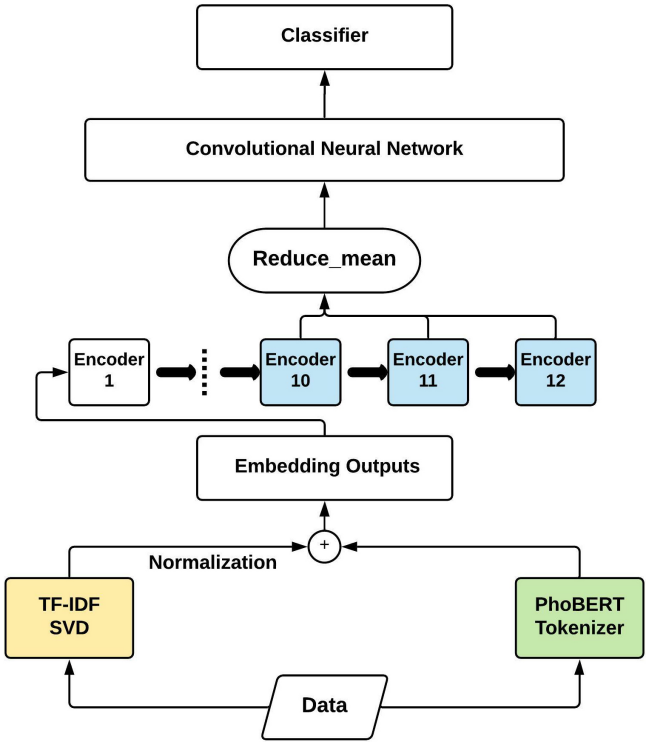

Figure 4. Our propesed model architecture.

6) Multi-Input Models: With the above NLP model approaches, we also incorporate additional information-carrying attributes as presented in section III. Besides the post_mesage feature, we use 13 extra features including: num_char, num_url, num_hashtag, num_post, num_like, num_cmt, num_share, pixel, num_image, hour, weekday, day, month. These additional features bring significant efficiency to the classification model. We have presented the user penalty point, but in the small data experiment, it caused overfitting so we removed it. Details of the results are presented in the next section.

V. Experiments

A. Evaluation metrics

To evaluate the performance of algorithms for fake news detection problem [17], we have some concepts:

- True Positive (TP): When fake news is correctly predicted to be fake news.
- True Negative (TN): When real news is correctly predicted to be real news.
- False Negative (FN): When fake news is wrongly predicted to be real news.
- False Positive (FP): When real news is wrongly predicted to be fake news.

Because fake news datasets are skewed [10], in this study we used the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) to evaluate the performance of the model. The Receiver Operating Characteristics (ROC) curve compares the performance of classifiers by varying the class distribution through a certain threshold based on True Positive Rate (TPR) and False Positive Rate (FPR). The ROC curve is defined by plotting FPR on the x-axis and TPR along the y-axis. TPR and FPR are defined as follows:

$$TPR = \frac{|TP|}{|TP| + |FN|}$$

$$FPR = \frac{|FP|}{|FP| + |TN|}$$

The performance of the classification model is calculated by the Area Under the Curve (AUC). The AUC score is between 0 and 1. The higher AUC score, the better the performance of the model at distinguishing between the positive and negative classes. AUC is defined as below:

$$AUC = \frac{\sum(n_0 + n_1 + 1 - r_i) - n_0(n_0 + 1)/2}{n_0 n_1}$$

where r_i is the rank of ith fake news piece and n_0 (n_1) is the number of fake (true) news pieces.

B. Results and Discussion

In the first experiment, we trained the model on text data under two scenarios: processed text and raw text. Table II shows that all methods give better results with raw text. This is clear evidence supporting the hypothesis that we should consider processing text data on social network problems we mentioned in section III.

Table II
AUC scores of text data processing methods.

Model	Processed	Raw
TF-IDF + LightGBM	0.8827	0.8880
TF-IDF + SVM	0.9086	0.9111
TF-IDF + weighted SVM + LightGBM	0.9110	0.9129
GloVe + BiLSTM	0.8556	0.8655
GloVe + BiLSTM + CNN	0.8755	0.8779
ViElectra	0.8970	0.9086
PhoBERT	0.9278	0.9348
PhoBERT + TF-IDF + CNN	**0.9399**	**0.9476**

Table II showed that model PhoBert+TF-IDF+CNN achieved the highest results with 0.9476 AUC score. From there, it can be seen that combining the pre-trained model with other methods brings higher efficiency instead of just fine-tuning the model normally. Transformer-based models achieve good results that outperform other methods. However, due to limited computational resources, the ViElectra model is not efficient. The traditional approach using TF-IDF and SVM still give quite high results on both data scenarios. The weighted ensemble SVM+LightGBM model obtained better results than SVM and LightGBM models. It proves that training models with traditional methods are always effective, especially when using ensemble learning.

Table III
Efficiency of combining CNN and TF-IDF.

Model	AUC score
PhoBERT + TF-IDF	0.9383
PhoBERT + CNN	0.9458
PhoBERT + TF-IDF + CNN	**0.9476**

Table III shows the efficiency of using TF-IDF and CNN when combined with the pre-trained PhoBERT. The use of TF-IDF helps overcome the limitations of the OOV problem of pre-trained models, helps model to learn rare words thereby improving the performance of the model. With powerful extraction capabilities, CNN filters will extract important features that are useful for classification.

Table IV
Result of adding more features.

Model	AUC score
SVM	0.9343
LightGBM	0.9449
Weighted ensemble SVM + LightGBM	0.9532
PhoBERT + TF-IDF + CNN	**0.9538**

Through data analysis, we found that there is a high correlation between numerical features and labels. So we decided to train the model on more features. The results in table IV show that traditional machine learning models had significantly better performance. When training model on text data, SVM achieves higher results than LightGBM, but when combining features, LightGBM achieves significantly higher results. Especially, the weighted ensemble SVM+LightGBM model obtained outstanding results, increasing 5% AUC score compared to using only text features. This proves that the ensemble model is extremely efficient when trained on a dataset combining features. However, PhoBERT+TF-IDF+CNN model has still obtained the best results with 0.9538 AUC score.

VI. Conclusions

In this paper, we have experimented with traditional machine learning and deep learning methods as well as combined the two methods, finally proposing a superior model to solve the information detection problem. Our proposed model inherits and develops ideas from two papers by Rohit Kumar et al.

and Sunil Gundapu et al. We further improved the use of TF-IDF for word representation to solve the OOV problem in pre-trained models. The models were evaluated on the dataset from the ReINTEL shared task. Experimental results show that our proposed model has the highest results with 0.9538 AUC score on the test set. Through experimentation, we realized that data preprocessing for social media posts need to be considered because it may lose important information. Besides, in addition to text data, other features also contain a lot of important information to make the model classify more accurately.

VII. Acknowledgment

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup, and supported by Vingroup Innovation Foundation (VINIF) under project code DA137_15062019.

References

[1] Ahmed, H., Traore, I., Saad, S.: Detection of online fake news using n-gram analysis and machine learning techniques. In: International conference on intelligent, secure, and dependable systems in distributed and cloud environments. pp. 127–138. Springer (2017)

[2] Akhtar, M.S., Chakraborty, T.: Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In: Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT. p. 42 (2021)

[3] Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). pp. 1–6 (2017). https://doi.org/10.1109/ICEngTechnol.2017.8308186

[4] Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of economic perspectives **31**(2), 211–36 (2017)

[5] Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N.: Stop clickbait: Detecting and preventing clickbaits in online news media. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). pp. 9–16. IEEE (2016)

[6] Clark, K., Luong, M.T., V. Le, Q., D. Manning, C.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)

[7] Conroy, N.K., 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)

[8] Dat Quoc, N., Anh Tuan, N.: Phobert: Pre-trained language models for vietnamese. arXiv preprint arXiv:2003.00744 (2020)

[9] 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)

[10] Duc-Trong, L., Xuan-Son, V., Nhu-Dung, T., Huu-Quang, N., Thuy-Trinh, N., Linh, L., Anh-Tuan, N., Minh-Duc, H., Nghia, L., Huyen, N., Hoang, D.N.: Reintel: A multimodal data challenge for responsible information identification on social network sites. arXiv preprint arXiv:2012.08895 (2020)

[11] Gundapu, S., Mamidi, R.: Transformer based automatic covid-19 fake news detection system. arXiv preprint arXiv:2101.00180 (2021)

[12] Kahneman, D.: Prospect theory: An analysis of decisions under risk. Econometrica **47**, 278 (1979)

[13] Kaliyar, R.K., Goswami, A., Narang, P., Sinha, S.: Fndnet – a deep convolutional neural network for fake news detection. Cognitive Systems Research **61**, 32–44 (2020)

[14] Mohammad, F.: Is preprocessing of text really worth your time for online comment classification? arXiv preprint arXiv:1806.02908 (2018)

[15] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 (2017)

[16] Ruchansky, N., Seo, S., Liu, Y.: Csi: A hybrid deep model for fake news detection. arXiv preprint arXiv:1703.06959 (2017)

[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] Viet, B.T., Oanh, T.T., Phuong, L.H.: Improving sequence tagging for vietnamese text using transformer-based neural models. arXiv preprint arXiv:2006.15994 (2020)

[19] Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. In: Proceedings of the ACL 2014 workshop on language technologies and computational social science. pp. 18–22 (2014)

[20] Zhang, S., Zheng, D., Hu, X., Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. pp. 73–78. Shanghai, China (Oct 2015), https://aclanthology.org/Y15-1009