ReINTEL Challenge 2020: Exploiting Transfer Learning Models for Reliable Intelligence Identification on Vietnamese Social Network Sites

Kim Thi-Thanh Nguyen[1,2], **Kiet Van Nguyen**[1,2]

[1]University of Information Technology, Ho Chi Minh City, Vietnam

[2]Vietnam National University, Ho Chi Minh City, Vietnam

18520963@gm.uit.edu.vn, kietnv@uit.edu.vn

Abstract

This paper presents the system that we propose for the Reliable Intelligence Identification on Vietnamese Social Network Sites (ReINTEL) task of the Vietnamese Language and Speech Processing 2020 (VLSP 2020) Shared Task. In this task, the VLSP 2020 provides a dataset with approximately 6,000 training news/posts annotated with reliable or unreliable labels, and a test set consists of 2,000 examples without labels. In this paper, we conduct experiments on different transfer learning models, which are bert4news and PhoBERT fine-tuned to predict whether the news is reliable or not. In our experiments, we achieve the AUC score of 94.52% on the private test set from ReIN-TEL's organizers.

Index Terms: Reliable, news, transfer learning, bert4news, PhoBERT.

1 Introduction

With the explosion of The Fourth Industrial Revolution in Vietnam, SNSs such as Facebook, Zalo, Lotus have attracted a huge number of users. SNSs have become an essential means for users to not only connect with friends but also freely share information and news. In the context of the COVID-19 pandemic, as well as prominent political and economic events that are of great interest to many people, some people tend to distribute unreliable information for personal purposes. The discovery of unreliable news has received considerable attention in recent times. Therefore, VLSP opens ReIN-TEL (Le et al., 2020) shared-task with the purpose of identifying being shared unreliable information on Vietnamese SNSs.

Censoring news to see if it is trustworthy is tedious and frustrating. It is sometimes difficult to determine whether the news is credible or not. Fake news discovery has been studied more and more by academic researchers as well as social networking companies such as Facebook and Twitter. Many shared-task to detect rumors were held, such as SemEval-2017 Task 8: Determining rumour veracity and support for rumours (Derczynski et al., 2017) and SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours (Gorrell et al., 2019).

In this task, we focus on finding a solution to categorize unreliable news collected in Vietnamese, which is a low-resource language for natural language preprocessing. Specifically, we implement deep learning and transfer learning methods to classify SNSs news/posts. The problem is stated as:

- **Input**: Given a Vietnamese news/post on SNSs with the text of news/post (always available), some relative information, and image (may be missing).

- **Output**: One of two labels (unreliable or reliable) that are predicted by our system.

Figure 1 shows an example of this task.

The rest of the paper is organized as follows. In Section 2, we present the related work. In Section 3, we explain some proposed approaches and its result. In Section 4, we present the experimental analysis. Finally, Section 5 draws conclusions and future work.

2 Related work

Ruchansky et al. (2017) used a hybrid model called CSI to categorize real and fake news. The CSI model includes three components: capture, source, and integrate. The first module is used to detect a user's pattern of activity on news feeds. The second module learns the source characteristics of user behavior. The last module combines both previous modules to categorize news is real or fake. The CSI model does not make assumptions about user behavior or posts, although it uses both user-profiles and article data for classification.

Id: 0.
User id: 2167074723833130000.
Post message: Cần các bậc phụ huynh xã Ngũ Thái lên tiếng, không ngờ xã mình cũng nhận thịt nhiễm sán... Cho các cháu Mầm non ăn uống thế này thật vô nhân tính! VTV đăng tin rồi nhé các anh chị.
English translation: *Needing the parents of Ngu Thai commune to speak up, astonishing my commune accept contaminated meat ... Feeding preschool children like this is so inhumane! VTV posted the news, guys.*
Timestamp post: 1584426000.
Number of post's like: 45.
Number of post's comment: 15.
Number of post's share: 8.
Label: 1 (unreliable).
Image: NAN.

Figure 1: An example extracted from the dataset.

Slovikovskaya (2019) focused on improving the results of the Fake News Challenge Stage 1 (FNC-1) stance detection task using transfer learning. Specifically, this work improved the FNC-1 best performing model adding BERT (Devlin et al., 2018) sentence embedding of input sequences as a model feature and fine-tuned XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019b) transformers on FNC-1 extended dataset.

3 Approaches

In this study, we concentrate on SOTA models, including deep neural network models and transfer learning models.

3.1 Experimental approaches

3.1.1 Deep neural network models

In studying the fundamental theories and methods of detecting fake news, Zhou and Zafarani (2020) have come up with some fundamental theories of detecting fake news. The authors wrote, *"theories have implied that fake news potentially differs from the truth in terms of, e.g., writing style and quality (by Undeutsch hypothesis)"*. Therefore, we choose text-feature as the primary input of our experimental models. Firstly, we run deep learning models like Text CNN (Kim, 2014), BiLSTM (Zhou et al., 2016) combine with some pre-trained word embed-

Model	AUC
Text CNN + FastText	0.865996
Text CNN + PhoW2V	0.846567
BiLSTM + FastText	0.863183
BiLSTM + PhoW2V	0.854487

Table 1: Experimental results of deep neural models on public test set.

Feature name	Train set	Public test set	Private test set
Id	0	0	0
User name	0	0	0
Post message	1	0	0
Timestamp post	96	28	34
Number of like	115	41	616
Number of comment	10	7	677
Number of share	725	280	742
Label	0	0	0
Image	3,085	1,148	1,138

Table 2: Statistics of missing values in the dataset.

ding models such as FastText[1] (Bojanowski et al., 2016) and PhoW2V[2] (Tuan Nguyen et al., 2020) to predict the credibility of news. The results of this approach get an AUC score of 0.84 to 0.86, as shown in Table 1. We also plan to experiment with incorporating other features that ReINTEL's organizers provide, such as user id, the number of likes, shares, comments, and image, but the lack of information (shown in Table 2) leads to enormous dynamic causes us to ignore this approach.

3.1.2 BERT and RoBERTa for Vietnamese

One of the problems of deep learning is its massive data requirements as well as the need for computing resources. This has spurred the development of large models and transfer learning methods. Nguyen et al. (2020) presents two BERT fine-tuning methods for the sentiment analysis task on datasets of Vietnamese reviews and gets slightly outperforms other models using GloVe and Fast-Text. Liu et al. (2019a) fine-tuned BERT under the multi-task learning framework and obtains new state-of-the-art results on ten NLU tasks, including SNLI, SciTail, and eight out of nine GLUE tasks, pushing the GLUE benchmark to 82.7% (an improvement of 2.2%)[3]. Therefore, we attempt to fine-tune PhoBERT[4] (Nguyen and Tuan Nguyen, 2020) and bert4news[5], pre-trained models for Vietnamese which is based on BERT architecture. And transfer learning shows strength in these experi-

---

[1]https://fasttext.cc/docs/en/crawl-vectors.html
[2]https://github.com/datquocnguyen/PhoW2V
[3]As of February 25, 2019 on the latest GLUE test set
[4]https://github.com/VinAIResearch/PhoBERT
[5]https://github.com/bino282/bert4news

Model	AUC
PhoBERT	0.932424
bert4news	0.935163
PhoBERT+bert4news	0.945169

Table 3: Results of PhoBERT, bert4news, and results that combine these two models on the private test set.

Model	Epochs	Random Seed	Learning Rate	AUC
PhoBERT	5	42	1.00e-5	0.901628
PhoBERT	6	42	3.00e-5	0.920835
PhoBERT	7	38	2.00e-5	0.924961
PhoBERT	7	42	2.00e-5	0.932424
bert4news	5	42	3.00e-5	0.930596
bert4news	6	24	2.00e-5	0.922787

Table 4: Parameter changes lead to a change of results on the public test set.

ments, we get an AUC score of between 0.92 to almost 0.95, as shown in Table 3.

3.2 Fine-tuning BERT and RoBERTa for Vietnamese

After many experiments, we find that the deep learning models do not achieve higher performance than fine-tuned bert4news and PhoBERT. Therefore, we decided to focus on only improving the results on transfer learning methods. Besides, we also try to combine the results of these two models.

The fine-tuning idea is taken from the study (Sun et al., 2019). The BERT base model creates an architecture of 12 sub-layers in the encoder, 12 heads in multi-head attention on each sub-layer. BERT input is a sequence of not more than 512 tokens; the output is a set of self-attention vectors equal to the input length. Each vector is 768 in size. The BERT input string represents both single text and text pairs explicitly, where a special token [CLS] is used for string sorting tasks, and a special token [SEP] marks the end position of the single text or the position that separates the text pair. For fine-tuning the BERT architecture for text classification, we concatenated the last four hidden representations of the [CLS] token, which will be passed into a small MLP network containing the full connection layers to transform into the distribution of discrete label values.

Our fine-tuning process consists of two main steps: tokenize the text content and retrain the model on the dataset. For PhoBERT, we use VN-coreNLP (Vu et al., 2018) library to tokenize content, while for bert4news, we use BertTokenizer.

4 Experiments

4.1 Experimental settings

In this paper, we conduct various experiments on Google Colab (CPU: Intel(R) Xeon(R) CPU @ 2.20GHz; RAM: 12.75 GB; GPU Tesla P100 or T4 16GB with CUDA 10.1). We fine-tune PhoBERT and bert4news with different parameters as batch size, learning rate, epoch, random seed. To save
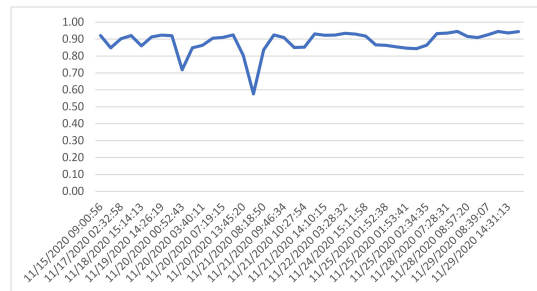

Figure 2: Performances of the team during the challenging task.

time and cost, we set batch size 32 for all models. With the same hyperparameter values, distinct random seeds can lead to substantially different results (Dodge et al., 2020). With the above configuration, we spend about 2.40 minutes per epoch for both bert4news and PhoBERT. Table 4 shows the parameter setting and the performance, respectively.

4.2 Performances over time

Figure 2 shows the results of our testing process. It is easy to see that our results are not stable in the first phase due to trying many methods. Our results are more stable in the later stage of the competition, but there are not many mutations.

5 Conclusion and future work

In summary, we have proposed the following methods for classifying untrustworthy news: combining deep learning model with pre-trained word embedding, fine-tune bert4news, and PhoBERT, combining text, numeric, and visual features. Accordingly, the best result belongs to the transfer learning models when achieving an AUC score of 94.52% for the combined model of bert4news and PhoBERT.

In the future, we plan to combine other features offered by ReINTEL's organizers with transfer learning models due to classifying based on news content alone is not enough (Shu et al., 2019). While we are doing well in transfer learning, we also aim to build a system for the fast and accu-

rate detection of fake news at the early stages of propagation, which is much more complicated than detecting long-circulated news. Besides, we hope to develop a system to score users based on the news they post and share to reduce unreliable news on Vietnam SNSs.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Duc-Trong Le, Xuan-Son Vu, Nhu-Dung To, Huu-Quang Nguyen, Thuy-Trinh Nguyen, Linh Le, Anh-Tuan Nguyen, Minh-Duc Hoang, Nghia Le, Huyen Nguyen, and Hoang D. Nguyen. 2020. Reintel: A multimodal data challenge for responsible information identification on social network sites.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Quoc Thai Nguyen, Thoai Linh Nguyen, Ngoc Hoang Luong, and Quoc Hung Ngo. 2020. Fine-tuning bert for sentiment analysis of vietnamese reviews.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news. *CoRR*, abs/1703.06959.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Valeriya Slovikovskaya. 2019. Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. *CoRR*, abs/1910.14353.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *CoRR*, abs/1611.06639.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).