ReINTEL Challenge 2020: A Multimodal Ensemble Model for Detecting Unreliable Information on Vietnamese SNS

Nguyen Manh Duc Tuan
Toyo Unversity, Japan
ductuan024@gmail.com

Pham Quang Nhat Minh
Aimesoft JSC, Vietnam
minhpham@aimesoft.com

Abstract

In this paper, we present our methods for unrealiable information identification task at ReINTEL Challenge 2020. The task is to classify a piece of information into reliable or unreliable category. We propose a novel multimodal ensemble model which combines two multimodal models to solve the task. In each multimodal model, we combined feature representations acquired from three different data types: texts, images, and metadata. Multimodal features are derived from three neural networks and fused for classification. Experimental results showed that our proposed ensemble model improved against single models in term of AUC score. We obtained 0.9445 AUC score on the private test of the challenge.

1 Introduction

Recently, fake news detection have received much attention in both NLP and data mining research community. This year, for the first time, VLSP 2020 Evaluation Campaign held ReINEL Challenge (Le et al., 2020) to encourage the development of algorithms and systems for detecting unreliable information on Vietnamese SNS. In ReINTEL Challenge 2020, we need to determine whether a piece of information containing texts, images, and metadata is reliable or unreliable. The task is formalized as a binary classification problem and training data with annotated labels was provided by VLSP 2020 organizers.

In this paper, we present a novel multimodal ensemble model for identifying unreiable information on Vietnamese SNS. We use neural networks to obtain feature representations from different data types. Multimodal features are fused and put into a sigmoid layer for classification. Specifically, we use BERT model to obtain feature representations from texts, a multi-layer perceptron to encode metadata and text-based features, and a fine-tuned VGG-19 network to obtain feature representations from images. We combined two single models in order to improve the accuracy of fake news detection. Our proposed model obtained 0.9445 ROC AUC score on the private test of the challenge.

2 Related Work

Approaches to fake news detection can be roughly categorized into categorises: content-based methods, user-based methods and propagation-based methods.

In content-based methods, content-based features are extracted from textual aspects, such as from the contents of the posts or comments, and from visual aspects. Textual features can be automatically extracted by a deep neural network such as CNN (Kaliyar et al., 2020; Tian et al., 2020). We can manually design textual features from word clues, patterns, or other linguistic features of texts such as their writing styles (Ghosh and Shah, 2018; Wang et al., 2018; Yang et al., 2018). We can also analyze unreliable news based on the sentiment analysis (Wang et al., 2018). Furthermore, both textual and visual information can be used together to determine fake news by creating a multimodal model (Zhou et al., 2020; Khattar et al., 2019; Yang et al., 2018).

We can detect fake news by analysing social network information including user-based features and network-based features. User-based features are extracted from user profiles (Shu et al., 2019; Krishnan and Chen, 2018; Duan et al., 2020). For example, number of followers, number of friends, and registration ages are useful features to determine the credibility of a user post (Castillo et al., 2011). Network-based features can be extracted from the propagation of posts or tweets on graphs (Zhou and Zafarani, 2019; Ma et al., 2018).

3 Methodology

In this section, we describe methods which we have tried to generate results on the private test dataset of the challenge. We have tried three models in total and finally selected two best models for ensemble learning.

3.1 Preprocessing

In the pre-processing steps, we perform following steps before putting data into models.

- We found that there are some emojis written in text format such as ":)", ";)", "=]]", ":(", "=[", etc. We converted those emojis into sentiment words "happy" and "sad" in Vietnamese respectively.

- We converted words and tokens that have been lengthened into short form. For example, "Coooool" into "Cool" or "*****" into "**".

- Since many posts are related to COVID-19 information, we changed different terms about COVID-19 into one term, such as "covid", "ncov" and "convid" into "covid", for consistency.

Since meta-data of news contains a lot of missing values, we performed imputation on four original metadata features. We used the mean values to fill missing values for three features including the number of likes, the number of shares, and the number of comments. For the timestamp features, we applied the MICE imputation method (Azur et al., 2011).

We found that there are some words written in incorrect forms, such as 's.áthại' instead of 'sát hại'. One may try to convert those words into standard forms, but as we will discuss in Section 4, keeping the incorrect form words actually improved the accuracy of models.

We converted the timestamp feature into 5 new features: day, month, year, hour and weekday. In addition to metadata features provided in the data, we extracted some statistic information from texts: number of hashtags, number of urls, number of characters, number of words, number of question-marks and number of exclaim-marks. For each user, we counted the number of unreliable news and the number of reliable news that the user have made and the ratio between two numbers, to indicate the sharing behavior (Shu et al., 2019). We also created

a Boolean variable to indicate that a post contains images or not. In total, we got 17 features including metadata features. All the metadata-based features will be standardized by subtracting the mean and scaling to unit variance, except for the Boolean feature.

3.2 Model Architecture

Figure 1 shows the general model architecture of three models we have tried. In all models, we applied the same strategy for image-based features and meta-data based features. For metadata-based features, we passed it into a fully-connected layer layer with batch normalization. We found that there are posts having one or more images and there are posts having have no image. For posts containing images, we randomly chose one image as the input. For other posts, we created a black image (all pixels have zero values) as the input. We then fine-tuned VGG-19 model on the images of the training data. After that, we used the output prior the fully-connected layer as image-based features. Instead of taking averages of all vectors of pixels, we applied the attention mechanism as shown in Figure 1b to obtain the final representation of images.

In the following sections, we describe three variants that we made from the general architecture.

Model 1

In the first model (Figure 2a), we obtained the embedding vector of a text using BERT model (Devlin et al., 2019). After that, we used 1D-CNN (Kim, 2014) with filter sizes 2, 3, 4, and 5. By doing that, we can use more information from different sets of word vectors for prediction. We flattened and concatenated all the output from 1D-CNN and passed into a fully-connected layer with with a batch normalization layer. Finally, we took averages of features of texts, images and metadata and passed them into a sigmoid layer for classification.

Model 2

In the second model (Figure 2b), there are some changes in comparison with the first model. After passing the embedding vectors through various layers of 1D-CNN, we stacked those outputs vertically and passed into three additional 1D-CNN layers.

Model 3

In the third model (Figure 2b), we just slightly changed the second model by adding a shortcut connections between input and the output of each 1D-CNN layer.
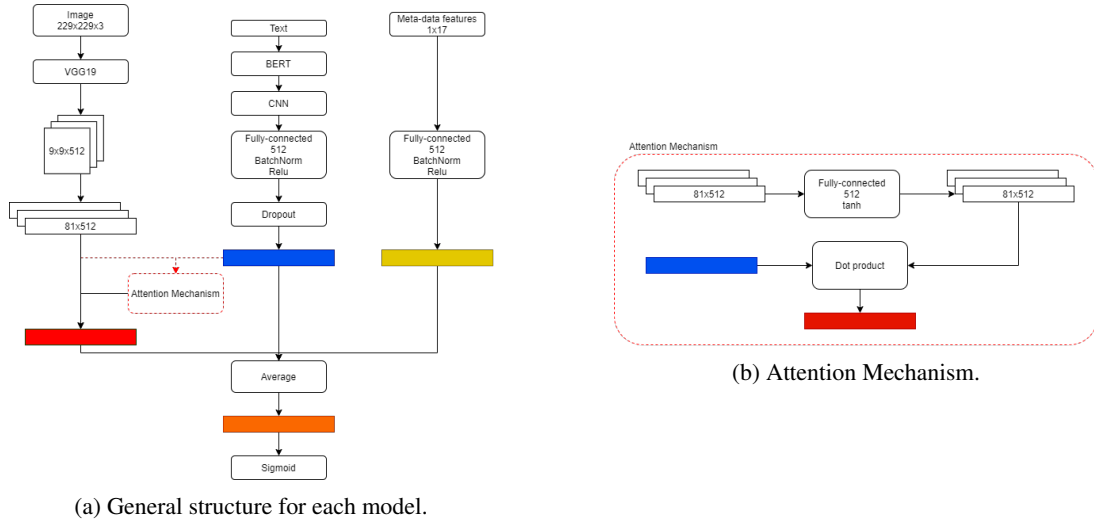
(a) General structure for each model.

(b) Attention Mechanism.

Figure 1: General Model Architecture

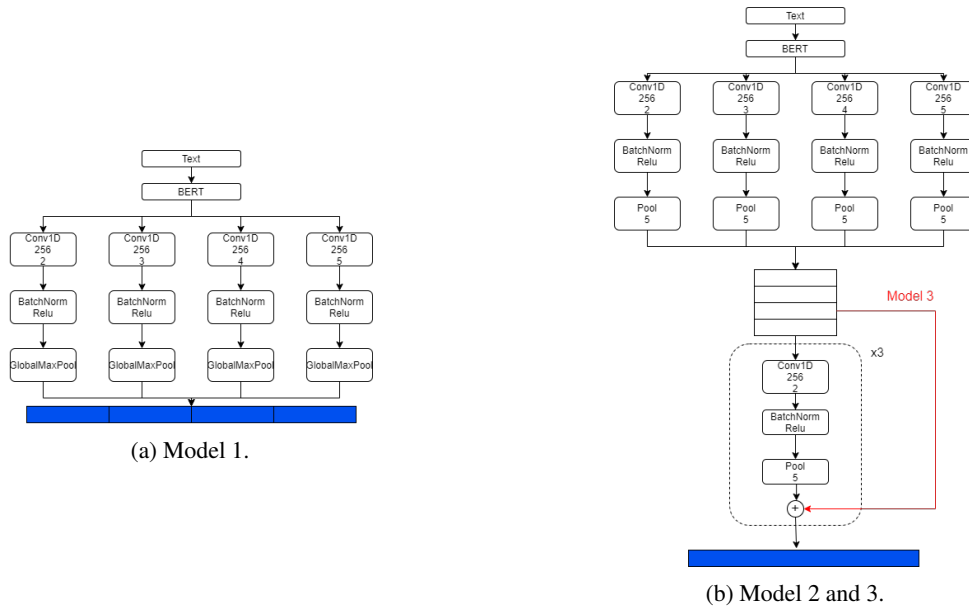(a) Model 1.

(b) Model 2 and 3.

Figure 2: Text-based features extractor for each model.

Ensemble Model

For the final model, we selected two best models among three above models and took averages of probabilities returned by the two models to obtain the final result.

4 Experiments and Results

In experiments, we used the same parameters as showed in Table 1 for all proposed models[1]. We reported ROC-AUC scores on the private test data.

In the first experiment, we compared two ways of preprocessing texts: 1) converting words in incorrect forms into corrected forms; and 2) keeping the

incorrect forms of words. The text is put through PhoBERT (Nguyen and Nguyen, 2020) to get the embedded vectors. In this experiment, we did not apply the attention mechanism. Table 2 shows that keeping the original words obtained better ROC-AUC score.

Next, we compared the effects of two different pre-trained BERT models for Vietnamese: PhoBERT and Bert4news[2]. Table 3 shows that Bert4news model is significantly better than PhoBERT model. Furthermore, when we added the proposed attention mechanism to get feature representations for images, we obtained 0.940217

---

[1]Our code: https://github.com/dt024/vlsp2020_toyoaimesoft

[2]Bert4News is available on: https://github.com/bino282/bert4news

Hyper-parameter	Value
FC layers	512
Dropout	0.2
Pooling size	5
1D-Conv filters	256
Learning parameter	2e-5
Batch size	16

Table 1: Parameters Setting

Exp	ROC-AUC
Convert words to correct forms	0.918298
Keep words in incorrect forms	0.920608

Table 2: Two ways of preprocessing texts.

Exp	ROC-AUC
PhoBERT	0.920608
Bert4news	0.927694
Bert4news + attention	0.940217

Table 3: Comparison of different pre-trained models and using attention mechanism

Exp	ROC-AUC
Model 1	0.939215
Model 2	0.919242
Model 3	0.940217
Ensemble	0.944949

Table 4: Final results

AUC score.

Table 4 shows results for three models which we have described in section 3. We got 0.939215 with model 1, 0.919242 with model 2, and 0.940217 with model 3. The final model is derived from model 1 and model 3 by calculating the average of results returned by model 1 and model 3. We obtained 0.944949 of ROC-AUC using that simple ensemble model.

5 Discussion

Since there may be more than one images in a post, we have tried to use one image as input or multiple images (4 images at most) as input. In preliminary experiments, we found that using only one image for each post obtained higher result in development set, so we decided to use one images in further experiments.

We have showed that keeping words in incorrect forms in the text better than fixing it to the correct forms. A possible explanation might be that those texts may contain violent contents or extreme words and users use that forms in order to bypass the social media sites' filtering function. Since those words can partly reflect the sentiment of the text, the classifier may gain benefit from it. The reason is that unreliable contents tend to use more subjective or extreme words to convey a particular perspective (Wang et al., 2018).

We also showed that by using the proposed attention mechanism, the result improved significantly. This result indicates that images and texts are co-related. In our observation, images and texts of reliable news are often related while in many unreliable news, posters use images that do not relate to the content of the news for click-bait purpose.

We found that convolution layers are useful and textual features can be well extracted by CNN layers. Conneau et al., 2017 has showed that a deep stack of local operations can help the model to learn the high-level hierarchical representation of a sentence and increasing the depth leads to the improvement in performance. Also, deeper CNN with residual connections can help to avoid overfitting and solves the vanishing gradient problem (Kaliyar et al., 2020).

6 Conclusion

6.1 Summary

We have presented a multimodal ensemble model for unreliable information identification on Vietnamese SNS. We combined two neural network models which fuse multimodal features from three data types including texts, images, and metadata. Experimental results confirmed the effectiveness of our methods in the task.

6.2 Future work

As future work, we plan to use auxiliary data to verify if a piece of information is unreliable or not. We believe that the natural way to make a judgement in fake news detection task is to compare a piece of information with different information sources to find out relevant evidences of fake news. We also want to see whether or not choosing one image randomly can affects the results and find solution to use more than one image.

References

Melissa Azur, Elizabeth Stuart, Constantine Frangakis, and Philip Leaf. 2011. Multiple imputation by chained equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20:40–9.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Xinhuan Duan, Elham Naghizade, Damiano Spina, and Xiuzhen Zhang. 2020. RMIT at PAN-CLEF 2020: Profiling Fake News Spreaders on Twitter. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR Workshop Proceedings.

Souvick Ghosh and Chirag Shah. 2018. Towards automatic fake news classification. *Proceedings of the Association for Information Science and Technology*, 55(1):805–807.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. Fndnet – a deep convolutional neural network for fake news detection. *Cogn. Syst. Res.*, 61(C):32–44.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA. Association for Computing Machinery.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

S. Krishnan and M. Chen. 2018. Identifying tweets with fake news. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 460–464.

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.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

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.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The role of user profile for fake news detection.

Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. 2020. Early detection of rumours on twitter via stance transfer learning. In *Advances in Information Retrieval*, pages 575–588, Cham. Springer International Publishing.

L. Wang, Y. Wang, G. de Melo, and G. Weikum. 2018. Five shades of untruth: Finer-grained classification of fake news. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 593–594.

Yaqing Wang, Fenglong Ma, Z. Jin, Ye Yuan, G. Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. 2018. Ti-cnn: Convolutional neural networks for fake news detection.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach.