VIETNAMESE SMS SPAM DETECTION WITH DEEP LEARNING AND PRE-TRAINED LANGUAGE MODEL

Vu Minh Tuan[*], Nguyen Xuan Thang[*], Tran Quang Anh[+]
[*] Hanoi University
[+] Posts and Telecommunications Institute of Technology

Abstract: Despite of the strong development of OTT message applications and social networks, Short Service Message (SMS) keeps its undeniable role in the marketing industry. As a top level of effective and cost-saving advertising tool, SMS has also given rise to SMS spam. To contribute for the fight against SMS spam, we suggested a model which is the combination of deep learning neural network model and pre-trained language technique – PhoBERT, a variant of BERT. Making full usage of the pre-training Vietnamese data, the proposed model achieved good accuracy at 99.53% in detecting Vietnamese spam messages.

Keywords: Deep learning neural network, PhoBERT, SMS, spam detection, transfer learning, Vietnamese spam.

I. INTRODUCTION

Along with emails and social networks, short service message (SMS) system has contributed to revolutionizing information exchange in our modern life. According to a survey by the Cellular Telecommunications Industry Association (CTIA) published in 2019, the number of sent and received SMS in 2018 was about 2000 billion messages (above 5.5 billion ones in transactions per day on average) [1]. This number continues to increase gradually around 16% each year.

The development and popularity of information exchange via SMS also leads to an increase in the number of spam messages every year. The reason for this statement is totally straightforward and explainable. In comparing to email which only has a 20% open rate, that of SMS is 98% [2]. As a result, SMS users will quickly become the target of spammers, scammers, service providers, advertising services... In Vietnam, according to statistics of the Viet Nam Cybersecurity Emergency Response Teams (VNCERT)[1], everyday millions of spam messages are sent, equivalent to millions of subscribers being attacked by irrelevant information. Not only subscribers are suffering from the spam in term of extra fee charging, online phishing or malicious codes when following the messages but also the domestic mobile telecommunications infrastructure gets more pressure from spamming traffic.

Although current machine learning models could be helpful, they require manually extracting spam features and then using simple classifiers to detect spam from legitimate messages. These processes are time-consuming and not highly effective. Hence, there are still opportunities for future enhancement.

In this paper, authors proposed an approach using the deep learning neural network combining with a pre-trained language model known as PhoBERT to deal with the problem of Vietnamese spam messages.

The remaining sections of the paper are organized as follows: Section II reviews the related works and background knowledge; Section III presents the methodology for the research that includes the data collection, data preprocessing and the proposed model; The experimental setup and result analysis are discussed in Section IV before concluding and suggesting some future work in the last Section.

II. RELATED WORKS

A. Spam detection techniques

Over the years, machine learning techniques have been applied to detect spam from legitimate messages by researchers. Nilam Nur Amir Sjarif et al [3] proposed a technique combining TF-IDF with a random forest classifier and achieved the result at 97.5% in term of the accuracy. Pavas Navaney et al [4] showed good results by applying various machine learning algorithms. Especially, the SVM model brought the best result with an accuracy of 97.4%. Tian Xia and Xuemin Chen [5] proposed the Hidden Markov Model (HMM) for spam detection. The authors achieved an accuracy of 98% with that model.

Besides classical machine learning algorithms, researchers have used deep learning approaches for spam detecting. Most popular models were CNN, RNN or LSTM. A CNN architecture was proposed by Popovac et al [6] with one layer of convolution and pooling to detect SMS spam. The result was an accuracy of 98.4%. Jain et al [7] applied LSTM model to spam detection. Their model reached the accuracy of 99.01% with 6000 features and 200

Contact author: Vu Minh Tuan
Email: minhtuan_fit@hanu.edu.vn
Manuscript received: 31/10/2021, revised: 28/4/2022, accepted: 11/5/2022.

[1] http://www.vncert.gov.vn

LSTM nodes. In a recent publication, Wael Hassan Gomaa [8] presented a methodology based on the automatic feature extraction resulting a fabulous accuracy of 99.26%.

B. Vietnamese SMS spam filtering

In a recent study, Shafi'l, et al. went through more than fifty papers to review major research findings on mobile SMS spam [9]. According to the summary, SMS spam filtering solutions are designed to work on either in the Access Layer (end-user layer) or Service Provider Layer. Filtering techniques are allocated in some groups such as SVMS Based, Bayesian Network, Evolution Algorithms, Machine Learning, Writing Style, and other techniques. However, most of these worked based on English messages. There are some exceptions with non-English ones such as Turkish [10], Chinese [11], Korean [12] but no Vietnamese language at all.

By doing the literature review on the same research topic in Vietnam, the author realized that a few findings on SMS spam detection were published. Vu, et al. proposed an anti-SMS-spam model on the top of SpamAssassin with the TP rate at 94% and the FN rate at 0.15% [13]. This publishment confirmed with Vietnamese data input. Thai-Hoang Pham and Phuong Le-Hong presented a system for Vietnamese spam SMS filtering with a pretty good accuracy [14]. One of their research highlights was the method for data preprocessing step because existing tools for Vietnamese preprocessing cannot give good accuracy on their dataset. The authors admitted that the system achieved an accuracy of about 94% detecting spam message while the misclassification rate of ham was about 0.4%.

Most of proposed models were developed based on the machine learning approach with manual feature extraction process and classical classifiers such as SVM, Naïve Bayes, kNN… which were not efficient when the spammers keep tweaking their technology day by day. Thus, it could be a good idea to apply transfer learning technique to improve the performance of Vietnamese SMS spam filtering system.

C. Pre-trained language models

The pre-trained language models have been proven to be an effective solution for different NLP problems [15] [16] [17] [18]. BERT (known as Bidirectional Encoder Representations from Transformers) has recently become the most successful and popular representative of the model enabling anyone building a machine learning model involving language processing a readily available component [19]. It saves the time, energy, knowledge, and resource being taken to train a language-processing model from scratch. There are two main steps in the proposed approach: pre-training and fine-tuning.

In the first process – pre-training, the model is trained with unlabeled data through training sessions to obtain the configuration parameters for the model, then all this data will be passed through the training process. The two main techniques used in the pre-training process to build the language model of BERT are "Masked language modeling (MLM)" and "Next Sentence Prediction (NSP)". Based these techniques, the vectors built through the BERT model will represent the data segment or specifically in this study, the messages in the vector form in more detail. The quantities in the vector will be specific to each word in the sentence not repeating with those words in another sentence with different context.

All pre-trained data will go through the fine-tuning process for comparing with pre-labeled data for the purpose of evaluating changes in configuration parameters as well as applying to some specific problems. The fine-tuning process will be demonstrated by changing the hidden layers in the BERT deep learning model as well as the output layer so that the results will be returned in binary form 0 and 1 (0 for legitimate messages; 1 for spam).

However, the default BERT model is trained with English texts. This is also a huge limitation for detecting Vietnamese spam messages when the context cannot be shown or misrepresented if pre-learned models built on English vocabulary. Thus, the author suggested to use a variant of BERT trained with Vietnamese language – PhoBERT [20]. This model is based on RoBERTa optimizing the BERT pre-training procedure to improve the performance [21].

For the pre-training data, Nguyen et al., used 20GB dataset from Vietnamese Wikipedia (about 1GB) and from a Vietnamese news corpus (about 19GB after cleaning). Then the tool RDRSegmenter [22] was used to perform word and sentence segmentation in the data, resulting in nearly 145 million segmented sentences that serve as vocabulary for the model building process. PhoBERT provides two models respectively, $\text{PhoBERT}_{\text{base}}$ with 540,000 training steps in 3 weeks and $\text{PhoBERT}_{\text{large}}$ with 1.08 million training steps in 5 weeks.

III. METHODOLOGY

A. Data Collection

The dataset used in this study includes 5113 messages with 52% spam and 48% legitimate messages. This data was collected from individual users across various operating systems and devices focusing on messages sent via the telecommunication network to preserves the SMS's features such as the meta data, message's length limit… All messages are in Vietnamese language to maximize the efficiency of the proposed model. The raw data was collected and fetched into the tabular form so that could be processed easily with Pandas library[2].

	is_spam	message	num_words	message_len
0	0	Ả nhà châu cbi đi ăn thôi ạ	8	28
1	0	Hx nhà con đnag trả 1250 nhưng bọn con k bán	11	44
2	0	Đnag đòi 1290 k đồ	5	18
3	0	Đồ nào hợp con mang sang nhà mới còn đồ nào k ...	24	96
4	0	Cũng đỡ đc khối	4	15
...	...	...	...	...
3006	1	M/BAN 0884.888888-84 0882.888888-44 0838.83883...	35	160
3007	1	(QC) Dang ky dv Giai tri truyen hinh de co co ...	35	153
3008	1	LH 0828-228-228 MuaBanSim 0823888833=22 082448...	25	160
3009	1	CAU HG Lien tiep +24/4: 28-30 -> 30x2 +32/4: 3...	40	160
3010	1	GoldMark CiTy: Co Hoi So Huu va Dau Tu Chung C...	40	160

Figure 1 SMS dataset for the experiment

[2] https://pandas.pydata.org/

B. Data Preprocessing

The purpose of this step is to remove redundant data in the dataset. The redundant does not affect the classification of text messages at all. And of course, if we keep them when applying the model, it will lead to bad processing results. For SMS text message data, the noise data is paths, unnecessary phrases, characters that have no meaning and phone numbers. The removal was conducted with the regular expression filter.

Employing a Vietnamese word segmentation also makes the difference from the original BERT model comparing our approach. As suggested by the author of PhoBERT, the RDRSegmenter from VnCoreNLP[3] was used to process the input raw message.

In addition to the input message data and the vectors obtained after preprocessing, we added two quantities *num_word* (number of words in a message) and *message_len* (number of characters in a message). This adjustment is originated form the fact that legitimate messages among users will be shorter due to the character limit of each message, while spam messages will usually be longer to ensure full transmission of the advertisement content. The features of word count and character count in a message can be added to the input data to increase model accuracy. The graph below shows the difference between spam and regular messages. Most regular messages will contain between 0 and 10 words in a message, only a small number of regular messages will have up to 30 to 40 words. In contrast, most spam messages have a word count of between 30-40 words.

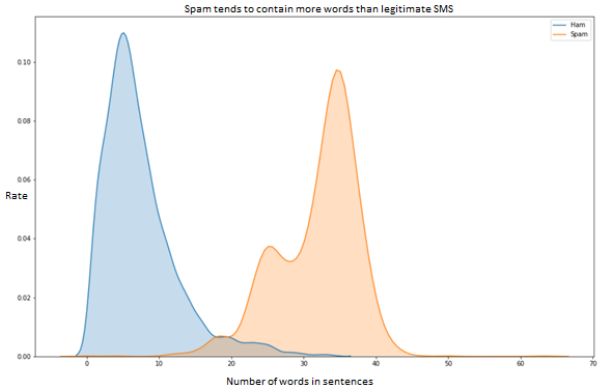

Figure 2 Spam and ham word-count comparison

However, adding these two features to the input data may cause the deviation when vectorizing the message using PhoBERT. Text preprocessed and encoded with PhoBERT will be converted into vector quantities with values in the range 0 - 1, while the two features related to word count and message length have average values as 35 (word count) and 150 (message length) on average. If not normalized, the range of these two features will directly affect the efficiency of the classification algorithm. Therefore, these two features need to be normalized to convert the range to about 0 - 1. The data normalization tool StandardScaler[4] in the *sklearn* library to achieve the task. The data normalization formula is presented as below:

$$x' = \frac{x - \mu}{\sigma}$$

In which:

- x': data standard point
- x: a single data
- μ: the mean value of all data
- σ: the standard deviation of all data

After preprocessing the raw data, we fed the processed text to the PhoBERT model which will eventually generate a contextualized word embedding for our training data.

C. Proposed Model

In this study, we proposed a Vietnamese SMS spam detection model as illustrated in Figure 3.

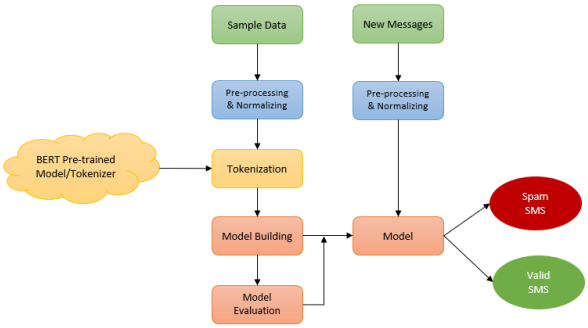

Figure 3 SMS Spam detection model with PhoBERT technique

To demonstrate the efficiency of the embeddings a deep learning neural network was created. The deep learning network model u has 04 hidden layers with the parameters of the components in the hidden layer being 256-256-128-10 respectively (Figure 4).

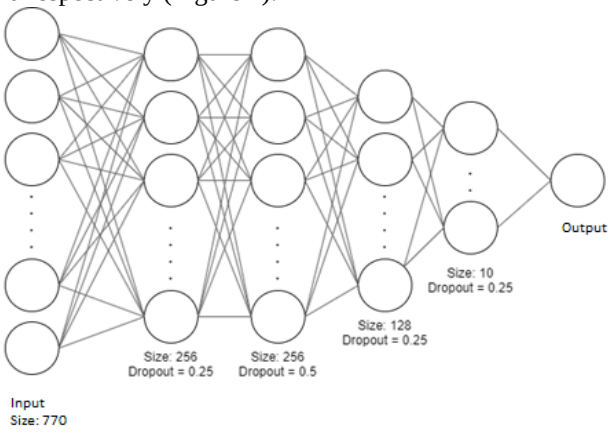

Figure 4 Structure of the DNN model

The input layer of the deep learning network has a size of 770 nodes to receive an input vector of length 770. The input vector length includes the encoding vector generated by PhoBERT of length 768 and 02 features (*word-count* and *message-length*). The values of these two features were normalized as described above. The use of Dropout parameter at each hidden layer skipping some random nodes is to avoid over-fitting during deep learning model training. The number of nodes of each layer is gradually reduced over each layer; so that the result can gradually

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

[4] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

converge at the output layer of the deep learning model. The output layer of the model returns a real number in the range 0 - 1 representing the probability that the input message is a legitimate message (close to 0) or a spam message (close to 1).

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Feature Normalization

As mentioned in III.B, adding two features *num_word* and *message_len* to the input data without normalizing may cause the deviation when vectorizing the message using PhoBERT. We firstly evaluated the performance of the model before and after normalizing added features. The diagram shows that the model's efficiency with the optimized quantities is more stable and more accurate.

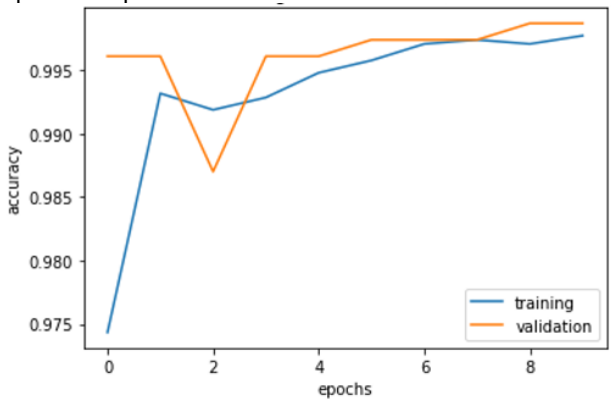

Figure 5 Model's Accuracy BEFORE normalizing features

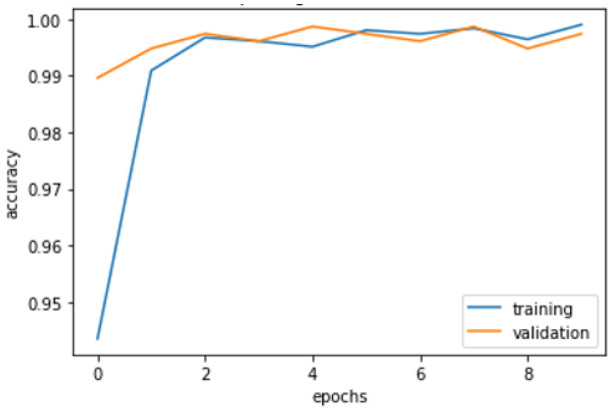

Figure 6 Model's Accuracy AFTER normalizing features

B. Classifiers

The proposed model was trained in 10 epochs. To evaluate the effectiveness of the model, the authors also did experiments with classical machine learning algorithms (SVM, NB and kNN) and a deep learning algorithm (CNN). These models were tested with the same training and validation dataset. The results were captured and represented in Table 1.

Table 1 Classifier testing results

	Accuracy	***Precision***	***Recall***
DNN PhoBERT	0.995	1.0	0.99
CNN	0.97	0.95	0.99
SVM	0.96	0.99	0.93
NB	0.93	0.96	0.90
kNN	0.89	0.99	0.78

The proposed model with the support of PhoBERT pre-trained language technique achieves higher accuracy (0.995) in comparing to that of CNN model (0.97). The SVM algorithm showed the best performance with the accuracy of 0.96 among the traditional machine learning ones. However, that result was still lower than that of DNN with PhoBERT pre-trained language model. We also recorded the same evaluating result in term of Precision. It seems that the context understanding of PhoBERT model is useful in supporting SMS spam detection.

V. CONCLUSION

In this study, a Vietnamese SMS spam detection approach with the deep learning neural network model has been suggested. The model's spotlight is the application of a pre-trained language technique – PhoBERT for data preprocessing (for Vietnamese language). We achieved positive results with the accuracy of **99.5%.**

In the future, we will not only focus on SMS but also other types of spam such as OTT application message, social network, and reviews… In term of the model improvement, the authors are planning to conduct more experiments with different parameters to optimize the model.

REFERENCES

[1] CTIA, "2019 Annual Survey Highlights," June 2019. [Online]. Available: https://www.ctia.org/news/2019-annual-survey-highlights. [Accessed October 2021].

[2] A. Doherty, "SMS Versus Email Marketing," July 2014. [Online]. Available: https://mobilemarketingwatch.com/sms-marketing-wallops-email-with-98-open-rate-and-only-1-spam-43866. [Accessed October 2021].

[3] Nilam Nur Amir Sjarif, "SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm," in *The 5th Information Systems International Conference 2019*, East Java, Indonesia, 2019.

[4] Pavas Navaney, Gaurav Dubey, Ajay Rana, "SMS Spam Filtering Using Supervised Machine Learning Algorithms," in *The 8th International Conference on Cloud Computing, Data Science & Engineering*, Noida, India, 2018.

[5] Tian Xia, Xuemin Chen, "A Discrete Hidden Markov Model for SMS Spam Detection," *Applied Sciences 10*, vol. 14, 2020.

[6] Milivoje Popovac, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, "Convolutional Neural Network Based SMS Spam Detection," in *2018 26th Telecommunications Forum (TELFOR)*, Belgrade, Serbia , 2018.

[7] Gauri Jain, Manisha Sharma, Basant Agarwal , "Optimizing semantic LSTM for spam detection," *International Journal of Information Technology*, vol. 11, pp. 239 - 250, 2019.

[8] W. Gomaa, "The Impact of Deep Learning Techniques on SMS Spam Filtering," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 544 - 549, 2020.

[9] Shafi'I Muhammad Abdulhamid; Muhammad Shafie Abd Latiff; Haruna Chiroma; Oluwafemi Osho; Gaddafi Abdul-

Salaam, "A Review on Mobile SMS Spam Filtering Techniques," *IEEE Access*, vol. 5, pp. 15650 - 15666, 2017.

[10] A. K. Uysal, S. Gunal, S. Ergin, and E. Sora Gunal, "The Impact of Feature Extraction and Selection on SMS Spam Filtering," *Elektronika ir Elektrotechnika*, vol. 19, pp. 67 - 72, 2012.

[11] Liumei Zhang, Jianfeng Ma, Yichuan Wang, "Content Based Spam Text Classification: An Empirical Comparison between English and Chinese," in *The 5th Intelligent Networking and Collaborative Systems (INCoS)*, 2013.

[12] D.-N. Sohn, J.-T. Lee, S.-W. Lee, J.-H. Shin, and H.-C. Rim, "Korean Mobile Spam Filtering System Considering Characteristics of Text Messages," *Journal of the Korea Academia-Industrial cooperation Society*, vol. 11, pp. 2595 - 2602, 2010.

[13] Vu Minh Tuan, Dang Dinh Quan, Nguyen Thanh Ha, Tran Quang Anh, "Lọc tin nhắn rác với Spam-Assassin," *Journal of Science and Technology on Information and Communications*, vol. 3, no. 4, pp. 34-41, 2017.

[14] Thai-Hoang Pham, Phuong Le-Hong, "Content-based approach for Vietnamese spam SMS filtering," in *2016 International Conference on Asian Language Processing (IALP)*, Tainan, Taiwan, 2017.

[15] Andrew M. Dai, Quoc V. Le, "Semi-supervised Sequence Learning," in *Advances in Neural Information Processing Systems* , 2015.

[16] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving language understanding with unsupervised learning," OpenAI, 2018.

[18] Jeremy Howard, Sebastian Ruder, "Universal language model fine-tuning for text classification," in *ACL. Association for Computational Linguistics*, 2018.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova., "BERT: Pre-training ofdeep bidirectional transformers for language understanding," in *Proceedings of NAACL*, 2019.

[20] Dat Quoc Nguyen, Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, "A Robustly Optimized BERT Pretraining Approach," *arXiv preprint, arXiv*, vol. 1907.11692 , 2019.

[22] Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, Mark Johnson, "A Fast and Accurate Vietnamese Word Segmenter," in *Proceedings of LREC*, 2018.

Vu Minh Tuan graduated from Hanoi University in 2010. He completed his Information System Design MSc in University of Central Lancashire, UK. His research interests include but not limited to spam detection, machine learning, system analysis and design…

Dr. Nguyen Xuan Thang is working for the Faculty of Information Technology at Hanoi University. His research interests span the fields of Wireless Networking, Software Engineering and Cyber Security. In Cyber Security, he has focused on the analysis of social, biological, and data networks to deal with security problems like spam, intrusion detection or firewall systems. He is the author and co-authors of over twenty papers on his fields.

Prof. Tran Quang Anh is currently the Vice President of Posts and Telecommunications Institute of Technology. He completed his Master and Doctoral programs at Tsinghua University, China. His research areas include network security, evolutionary algorithms, anti-s