Evaluating the efficiency of Vietnamese SMS spam detection techniques

Vu Minh Tuan, Nguyen Xuan Thang, Tran Quang Anh

***Abstract*— This paper is aimed at evaluating the efficiency of Vietnamese SMS spam detection methods on different variants of Vietnamese datasets by utilizing both traditional machine learning models and deep learning models. The researchers experimented with five algorithms, which were Support Vector Machine (SVM), Naive Bayes (NB), Random Forests (RF), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM), on three different Vietnamese datasets. The findings reveal that the LSTM and CNN, supported by a transformer learning model - PhoBert, were more efficient than the traditional machine learning models. The LSTM model showed the highest accuracy of 97,77% when operating on the full-accent Vietnamese dataset. Similarly, the CNN model and PhoBert model showed the highest accuracy of 95,56% when dealing with non-diacritic Vietnamese dataset.**

***Tóm tắt*— Bài báo sử dụng cả mô hình học máy truyền thống và mô hình học sâu để phát hiện tin nhắn rác SMS tiếng Việt nhằm đánh giá hiệu quả các mô hình trên các dạng tập dữ liệu biến thể tiếng Việt khác nhau. Nhóm tác giả đã thử nghiệm 5 thuật toán, bao gồm Support Vector Machine (SVM), Naive Bayes (NB), Random Forests (RF), Convolutional Neural Network (CNN) và Long Short-Term Memory (LSTM) trên bộ dữ liệu tiếng Việt có dấu, tiếng Việt không dấu và tập hỗn hợp. Các phát hiện cho thấy LSTM và CNN, được hỗ trợ kỹ thuật học chuyển đổi PhoBert, hiệu quả hơn các mô hình học máy truyền thống. Mô hình LSTM cho độ chính xác cao nhất là 97,77% khi thử nghiệm trên bộ dữ liệu tiếng Việt đầy đủ dấu. Tương tự, mô hình CNN kết hợp với PhoBert cho độ chính xác cao nhất là 95,56% khi xử lý bộ dữ liệu tiếng Việt không dấu.**

***Keywords*— SMS spam; Vietnamese SMS spam; machine learning; deep learning; transfer learning; PhoBert.**

***Từ khóa*— tin nhắn rác; tin nhắn rác tiếng Việt; học máy; học sâu; học chuyển đổi; PhoBert.**

This manuscript is received on March 08, 2023. It is commented on April 19, 2023 and is accepted on June 07, 2023 by the first reviewer. It is commented on May 08, 2023 and is accepted on June 05, 2023 by the second reviewer.

I. INTRODUCTION

Short Message Service (SMS) has become one of the most popular communication channels for decades. In 2021, people in US sent and received about 2.2 trillion SMS and MMS messages [1]. Main reasons for the popularity are the effectiveness of information delivering and being a low-cost solution. Nearly 96% of marketers using text messaging say it's helped them drive more revenue and almost 60% say it's significantly or overwhelmingly increased revenue generation [2]. Indeed, the more conveniently SMS messages were delivered to customers, the higher the likelihood of them being targeted by spammers. In Vietnam, according to the statistic of the Authority of Information Security (Ministry of Information and Communications), the carriers were requested to block about 9,6 million spam messages in only one week of Tet Holiday in 2022. Undoubtedly, while SMS brings significant benefits to businesses and communications, SMS spam has been recognized as having negative effects on the economy and customer behavior.

To mitigate the abuse of spammers, researchers have employed both traditional machine learning and deep learning techniques for detecting SMS spam. In a recent publication, Shafi'l et al. conducted a comprehensive review of over fifty studies to provide an overview of SMS anti-spam technology and its advancements [3]. These methods are relatively effective in detecting and preventing spam messages in English. However, the efficiency of content-based approaches such as traditional machine learning or even deep learning ones depends heavily on the language of the dataset.

Especially, for the language with different variants like Vietnamese, there is a research gap on how well these techniques work with each variant dataset.

The objective of this paper is to evaluate the efficiency of SMS spam detection algorithms including SVM, NB, RF, CNN, LSTM with different forms of Vietnamese dataset. Our main contributions are:

- We propose three sets of Vietnamese SMS datasets corresponding to two variants of Vietnamese (with/without diacritics and mixed type).
- These datasets are experimented with different anti-spam algorithms to figure out the most effective combination between these datasets and algorithms.

This paper is structured as follows: Research related to SMS spam detection techniques is reviewed in Section II; Section III describes the details of methodology with the data collection and preprocessing, vector representation and classifiers; The experiment setup and comparison results are discussed in Section IV. The author sums up the main achievement and suggests future of work in the last section.

II. Related Works

A. Spam detection techniques

For decades, a lot of filtering methodologies have been researched and applied to fight against the spam spread in general and SMS spam in detail. In this section, the author reviews some typical techniques to detect SMS spam from the classical to the modern ones.

Yadav et al. suggested an SMS spam filtering technique with Bayesian machine learning algorithm [4]. The authors used the user-generated features such as blacklist, whitelist, or preferred keywords to detect the unwanted message. A spam detection platform based on both SMS and emails was proposed by El-Alfy and AlHasan [5]. Two classifiers SVM and NB with 11 features were used in their experiments. They evaluated the efficiency of these algorithms on five datasets of SMS and emails. Narayan et al. developed a double layer SMS spam classifier [6]. The first layer captures the collection of words which have a higher probability than a pre-defined threshold. The second one is activated to select the word. Different pairs of machine learning algorithms (SVM - Bayesian, Bayesian - SVM, SVM - SVM, Bayesian - Bayesian) were combined to figure out the best one.

Although both machine-learning-based and deep-learning-based approaches were proposed, the later one brought more promising results. Deep learning models have been widely applied in the field of security [7, 8]. Some popular models were CNN, RNN or LSTM. A CNN architecture was proposed by Popovac et al. with one layer of convolution and pooling to detect SMS spam [9]. The result was an accuracy of 98.4%. Jain et al. applied LSTM model to spam detection. Their model reached the accuracy of 99.01% with 6000 features and 200 LSTM nodes [10]. In a recent publication, Wael Hassan Gomaa presented a methodology based on the automatic feature extraction resulting a fabulous accuracy of 99.26% [11]. Another neural network was proposed by Aliaksandr Barushka, Petr Hajek [12]. Their proposed model achieved an accuracy of 98.51%.

Upon conducting a literature review, the author observed a limited number of publications focusing SMS spam detection. In the past, the author proposed a rule-based approach with SpamAssassin application [13, 14]. The model achieved a true positive (TP) rate of 94% and a false negative (FN) rate of 0.15%. However, the main challenge faced by this proposed method is the inherent limitation of short content in SMS messages. Unlike emails, which can deliver rich content with longer text, SMS messages have a restricted character limit. Consequently, ensuring the quality of the generated rule becomes challenging in the context of SMS spam detection. Recently, Thai Hoang Pham and Phuong Le Hong presented a system for filtering Vietnamese spam SMS with commendable accuracy [15]. Notably, their research highlighted a novel data preprocessing method, as the existing tools for Vietnamese preprocessing did not yield satisfactory accuracy with their dataset. The authors acknowledged

that their system achieved approximately 94% accuracy in detecting spam messages, while the misclassification rate for legitimate messages (ham) was around 0.4%. However, this approach was solely based on traditional machine learning models such as NB, SVM, logistic regression (LR), decision tree (DT), and k-nearest neighbors (kNN). In contrast, deep learning methods have demonstrated greater promise in tackling text classification problems [16, 17]. In this article, experiments were conducted using the latest technology in Vietnamese SMS spam detection and data processing - a deep learning model combined with transfer learning (PhoBert) to evaluate the effectiveness of spam detection with different variations of the Vietnamese language.

B. Vietnamese SMS with and without diacritics

One feature when studying SMS spam is the linguistic variations of the message text. Due to the limit on the number of characters in each message, users often tend to abbreviate or use symbols and emoticons. Each language has its own set of symbols. In addition to emotional symbols and abbreviations, there exist at least two forms of language expression: Standard Vietnamese with diacritics and Vietnamese without diacritics. The use of these two forms has no definite rule, completely depends on the user's habits and the Unicode support of the device they use.

Most of messages in the spam dataset used in this paper are in the second form – no diacritics. The reason for this fact is explainable. It is interesting that Vietnamese users can easily understand the content of the message no matter it comes with or without diacritics. However, a message in standard Vietnamese contains more characters than the "no-sign" one. As a result, the spammers prefer to spread SMS Spam with the second form for cost-saving purpose while the effectiveness of content delivering mostly remains unchanged.

In this paper, we aim at building SMS datasets in both diacritic and non-diacritic Vietnamese languages. Therefore, a comprehensive literature review was conducted to explore methods for diacritic restoration in the original text. Nguyen an Ock proposed a machine learning approach with AdaBoost and C4.5 algorithm to achieve an accuracy of 94.7% on their dataset [18]. Deep learning methods were also applied as a novel combination of a character-level recurrent neural network-based model and a language model applied to diacritics restoration and reached the highest accuracy of 97.73% on Vietnamese by Náplava et al [19]. There are some limitations of those methods such as producing non-existent outputs and incorrect recovering in some special contexts. Moreover, there is no proof of the efficiency when being applied to other data. Thus, the author decided to restore the diacritic of the text manually with the help of volunteers. With two levels of control, manual data processing ensured that the data was handled with high accuracy and provides more oversight compared to machine learning methods.

III. METHODOLOGY

A. Preprocessing

Before training any model, it is essential to preprocess the data. This step will help correct the data content, remove redundant content. This step done well will significantly increase the accuracy of the model.

1. Data cleaning: The purpose of this step is to remove redundant data in the data. These redundant data do not affect the classification of text messages at all, and of course, if they are left alone, the application of the model will lead to bad processing results. For SMS text message data, the noise data is usually paths, possibly unnecessary phrases, or characters that have no meaning.

2. Stemming: The goal is to bring documents from heterogeneous forms into the same form. From the perspective of optimizing storage memory and accuracy is also very important. For example: [Q.C], [QC], (QC), Q.C, Q&C, quảng cáo, quang cao → QC.

3. Stop-word removal: The model of the algorithm implementation has a segmentation treatment that removes StopWords in the data. The list of stopwords in Vietnamese used includes 1,942 words suggested by Le Van Duyet, who created a dictionary of Vietnamese Stopwords.

4. Word segmentation: Word segmentation technique is a processing process for the purpose of determining the boundaries of words in a sentence. It can also be understood simply that word separation is the process of identifying single words, compound words, etc. in a sentence. For Vietnamese processing, to determine the grammatical structure of a sentence and the type of a word in a sentence, it is necessary to determine which word is in the sentence.

B. Vector representation

The message raw data is text and needs to be converted to vector for use with machine learning models.

1. Bag of Word (BoW): BoW learns a set of words from all the documents, then models the documents by counting the number of occurrences of each word in the text. BoW doesn't care about the word order in sentences and word semantics.

2. Term frequency - Inverse document frequency (TF-IDF): BoW still has shortcomings, so TF-IDF methods are used in combination to overcome this problem. TF-IDF is used to evaluate and rank a word in a message body. TF-IDF value is a parameter obtained through statistical methods, showing the importance of a word in the message content in a set of message contents [20].

3. PhoBert: PhoBert is a pre-trained language model developed for Vietnamese language processing tasks, based on the transformer architecture. It was trained on a large corpus of Vietnamese text data and fine-tuned on specific NLP tasks such as named entity recognition, part-of-speech tagging, and sentiment analysis [21]. In this paper, we use PhoBert as a technique to generate contextualized word representations for Vietnamese SMS message input of the deep learning models.

In this paper, the authors used TF-IDF for traditional machine learning algorithms while PhoBert was applied with the deep learning ones.

C. Classifiers

To evaluate the efficiency of spam detection algorithm with different variants of Vietnamese datasets, both traditional machine learning algorithms and deep learning ones were applied in the scope of this paper.

1. Traditional machine learning algorithms

After preprocessing the data, we extracted occurrence words as features and used TF to select the features. The term frequency (TF) of each word in a document is weight dependent upon the distribution of each word in the files [22]. Adding this feature to the feature matrix is an important stage in the general traditional machine learning model to detect spam with SVM [23], NB [24] and Random Forest [25] as shown in Figure 1.

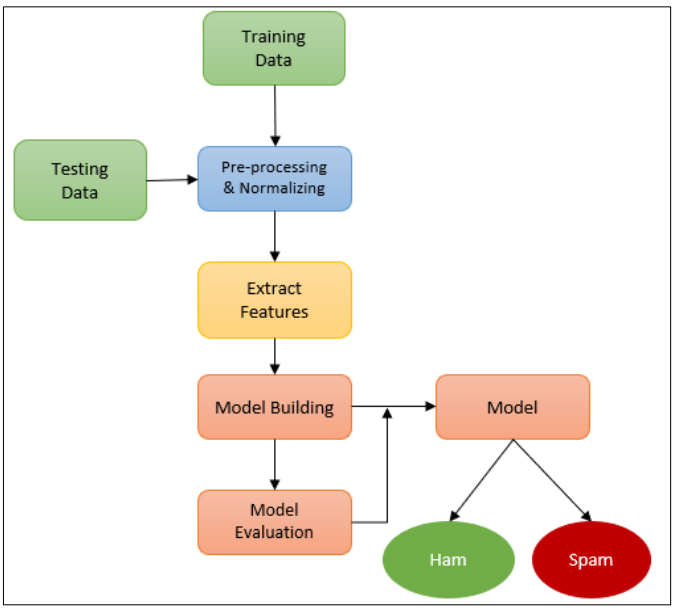

Figure 1. Traditional Machine Learning Model

2. Deep learning algorithms

- Convolutional Neural Network (CNN) model: The CNN model has become more and more popular algorithm in machine learning. The CNN is one of the deep learning models that can automatically extract relevant features from data. A CNN-based spam filtering model typically consists of three processes as depicted in Figure 2:

- Creating word matrix.
- Identifying hidden feature from text.
- Performing message classification into predetermined classes.

We use an embedding layer with PhoBert and convolutional layer, dropout layer, max

pooling layer. Relu method is applied as the activation function.

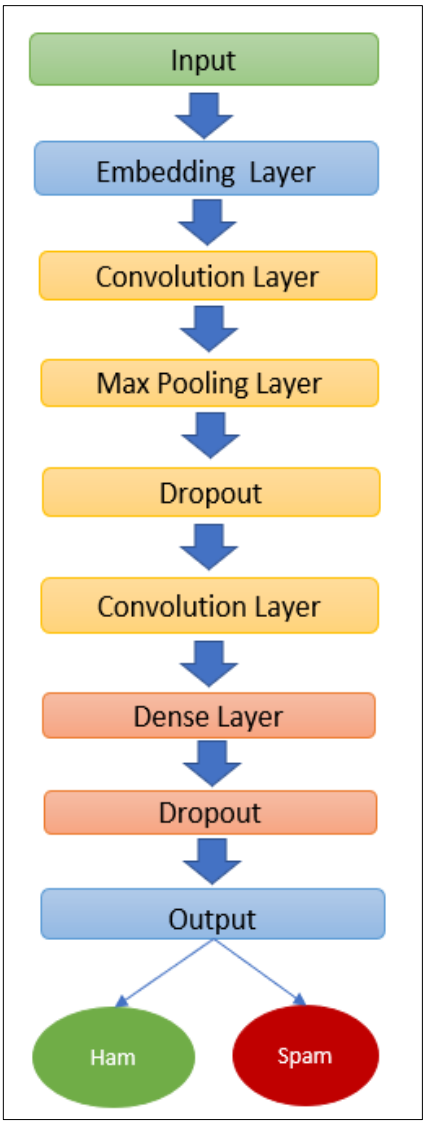

Figure 2. Convolutional Neural Network Model

- *LSTM model:* The LSTM is a commonly used type of RNN because it can solve the problem of gradients disappearing after a few steps of classical RNNs. The CNN model can extract hidden features from text data. However, it cannot remember long strings of text. The LSTM network can remember long text sequences.

The first layer maps each word to an N-dimensional vector of real numbers and is known as a pre-trained embedding layer. The second layer is an RNN with LSTM units. The final layer is the output layer, with two neurons corresponding to "spam" or "ham" with sigmoid activation functions. In this experiment, we use PhoBert and 100-dimensional vector with Softmax activation function.

D. Data Collection

The original dataset (dataset 1) in this paper includes about 10.000 spam SMS provided by Vietnam Computer Emergency Team (VNCERT) and 6.000 legitimate messages collected from the author's daily usage and volunteers. All messages are in Vietnamese. However, most of spam messages are non-diacritic while as all ham messages are standard Vietnamese with full accents. Two variants of datasets were created by restoring and removing the diacritics of the original messages. Experiments were performed with dataset variations in turn to determine the most effective data and algorithm pairs in detecting Vietnamese spam messages. For the dataset without accents (dataset 2), the author uses the *unidecode* library in Python to convert standard Vietnamese (which uses diacritics) to unaccented Vietnamese. For the dataset with full accents (dataset 3), non-diacritic messages were recovered manually by the volunteer to assure the accuracy of the conversion.

IV. Experimentation and Results

A. Evaluation Metric

We use Precision, Recall and Accuracy as measuring criteria for evaluating the performance of algorithms.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

B. Results and Discussion

For evaluating the efficiency of spam detection models with different variations of Vietnamese dataset, we implemented both traditional machine learning algorithms and deep learning algorithms with the support of PhoBert

TABLE 1. EXPERIMENT RESULTS

Models	Dataset 1			Dataset 2			Dataset 3		
	Acc.	***Rec.***	***Prec.***	***Acc.***	***Rec.***	***Prec.***	***Acc.***	***Rec.***	***Prec.***
SVM	85,43	83,75	84,85	87,43	87,56	87,25	87,9	87,14	89,42
NB	79,77	81,02	81,36	81,15	80,17	81,45	81,23	82,13	81,03
RF	87,58	85,86	86,54	88,32	87,15	86,06	88,55	86,02	87,55
CNN	95,03	94,35	95,15	**95,56**	93,57	94,72	96,04	96,3	96,68
LSTM	**95,66**	94,05	96,36	94,89	93,12	95,16	**97,77**	96,85	97,32

- a transfer learning technique. Five models were trained and tested with three datasets mentioned in Section III.D.

The data was divided into 3 subsets: 70% for training, 15% for validation and 15% for testing. We applied the cross-validation technique to evaluate the performance of the model and reduce the risk of overfitting with n-folds. The n here is assigned to 10. Table 1 shows the accuracy, recall and precision of Vietnamese SMS spam detection models on different datasets.

According to Figure 3, the accuracy of deep learning models with the support of PhoBert are higher than that of the traditional machine learning models. For experiments carried on dataset 1 and dataset 3, the LSMT model also occupies the highest position with the rate of accuracy 95,66% and 97,77% namely. However, when models were trained and tested with the non-diacritic Vietnamese dataset (dataset 2), the CNN model is ranked first among five classifiers. The accuracy reaches up to 95.56% in comparing to 94.89% which is the accuracy rate of the LSTM model. The high accuracy of the LSTM on dataset 3 (97.77%) makes the results highly reasonable. It proves that the contextual representation of the input text by PhoBert works more effectively with diacritic Vietnamese dataset.

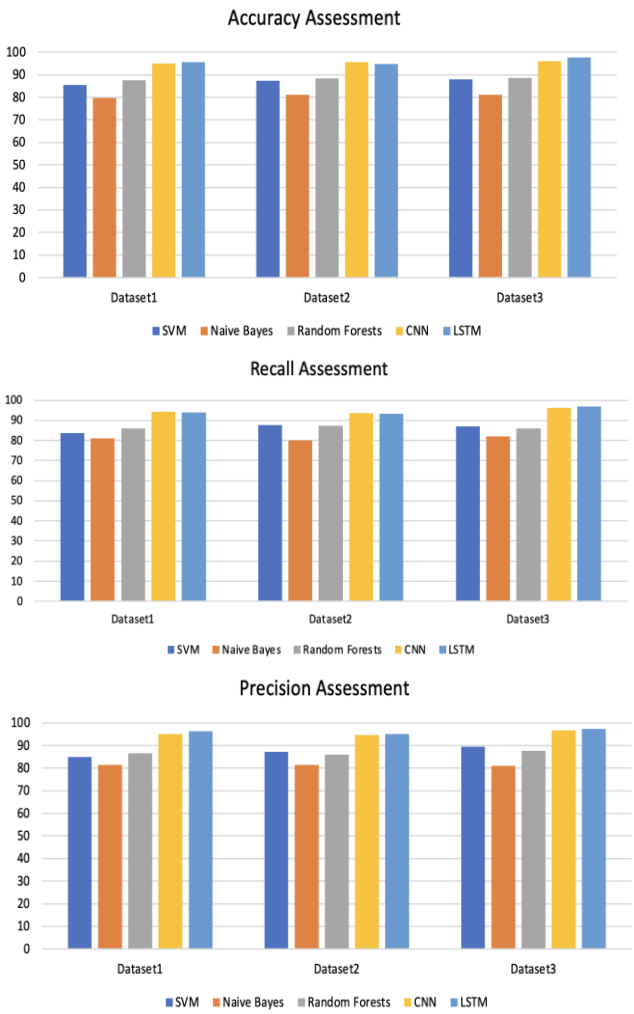

Figure 3. Comparison of five models' performance on 3 variants of Vietnamese datasets

V. CONCLUSION

In this paper, both traditional machine learning models and deep learning models were deployed against the task of detecting Vietnamese SMS spam. Five selected algorithms including SVM, NB, RF, CNN, and LSTM were in turn experimented on three variants of Vietnamese dataset. Our results show that the LSTM and CNN with the support of PhoBert transformer model perform more efficiently than the group of traditional machine learning models. The achievements are even better when the LSTM

and PhoBert works on the full-accent Vietnamese dataset with 97,77% accuracy. The CNN model and PhoBert achieve the highest score of accuracy at 95,56% with non-diacritic Vietnamese dataset.

For the future work, the automation technique to restore the diacritic of Vietnamese SMS should be evaluated and applied to reduce the manual works as the fact profs that the highest rate of accuracy accomplished with the full accent Vietnamese data. Besides, the authors are also working on some AI techniques such as GPT-3 to leverage the result of spam detection methods.

REFERENCES

[1] CTIA, "2021 Annual Survey HIGHLIGHTS," 2021. [Online]. Available: https://www.ctia.org/news/2021-annual-survey-highlights.

[2] Attentive, "2021 SMS Marketing Benchmarks Report," 2021. [Online]. Available: https://www.attentivemobile.com/2021-sms-marketing-benchmarks-report. [Accessed 2022].

[3] 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.

[4] K. Yadav, S. K. Saha, P. Kumaraguru, and R. Kumra, "Take control of your smses: Designing an usable spam sms filtering system," in 2012 IEEE 13th International Conference on Mobile Data Management, Bengaluru, India, 2012.

[5] El-Alfy, E.-S.M. and AlHasan, A.A., "Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm," Future Generation Computer Systems, vol. 64, pp. 98-107, 2016.

[6] A. Narayan and P. Saxena, "The curse of 140 characters: evaluating the efficacy of sms spam detection on android," in Third ACM workshop on Security and privacy in smartphones & mobile devices, Berlin, Germany, 2013.

[7] Doan Trung Son, Nguyen Thi Khanh Tram, Pham Minh Hieu, "Deep Learning Techniques to Detect Botnet," Journal of Science and Technology on Information Security, vol. 1, no. 15, pp. 85-91, 2022.

[8] Tran Ngoc Quy, Nguyen Thanh Tung, Do Quang Trung, Dang Hung Viet, "Convolutional neural network based sidechannel attacks," Journal of Science and Technology on Information Security, vol. 1, no. 15, pp. 26-37, 2022.

[9] 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.

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

[11] 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.

[12] Aliaksandr Barushka, Petr Hajek , "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks," Applied Intelligence , vol. 48, p. 3538–3556, 2018.

[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] Vu Minh Tuan, Quang Anh Tran, Minh Quang Ha, Lam Bui Thu, "A Multi-objective Approach for Vietnamese Spam Detection," in Knowledge and Systems Engineering 2013, Hanoi, 2014.

[15] Hoang Thai Pham, Le-Hong Phuong, "Content-based Approach for Vietnamese Spam SMS Filtering," in The 20th International Conference on Asian Language , Taiwain, 2016.

[16] R. Johnson, T. Zhang, "Supervised and semi-supervised text cat- egorization using lstm for region embeddings," in The 33rd International Conference on Machine Learning, New York, 2016.

[17] X. Zhang, J. Zhao, Y. LeCun, "Character-level convolutional networks for text classification," in The 28th Advances in Neural Information Processing Systems, Quebec, 2015.

[18] Kiem-Hieu Nguyen, Cheol-Young Ock, "Diacritics Restoration in Vietnamese: Letter Based vs. Syllable Based Model," in PRICAI 2010: Trends in Artificial Intelligence, Berlin, Heidelberg, 2010.

[19] Jakub Náplava, Milan Straka, Pavel Straňák, Jan Hajič, "Diacritics Restoration Using Neural Networks," in the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018.

[20] Hilal Tekgöz; Halil İbrahim Çelenli; Sevinç İlhan Omurca, "Semantic Similarity Comparison of Word Representation Methods in the Field of Health," in 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021.

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

[22] G. Forman, "BNS feature scaling: an improved representation over tf-idf for svm text classification," in Proceedings of the 17th ACM conference on Information and knowledge management, Napa Valley California USA, 2008.

[23] J.A.K. Suykens; J. Vandewalle , "Least Squares Support Vector Machine Classifiers," Neural Processing Letters , vol. 9, pp. 293 - 300, 1999.

[24] George H. John, Pat Langley, "Estimating Continuous Distributions in Bayesian Classifiers," in Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995), Quebec, Canada, 1995.

[25] L. Breiman, "Random Forests," Machine Learning volume , vol. 45, no. 1, pp. 5-32, 2001.

ABOUT THE AUTHOR

Vu Minh Tuan

Workplace: Hanoi University.

Email: minhtuan_fit@hanu.edu.vn

Education: He graduated from Hanoi University in 2010, completed his Information System Design MSc in University of Central Lancashire, UK.

Recent research detection: Spam detection; Machine learning; System analysis and design.

Cơ quan làm việc: Đại học Hà Nội.

Email: minhtuan_fit@hanu.edu.vn

Quá trình đào tạo: Tốt nghiệp Đại học Hà Nội năm 2010, hoàn thành Thạc sĩ Thiết kế Hệ thống Thông tin tại Đại học Central Lancashire, Vương quốc Anh.

Hướng nghiên cứu hiện nay: Phát hiện thư rác; Học máy; Phân tích và thiết kế hệ thống.

Nguyen Xuan Thang

Workplace: Hanoi University.

Email: nxthang@hanu.edu.vn

Education: He received his PhD degree from Kassel University, Germany.

Recent research detection: Wireless Networking; Software Engineering; Cyber Security; Spam detection; Intrusion detection; Firewall systems.

Cơ quan làm việc: Đại học Hà Nội.

Email: nxthang@hanu.edu.vn

Quá trình đào tạo: Nhận bằng Tiến sĩ tại Đại học Kassel, Đức.

Hướng nghiên cứu hiện nay: Mạng không dây; Kỹ thuật phần mềm; An ninh mạng; Phát hiện thư rác; Phát hiện xâm nhập; Hệ thống tường lửa.

Tran Quang Anh

Workplace: Posts and Telecommunications Institute of Technology.

Email: tqanh@ptit.edu.vn

Education: He received his MSc, PhD degrees from Tsinghua University, China.

Recent research detection: Network security; evolutionary algorithms, Anti-spam.

Cơ quan làm việc: Viện Công nghệ Bưu chính Viễn thông.

Email: tqanh@ptit.edu.vn

Quá trình đào tạo: Nhận bằng Thạc sĩ, Tiến sĩ tại Đại học Thanh Hoa, Trung Quốc.

Hướng nghiên cứu hiện nay: An ninh mạng; thuật toán tiến hóa, chống thư rác.