An Empirical Investigation of Online News Classification on an Open-domain, Large-scale and High-quality Dataset in Vietnamese

Khanh Quoc Tran[1,2,*], Phap Ngoc Trinh[1,2,*], Khoa Nguyen-Anh Tran[1,2,*],
Luan Van Ha[1,2,*], An Le-Hoai Tran[1,2,*], and Kiet Van Nguyen[1,2,†]

[1]University of Information Technology, Ho Chi Minh City
[2]Vietnam National University, Ho Chi Minh City
Email:*{*18520908, 18521227, 18520938, 18521062, 18520426*}*@gm.uit.edu.vn*
†*kietnv@uit.edu.vn*

Abstract. In this paper, we build a new dataset UIT-ViON (**Vi**etnamese **O**nline **N**ewspaper) collected from well-known online newspapers in Vietnamese. We collect, process, and create the dataset, then experiment and evaluate on it using different types of machine learning. In particular, we propose an open-domain, large-scale, and high-quality and dataset consisting of 260,000 textual data points annotated with multiple labels for evaluating Vietnamese short text classification. In addition, we evaluate three types of machine learning techniques including traditional machine learning (Naive Bayes, Logistic Regression), deep learning (Text-CNN, LSTM), and transfer learning (PhoBERT) for Vietnamese short text classification on the dataset. As a result, our best model achieves the F1-score of 80.62%. In the future, we will propose solutions to improve the quality of the data set and improve the performance classification of future models.

Keywords: Building dataset · Vietnamese news classification · Machine Learning · DeepLearning · Transfer Learning · Comparision

1 Introduction

Undergoing dramatic changes in the past two decades with the advent of digital technology and the dissemination of information on the Internet has led to online newspapers' birth. Online newspapers are types of writing newspapers built on a website and published on the Internet platform. This has created a change in viewing newspapers; people are more likely to read news through smartphones and electronic devices. Online newspapers update news regularly and allow people worldwide to access them quickly, regardless of time or space. With the number of articles being updated to the website every day, up to thousands of articles with various life topics, manually categorizing each article's headings into different topics is not an easy job. Instead, the automatic classification of documents into a suitable topic makes it easier to organize, archive, and retrieve documents in the future.

We use traditional machine learning models and neural network models to automatically classify messages into specific topics to solve the problem. Nevertheless, to be able to do that job, the above models must be learned through classification. They learn that through datasets with various features, the models' ability to classify also depends on the dataset training them. Since then, we have collected data from reputable online newspapers in Vietnam to create a large-scale and high-quality dataset with 260,000 data points divided equally among 13 different topics. We focus on developing and improving datasets and evaluating classification capabilities using traditional machine learning models: Logistic Regression, MultinomialNB; Deep Neural Models: Text-CNN, LSTM; and Transformers Model - PhoBERT.

This problem aims to build and develop a large-scale, high-quality, and open-domain dataset of Vietnamese newsletter headlines from leading Vietnamese online newspapers. From there, apply and evaluate machine learning, deep learning, and transfer learning models' performance to classify newsletter headings by topic.

In this article, we focus on introducing information related to constructing the Vietnamese newspaper title dataset. In section 2, we present some related research. Continuing in section 3, we present details about collecting, processing, and creating datasets. In Section 4, solutions and models are presented by us. The test results evaluated and analyzed in section 5. Finally, section 6 is the conclusion, and future development direction for Vietnamese news topic classification problems in general and recognition problems classify article titles in particular.

2 Related work

Hoang et al. (2007) [3] present a comparative study on Vietnamese text classification methods. Specifically, this study evaluated two popular approaches for text classification: Bag Of Words and Statistical N-Gram Language Modeling - N-Gram. The two approaches achieve an average accuracy of over 95% and take 79 minutes for 14,000 data points. They also point out some advantages and disadvantages of each method. The evaluation dataset was collected from four Vietnamese online newspapers: VnExpress, Tuoi Tre Online, Thanh Nien Online, Nguoi Lao Dong Online. The dataset contains more than 100,000 data points divided into two levels. This dataset is preprocessed using Teleport software, heuristics, and manually tagged.

Pham et al. (2017) [13] studied Vietnamese news classification based on BoW with Keywords Extraction and Neural Network. Using Bag of Words (BoW) with Keywords Extraction and Neural Network, they trained a machine learning model that could achieve an average accuracy of 99.75 % based on the dataset built by Hoang et al. [3]. This study shows that the feature extraction with Keywords Extraction and BoW is more efficient than other feature extraction methods. The study also shows that the Neural Network gives higher average accuracy than SVM, Random Forest, or SVC. They obtained higher results than the study of Hoang et al. [3] with the same algorithm and same dataset.

Nguyen et al. (2020) [10] released pre-trained language models for Vietnamese. They presented two versions of PhoBERT-base and PhoBERT-large. Experimental results show that PhoBERT consistently outperforms the recent best pre-trained multilingual model XLM-R [1] and improves state-of-the-art multiple Vietnamese-specific NLP tasks, including Part-of-speech tagging, Dependency parsing, Named-entity recognition, and Natural language inference. The PhoBERT pre-training approach is based on RoBERTa [8], optimizing the BERT pre-trained procedure for more robust performance. The pre-trained data is a concatenation of two corpora: the first one is the Vietnamese Wikipedia corpus (1GB). The second corpus (19GB) is generated by removing similar articles and duplication from a 50GB Vietnamese news corpus.

3 Dataset

In this section, we present basic information about the dataset, collection process and the challenges we face on the ViON: Vietnamese Online Newspapers dataset.

3.1 Data collection

We collect data based on the six largest circulation Vietnamese online newspapers[1]: VnExpress, Vietnamnet, Bao Moi Online, Dan Tri Online, Lao Dong Online, TuoiTre Online. The data is automatically collected according to the respective titles with Python using Beautiful Soup[2] library. There followed a stage of manual correction, we review and adjust the documents which are classified to the inappropriate topics with the guideline set, which was built with the support of one student in Journalism. With the above guideline set, we have improved the ViON dataset quality, contributing to increasing the models' classification efficiency. Finally, we obtain a relatively large and reliable dataset with 260,000 data points.

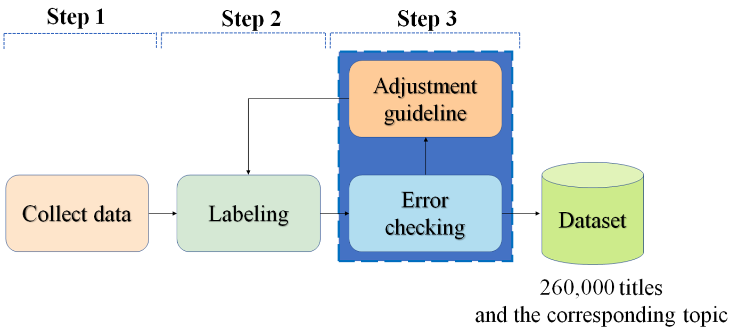


Fig. 1: The process of building ViON dataset

We build the **ViON: Vietnamese Online Newspapers** dataset to create a dataset with a larger size, higher timeliness, and higher quality than another that has the same category on Vietnamese news classification done in the past by other authors [3]. The ViON dataset with 260,000 data points are divided equally into 13 topics corresponding to the content that it covers. This dataset is suitable for the development of Machine Learning models and Deep Learning models in the field of natural language processing in general and Vietnamese short text classification in particular. We intend to develop the dataset in the future, such as: adding data, extracting the content of the news for classification, building ensemble models to increase accuracy and efficiency in classification.

3.2 The challenges

In the construction of the ViON dataset, we encounter three challenges:

- Lack of context: Initially, we collected 520,000 news titles on their respective topics. As a result, after data preprocessing and cleaning, we face a lack of context in some titles that made it difficult to classify.
  For example: "Đây là cái gì" → "(English: What is this)

[1] Top 30 Vietnam Newspaper - https://newspaperlists.com/vietnam
[2] Beautiful Soup - https://www.crummy.com/software/BeautifulSoup/bs4/doc/

- Noise data: The titles of news are always presented concisely, some of them do not cover the whole content of the news, so it is easy to confuse the classification of titles into appropriate topics;
  For example: "Rao trên mạng đòi giết và ăn thịt người, đối mặt án chung thân." (English: Claiming online to kill and eat people, face a life sentence)
  The title of this newspaper has the phrase "life sentence" usually belongs to the topic "Law" but the article is about a case in the US, so this article belongs to the issue "World."
- Topics have similar content: Newpapers in one issue could be on the other, making it difficult to classify.
  For example: "Top 10 thực phẩm tốt cho da" (English: Top 10 foods good for the skin)
  This newspaper could be under "Life" topic or "Health" topic.

3.3 Dataset information

The dataset includes 260,000 data points with three attributes: "title" (the title of newspaper), "link" (link to the website of newspaper), "label" (the topic of newspaper). The attribute "label" is the target attribute to predict, including 13 topics corresponding to the content of the newspapers that it mentions. We divide the data into training set, validation set and test set with the ratio of 8:1:1, as described in Table 2 and 1 below:

Table 1: Statistics of the ViON dataset

Label	**Training set**		**Validation set**		**Test set**	
	Number of titles	**Average length**	**Number of titles**	**Average length**	**Number of titles**	**Average length**
Technology	16,000	10.8	2,000	10.8	2,000	10.8
Travel	16,000	11.2	2,000	11.2	2,000	11.2
Educatioin	16,000	12.6	2,000	12.6	2,000	12.6
Entertainment	16,000	11.5	2,000	11.5	2,000	11.4
Science	16,000	11.1	2,000	11.2	2,000	11.1
Business	16,000	11.3	2,000	11.3	2,000	11.3
Law	16,000	13.4	2,000	13.4	2,000	13.3
Health	16,000	11.1	2,000	11.2	2,000	11.1
World	16,000	12.5	2,000	12.5	2,000	12.6
Sport	16,000	10.3	2,000	10.2	2,000	10.3
News	16,000	11.8	2,000	11.8	2,000	11.8
Vehicle	16,000	11.2	2,000	11.2	2,000	11.2
Life	16,000	11.1	2,000	11.0	2,000	11.1
Total	**208,000**	**11.5**	**26,000**	**11.5**	**26,000**	**11.5**

Table 2: Sample data from dataset on 13 classes

Title	**Label**
Fujifilm ’chọc’ Xiaomi vì điện thoại 108 megapixel (**English:** Fujifilm ’teased’ Xiaomi for its 108-megapixel phone)	Technology

Table 2: Sample data from dataset on 13 classes

Title	Label
7 món ăn "thử là mê" ở Malaysia (**English:** 7 "try to be mesmerizing" dishes in Malaysia)	Travel
Học sinh lớp 1 ở Mỹ được học những gì? (**English:** What are students in grade 1 in America learning?)	Education
Sao Hollywood cưỡi lạc đà trong ngày cưới (**English:** Hollywood star riding camels on the wedding day)	Entertainment
Những dấu tích cuối cùng của loài khủng long (**English:** The last traces of dinosaurs)	Science
Đề xuất giảm giá dịch vụ hàng không (**English:** Proposed discount airline service)	Business
Nhận tiền làm từ thiện có vi phạm pháp luật không? (**English:** Does accepting money for charity violate the law?)	Law
Bí quyết đơn giản giúp sống trẻ (**English:** Simple tips to help live young)	Health
Quốc hội Anh bác bỏ thoả thuận Brexit (**English:** British Parliament rejects the Brexit deal)	World
Đi tìm “chìa khóa” thành công của Arsenal (**English:** Finding the "key" of Arsenal’s success)	Sport
Thầy giáo cứu cậu bé chết đuối (**English:** The teacher saved the drowned boy)	News
Honda và Mitsubishi mua lại ô tô bị lỗi túi khí (**English:** Honda and Mitsubishi recall cars with defective airbags)	Vehicle
Làm gì khi tỏ tình không thành công? (**English:** What to do when confessing unsuccessful?)	Life

4 The Methodologies

4.1 Word embedding

Word embedding is a feature learning technique in Natural Language Processing (NLP) by mapping words or phrases from the set of vocabularies to a vector of real numbers [11]. The distributed representations of words, called word embedding, help improve various natural language models' accuracy.

This paper uses the Vietnamese word embedding ETNLP[3] provided by Vu Xuan Son et al. [14], which in dimension 300 with character n-grams. This pre-trained embedding model will be used as an embedding layer in the neural network models training.

4.2 Traditional machine learning models

This study conducts experiments on three different types of text classification models, including traditional machine learning, deep learning, and transfer learning. We describe these models that we use in detail as follows.

[3] Vietnamese Embedding ETNLP - https://github.com/vietnlp/etnlp

4.2.1 Multinomial Naive Bayes

This is an algorithm based on the Bayes theorem of probability theory to make predictions and classify data based on observed data and statistics [12], [6]. Multinomial Naive Bayes (MultinomialNB) Classification is one of the many algorithms used in machine learning to make the most accurate predictions on a collected dataset because it is relatively easy to set and give high accuracy. It belongs to the Supervised Machine Learning Algorithms group, which means machine learning from examples from existing data samples.

4.2.2 Logistic Regression

This is one of the basic and well-known methods of classification algorithms, especially binary classification. In text classification, it requires manual feature extraction [2,5].

In this paper, we have used Logistic Regression and MutinomialNB model with TF-IDF technique through TfidfVectorizer function in sklearn library with parameter ngram_range is (1,2).

4.3 Deep neural network models

4.3.1 Text-CNN

Convolutional neural network (CNN) is a multistage Neural network architecture developed for classification [7]. By using convolutional layers, it can detect combination features. In our experiments, we use 4 convolutional layers with 32 filters for each layer.

Filters perform convolution analysis on the input sentence, each creating a feature map. Feature maps go through max-pooling to obtain maximum value. Each vector has 1 element created in each feature map. Finally, the softmax function uses the result to predict the label for the text.

4.3.2 Long Short-Term Memory (LSTM)

LSTM [4] is a special kind of RNN [9], it is also a modern classification method. LSTM's network architecture includes memory cells and ports that allow the storage or retrieval of information. This method is strong in classification problems, and most of it has achieved high-performance classification results. Therefore, we decide to choose it to compare with other classification models in our works.

4.4 Transformers Model - PhoBERT

This library is based on HuggingFace's Transformers library[4]. Transformers provides thousands of pre-trained models to perform tasks on text such as: classification, extraction information, answering questions, summarization, translation, text generation in more than 100 languages. Its purpose is to make advanced natural language processing easier to use for everyone.

The use of Transformers Model, here is PhoBERT [10] allows a simpler implementation of complex model instruction from pre-built pipeline libraries at a high api level with the expectation that highly effective classification model for Vietnamese text classification problems, especially Vietnamese orthodox text classification.

5 Experiments

5.1 Data preparation

Based on the data described in Section 3, we first apply the preprocessing technique described below to create the cleaned dataset, for future model training:

[4] HuggingFace Transformers - https://github.com/huggingface/transformers

- Converting to lowercase text;
- Standardizing Vietnamese typing method;
- Vietnamese Word segmentation (Using vncorenlp[5]);
- Removing special characters and spaces;
- Removing stopword[6].

After data cleaning, we apply two traditional machine learning models and two deep learning models to evaluate and analyze models' performance on ViON. For each model, we divided the data into three sets: training, valid, and testing with the ratio of 8:1:1.

5.2 Experimenal configuration

All methods are compared on the same distributed dataset, and results are reported against the test set. Here, we refine the parameters to train all four models: MultinomialNB, Logistic Regression, Text-CNN, LSTM and Transformers Model - PhoBERT to compare and evaluate the multi-dimensional performance of the above models.

- **Multinomial Naive Bayes**: We use MutinomialNB with alpha = 1.0.
- **Logistic Regression (LR)**: Set the Logistic Regression model parameters with C = 1.0, solver = "lbfgs", multi_class = "auto" và max_iter = 10000.
- **Text-CNN**: We build a Text-CNN model with an embedded layer and 4 convolution layer. The word representation class has a size of 300, and the maximum number of objects is 40,000. Each convolution layer has 32 filters, and the filter sizes are 1, 2, 3, and 5. The active transition layer is ELU (Exponential Linear Unit). The output layer is Dense 13 with a softmax activation function, corresponding to 13 layers.
- **Long Short-Term Memory (LSTM)** This model has an embedding layer with an embedding size of 300, and the Bidirectional RNN layer is LSTM with 50 units. The output layer is Dense 13 with a softmax activation function, corresponding to 13 classes.
- **Transformers Model - PhoBERT** Using pre-trained to represent input data with custom parameter values such as 20 header length maximum, input string length 32.

5.3 Results and discussion

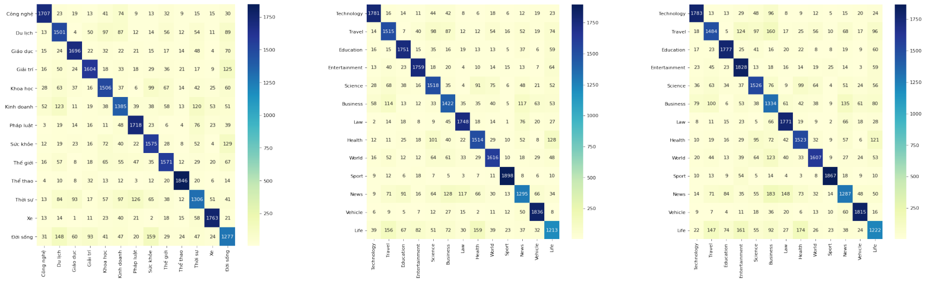
In this section, we will present the test results achieved with the parties updating the model topic. Classification of the results of the model classification is evaluated based on the F1-score scale. Table 3 below describes the accuracy of 4 models on the ViON dataset. Logistic Regression and PhoBERT models give the highest results among machine learning deep learning and transfer learning models that we tested with F1-score of 0.7867 (LR), 0.8026 (Text-CNN) and 0.8062 (PhoBERT).

[5] VnCoreNLP - https://github.com/vncorenlp/VnCoreNLP

[6] Vietnamese stopword - https://github.com/stopwords/vietnamese-stopwords

Table 3: Evaluation results on the ViON dataset. * indicates that the F1-score performances are measured on the ViON dataset with the pre-processing techniques.

Model		F1-score	F1-score*
Machine learning	MultinomialNB	0.7561	0.7816
	Logistic Regression	**0.7610**	**0.7867**
Deep learning	Text-CNN + W2V	0.7705	0.8007
	Text-CNN + FastText	**0.7730**	**0.8026**
	Text-CNN + ELMO	0.7702	0.8002
	LSTM + W2V	0.7793	0.7987
	LSTM + FastText	0.7712	0.7916
Transfer learning	**PhoBERT**	**0.7728**	**0.8062**



(a) Logistic Regression.

(b) Text-CNN.

(c) PhoBERT.

Fig. 2: Confusion matrix of experimental results on ViON dataset.

From the results shown in Figure 2a, 2b and 2c, we compare the prediction ability of our machine learning, deep learning and transfer learning models:

- The prediction results of the labels **Sport, Vehicles, Technology** have relatively good stability on all four models. Because these domains' content is highly specialized, it is less likely to be confused with other topics.
- The prediction result of the label **Life** was much wrong on all four models, with the mispredicted data mostly falling on labels **Health, Travel, Entertainment**.
- The prediction result of the label **News** was much wrong on all four models, with the mispredicted data mostly falling on labels **Business, Law, Education**
- The titles are mispredicted mainly because its content contains keywords related to other topics, confusing the training process.

In general, deep learning models obtain better results than machine learning models. On the machine learning models, Logistic Regression gives the best F1-score, but still hurt in some labels, giving less predictive results than MultinomialNB (Health, Life, Law). In deep learning models, TextCNN is a deep learning algorithm that achieves the best results. PhoBERT archieves the best results when making predictions on the ViON dataset. This can be explained because the PhoBERT model is a pre-trained model trained on the same formal data as the ViON set, so it has better classification performance.

PhoBERT is capable of performing parallel computations for words, reducing vanishing gradients, and helping the model learn better. Transformer model encoders are a form of feedforward neural nets, consisting of many other encoder layers, each of which encoder layer processes the words simultaneously.

In short, the Transformer model - PhoBERT is a good model for subject classification for Vietnamese news text.

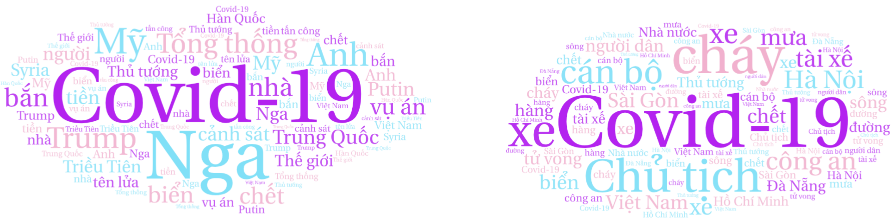
5.4 Error analysis

There are still some topics that are misclassified in the dataset due to the similarity between the contents. The table 4 shows examples of classification errors. From the table above, we can see that many misclassified titles are influenced by decision keywords, such as under the topic "Life" but was misclassified as the "Travel" topic due to the keyword "hotel".

Table 4: Several examples of classification error on ViON dataset.

Title	**Label**	**Predict**
Rao trên mạng đòi giết và ăn thịt người, đối mặt án chung thân (**English:** Classified on the Internet demanding to kill and eat people, face life sentence)	World	Law
Cho trẻ ăn gì để tăng đề kháng, hạn chế ốm vặt khi giao mùa? (**English:** What to feed children to increase resistance and limit disease during the season?)	Life	Health
Vì sao giường ngủ nên đặt 4 gối như phòng khách sạn (**English:** Why should the bed have 4 pillows like a hotel room?)	Life	Travel
Đề nghị chuyển trả lại 6.338 tỷ đồng vốn đầu tư công (**English:** It is proposed to refunds VND 6,338 billion of public investment capital)	News	Business
Phi vụ ’dùng xác chết thế thân’ của bí thư xã được toan tính như thế nào (**English:** How the communal secretary’s ’patsy’ mission was staged	News	Law

After looking at the dataset depicted in Figure 3, we noticed that there are many headers under the "World" label and the "Law" label containing keywords. same as: "Covid-19", "Prime Minister", "died". Furthermore, the titles in the "World" and "News" label not only contain the same keywords but also keywords like: "case", "police", "dead". These are potential keywords to predict the "Law" label titles. Our future studies will pay more attention to confusion cases, these decisive keywords to improve classification performance.



(a) World.

(b) News.

Fig. 3: Wordcloud for the keywords that appear most commonly on several labels.

6 Conclusion and future work

In this paper, we have present the process of building an open-domain, large-scale and high-quality dataset of 260,000 newsletter titles annotated with multiple labels for evaluating Vietnamese text classification. Then, apply traditional machine learning, deep learning, and transfer learning techniques to solve the classification of newspapers titles. We obtain positive results on all models experimented. With the prediction results of 4 models: MultinomialNB, Logistic Regression, Text-CNN, PhoBERT, evaluated by F1-score, the results are 0.7816, 0.7867, 0.8026, and 0.8062 on the test set, respectively. The results obtained a significant improvement (increased by more than 2% F1-score) after using the treatment of interference cases, using the dedicated Label Test Guideline. That shows that the dataset is built relatively well, suitable for training models to apply to solve the problem of text classification in general, automatically recognize the newsletter title, classify it into a suitable topic in particular, and save a lot of time and effort methods.

For natural language processing datasets, to create a useful dataset used to develop text classification models, data collection and pre-processing are essential. It is the decisive factor for the processing performance of the model when applied in practice. Be learned a good dataset, the model's ability to solve problems will be enhanced, making the model more accurate and reliable for prediction.

In the future, we plan to develop this dataset further by increasing the size, adding, filtering content, keywords related to the topic of the article; building combination models to increase accuracy and efficiency in classification.

References

1. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
2. Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.
3. Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen, and Hung Quoc Ngo. A comparative study on vietnamese text classification methods. In *2007 IEEE International Conference on Research, Innovation and Vision for the Future*, pages 267–273. IEEE, 2007.
4. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

5. David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
6. Sang-Bum Kim, Hae-Chang Rim, Dongsuk Yook, and Heui-Seok Lim. Effective methods for improving naive bayes text classifiers. In *Pacific rim international conference on artificial intelligence*, pages 414–423. Springer, 2002.
7. Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
8. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
9. Larry Medsker and Lakhmi C Jain. *Recurrent neural networks: design and applications*. CRC press, 1999.
10. Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*, 2020.
11. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
12. Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.
13. Toan Pham Van and Ta Minh Thanh. Vietnamese news classification based on bow with keywords extraction and neural network. In *2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)*, pages 43–48. IEEE, 2017.
14. Xuan-Son Vu, Thanh Vu, Son N Tran, and Lili Jiang. Etnlp: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. *arXiv preprint arXiv:1903.04433*, 2019.