Stock article title sentiment-based classification using PhoBERT*

Nguyen Son Tung[0000−0001−9244−7093], Nguyen Ngoc Long[0000−0002−6979−4473], Trang Tran[0000−0003−3370−6272], Nguyen Thu Thao[0000−0002−4026−7362], Duong T. Thu Phuong[0000−0002−4267−7162], and Tuan Nguyen**[0000−0002−3616−5267]

National Economics University, Hanoi, Vietnam
{209tungns,ngoclong1282001,huyentrang201ciel,thaothu2742001,
duongthithuphuong26122001}@gmail.com
nttuan@neu.edu.vn

Abstract. Text classification is a typical and important part of supervised learning, it has several applications in economics and attracted the attention of many stock market investors. For a long time, the news is frequently an unanticipated stock investment variable that instantaneously influences stock price directions. In front of an enormous volume of news, investors are always searching for models that automatically categorize news quickly and accurately. Thus, in this research, we have utilized different models like PhoBERT, SVM, Logistic Regression, LSTM, Random Forest, and Naive Bayes to classify news articles into three categories [negative, neutral, or positive] based on their titles. The results demonstrated that after training with a dataset of over 1000 news samples from CafeF.vn, the PhoBERT model outperformed other models with an accuracy up to 93%. The code and dataset is available at https://github.com/209sontung/Vietnamese-stock-article-classification.

Keywords: classification · PhoBERT · sentiment analysis · stock articles

1 Introduction

Text classification is a traditional word processing problem using machine learning. The first idea is to map a text to a known topic from a finite set of topics based on the semantics of the text. Text documents are typically used for classification, which is done based on selected documents and features. However, the

 In: N. D. Vo, O.-J. Lee, K.-H. N. Bui, H. G. Lim, H.-J. Jeon, P.-M. Nguyen, B. Q. Tuyen, J.-T. Kim, J. J. Jung, T. A. Vo (eds.): Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021), Da Nang, Viet Nam, 28-October-2021, published at http://ceur-ws.org

** Corresponding author.

classes are chosen before the experiment analysis, which is referred to as supervised machine learning operations. Due to the growing number of documents, the demand for text classification is expanding, and the tasks are getting increasingly diverse, such as sentiment analysis of reviews and news categorization, so on. An article in a newspaper, for example, could fall under one (or more) of these categories (such as sports, health, information technology, etc.). The use of spam filtering in email [1] [2], web services [3][4], fake currency identification [5], fake news identification [6], and opinion mining techniques [7] are also prominent and important applications in this field. Automatically categorizing text into a certain topic makes it easier to organize, store, and query documents later.

Forecasting is always a difficult task in the stock market because it is highly volatile and dynamic. Many methods have been proposed to forecast the future direction of the stock market. Financial news, for instance, has an immediate positive or negative impact on stock prices. Before purchasing a stock, investors, for instance, evaluate a company based on its activities on its official website and financial news about the company. However, investors can not fully assess such vast amounts of financial news data on their own. As a result, investors require a model that can assist them in quickly sorting through financial news articles.

In this research, we collected a dataset of over 1000 financial news articles about the stock market from the website CafeF.vn. Afterward, we used LSTM [8], PhoBert[9], SVM[10], and other models to categorize news articles from the above dataset into three categories [positive, negative, and neutral]. The PhoBERT model provided outstanding accuracy results of up to 93% after training.

The remainder of this paper is organized as follows. Section 2 introduces related works . The proposed model is introduced in Section 3. Section 4 discusses the results of experiments and is followed by a conclusion in Section 5.

2 Related works

Sentiment analysis, also referred to as a classification task, aims to forecast the overall sentiment of a text, which could be a tweet or a review of a movie or product. The main goal is to determine if the text's conveyed impression is positive or negative, in some cases, with a score or confidence metric.

In English, a lot of publications on sentiment analysis have been undertaken. For the problem of sentiment classification, Pang et al compared multiple supervised learning algorithms [11], including Naive Bayes [12], KNN [13], Maximum Entropy Models [14], and Support Vector Machines. They tested several types of features and achieved the highest accuracy of 82.9% on a corpus of movie reviews. Zhou utilized the Stanford Sentiment Treebank (SST) dataset [15] to describe the sentiment categorization of movie reviews. In comparison to multilayered CNN [16] and RNN [17] models, the architecture combining CNN and LSTM models produced better performance. The two classes (positive and negative) dataset had an accuracy of 87.7%, whereas the five classes (very positive, positive, neutral, very negative, negative) dataset had a 49.2%. In another sen-

timent analysis study performed on the SST dataset, Manish Munikar et al. [18] applied BERT [19] - the latest state-of-the-art in the NLP [20] field proposed by Google in 2018. The architecture contained a dropout regularization and softmax classifier layers on top of the pre-trained BERT layer. Their proposed model was presented that achieved the highest of 94.7% correct in SST-2 and 84.2% when performed in SST-5, surpassing every aforementioned technique.

In Vietnamese, Kieu and Pham [21] performed studies on a corpus of computer product reviews by offering a rule-based system for sentiment classification in Vietnam utilizing the GATE framework [22]. This approach reached 67.35% precision overall, but designing the rules seems to be a difficult and time-consuming task. Quan et al. [23] presented a multi-channel LSTM - CNN model for Vietnamese sentiment analysis that combines Long Short-Term Memory (LSTM) and CNN. This combination had an accuracy of 87.72% on the VS dataset and 59.61% on VLSP, which was also proposed in their research.

In this paper, we proposed a Vietnamese stock news sentiment classification model, which is a novel approach to sentiment analysis in Vietnamese. The proposed model achieved 93.12% accuracy and was constructed using a pre-trained PhoBERT, a state-of-the-art language model for Vietnamese based on BERT architecture. In addition, we built a dataset that included 1000 titles of financial articles taken from CafeF.vn and labeled them into three groups [negative, neutral, or positive].

3 Proposed Model

Our proposed model consists of 2 stages described in Fig. 1. The input is the financial article's headlines, which will proceed through the first stage to preprocess the data to convert it into a format that the PhoBERT model can understand and improve its accuracy. Following that, in the second stage, our PhoBERT-based model will be tasked with assessing content from the header broadcast and categorizing it into one of three classes represented as -1, 0, or 1 (i.e., -1 as negative, 0 as neutral, and 1 as positive direction).

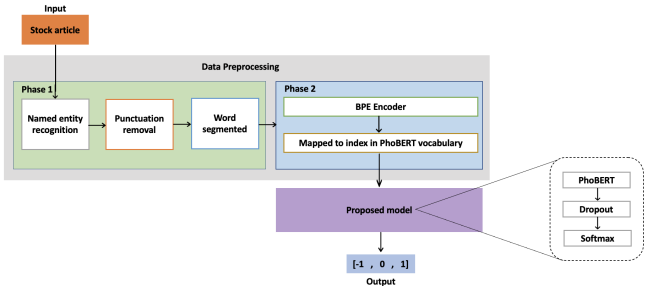

Fig. 1. Proposed model

3.1 Preprocessing

The preprocessing procedure was separated into two phases. In Phase 1, first, we applied VnCoreNLP's Named entity recognition [24] to extract all the proper nouns and replace those words that signify location with the word "loc" or "name" for the organization name, stock code, or person's name. To avoid any confusion when the model predicts, the punctuation was then removed, hence increasing the model's accuracy. Considering the fact that white space is also utilized to separate syllables that make up words in Vietnamese, in the last step of Phase 1, we adopted Rdrsegmenter from VnCoreNLP to separate words for input data. Furthermore, as an input for the PhoBERT model the title needed to be tokenized, therefore we utilized BPE tokenizer [25].

In Phase 2, we had the symbol vocabulary with the character vocabulary, and each word was represented as a sequence of characters with a unique end-of-word symbol " $< /s >$ " that allowed us to recover the original tokenization after translation. In example, we counted all symbol pairs iteratively and replaced each occurrence of the most common pair ("A", "B") with the new symbol "AB". Each merge process generates a new symbol that represents an n-gram of characters. BPE does not require a shortlist because frequently occurring character n-grams (or complete words) are finally combined into a single symbol. Thus, the amount of the final symbol vocabulary is equal to the original vocabulary.

Then we mapped each subword to its corresponding ID in the PhoBERT vocabulary, and because each title is varied in length, we employed pad sequences to match them all in length. i.e. sentences that shorter than 125 subwords are padded with 0 at the end, while longer are trimmed to produce 125.

3.2 Training details

First, we installed all the necessary materials included transformers library, VnCoreNLP Python wrapper and its word segmentation component (i.e. RDRSegmenter), then fastBPE to convert the input text into a list of subwords. After train and test dataset are done prepared as described in Section 3.1, DataLoader is created to load data into the model. Then we loaded $PhoBERT_{BASE}$ from HuggingFace's transformers library as the pre-trained model. The optimizer we chose for the training stage is AdamW optimizer, which is an improved version of Adam optimizer from the transformers library. In addition, this study used batch size = 32 with 10 epochs divided into two stages with different learning rate. The initially learning rate which we utilized was $\alpha = 5e - 6$ in the first 5 epochs in order for the loss to converge faster. Once the first training phase was completed, we reduced the learning rate to $\alpha = 5e - 7$ to achieved the smallest potential loss. The model was then trained for another 5 epochs since the loss on the validation dataset appears to stabilize after this number of cycles.

Finally, the softmax classification layer (which includes three nodes corresponding to three classes in the dataset) will output the probabilities of the input text belonging to each of the class labels, with the total of the probabilities equal to 1. The dense layer consists of a fully connected neural network with

the softmax activation function. The softmax function $\sigma : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is given in (1).

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad for\ \ i = 1, ..., K \tag{1}$$

where $\mathbf{z} = (z_1,...,z_K) \in \mathbb{R}^K$ is the softmax layer's intermediate output (also called logits). The predicted label for the input is then chosen from the output node with the highest likelihood. The output of the proposed model will be represented as -1, 0, or 1. The entire training process was deployed on PyTorch [26] framework.

3.3 Method

BERT. BERT stands for Bidirectional Encoder Representations from Transformer. It is a transformer-based [27] machine learning technique developed by Google for pre-training in natural language processing (NLP). In 2018, Jacob Devlin and his colleagues created and published BERT. BERT includes two original models in English. The first model is $BERT_{BASE}$: It consists of 12 encoders with 12 bidirectional self-attention heads. The second model is $BERT_{LARGE}$: 24 encoders with 16 bidirectional self-attention heads. BERT is designed for pre-training from unlabeled texts with 800 million words by BooksCorpus and 2,500 million words by Wikipedia.
BERT has performed more than 10 natural language processing tasks with good results. It has improved the GLUE benchmark to 80.5%, pushed MultiNLI accuracy to 86.7%, absolute 5.1 point improvement in SQuAD v2.0 Test F1, etc. With L is the number of sub-layer blocks in the transformer, H: the size of the embedding vector, A: the number of heads in the multi-head layer, model BERT has two architectures as follows:

- $BERT_{BASE}$(L=12, H=768, A=12): Total parameters are 110 million.
- $BERT_{LARGE}$(L=24, H=1024, A=16): Total parameters are 340 million.

PhoBERT. The BERT model's release marked a watershed moment in the NLP industry. Following the public release of the BERT model, a slew of open-source BERT training programs have sprung up. There are also numerous unilingual and multilingual BERT pre-train models that are commonly used. Since then, PhoBERT has been particularly trained for Vietnamese and released by VinAI Research in March 2020.
PhoBERT is based on the design and approach of RoBERTa [28], which was introduced by Facebook in 2019 and is an improvement over the original BERT. PhoBERT was trained from about 20GB of data, including approximately 1GB of the Vietnamese Wikipedia Corpus and 19GB remaining from the Vietnamese News Corpus. This type of data is also ideal for training a model like BERT.
PhoBERT, similar to BERT, is available in two versions. The first version is $PhoBERT_{BASE}$ with 12 transformer blocks and the second version with 24 transformer blocks is named $PhoBERT_{LARGE}$.

VnCoreNLP. VnCoreNLP (A Vietnamese Natural Language Processing Toolkit) is a Java Natural Language Processing toolkit designed to aid NLP research in Vietnam. Through essential NLP components such as word segmentation, POS tagging, and NER, VnCoreNLP provides extensive linguistic annotations. In the course of NLP research, the Vietnamese standard dataset was published. In early 2013, the first VLSP evaluation campaign used datasets for word segmentation and POS tagging. In 2014, a high-quality dependency treebank was published, and a NER dataset was published for the 2016 VLSP review campaign. The architectural system design is depicted in Fig. 2.

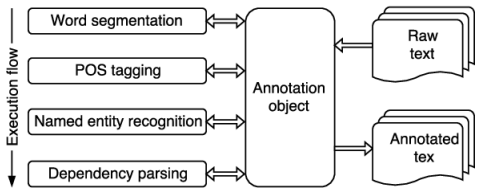

Fig. 2. In pipeline architecture of VnCoreNLP

4 Experimental results

4.1 Dataset

To be able to use PhoBERT to evaluate and categorize the news' impact, we provided a dataset that included 1000 titles of financial articles taken from CafeF.vn and labeled them into three groups [negative, neutral, or positive] with the help of experts. The dataset contains 187 articles having a negative impact, 248 articles with no impact, and 565 articles with a positive impact. After that, we divided the dataset into three sets, 80% for training, 10% for validation and 10% for testing. The training set was used to train the model, validation set was utilized to tune the hyper-parameter. Finally, the result of model was evaluated on testing set. The examples of our dataset are shown in Table 1.

4.2 Result

After experimenting with 6 different models on the same dataset using various preprocessing techniques in order to achieve the best possible results, Table 2 was obtained.

As seen in the table above, our model, PhoBERT, outperformed other popular and sophisticated NLP models with 93.12% accuracy and was 10.54% higher than the second-highest approach using Logistics Regression. Our model also achieved the best performances in other metrics such as precision, recall and F1 score.

Table 1. Dataset examples

Label	Titles	Titles (English)
Positive (1)	Vĩnh Hoàn (VHC): Doanh thu tháng 4/2021 đạt 800 tỷ đồng, các thị trường xuất khẩu đồng loạt tăng tốt	Vinh Hoan (VHC): April 4/2021 revenue reached VND 800 billion, export markets simultaneously increased well
Neutral (0)	Lịch sự kiện và tin vắn chứng khoán ngày 17/5	Calendar of events and short stocks news on May 17
Negative (-1)	Khối ngoại tiếp tục bán ròng gần 630 tỷ đồng trong phiên 18/5	Foreign investors continued to net sell nearly VND 630 billion in May 18

Table 2. Experimental result (%) of our models compared to other models

Model	Accuracy	Precision	Recall	F1 score
Logistics Regression	82.58	82.83	78.64	80.00
SVM	80.09	80.39	76.66	77.91
Naive Bayes	79.6	80.36	76.00	77.22
Random Forest	80.10	83.64	74.25	77.04
LSTM	75.56	72.67	73.42	72.22
PhoBERT	**93.12**	**90.97**	**92.64**	**90.63**

5 Conclusion

In this research, sentiment classification was performed on Vietnamese stock articles collected from the site CafeF.vn. The whole dataset consists of 1000 titles divided into positive, neutral, and negative news. Because of the fact that white space is also utilized to separate syllables that make up words in Vietnamese, we utilize Rdrsegmenter from VnCoreNLP (a word splitting library published by the author of PhoBERT) to separate words for input data. Then, using the BPE Encoder, convert text into a list of subwords, then map each subword to its ID in the PhoBERT vocabulary. The proposed model in this study employed a state-of-the-art language model for Vietnamese named PhoBERT, and the entire training process has been deployed on PyTorch. As a result, our approach achieved 93.12% accuracy, surpassing other popular and sophisticated NLP models when performed on the same dataset.

As previously stated, financial news can directly impact stock prices, so our proposed model could be used to assist with stock price forecasting problems using machine learning. In future work, we also want to explore the effect of using word embeddings on sentiment classification. Furthermore, we aim to investigate multiclass categorization of news data using various Deep Learning models, which will comprise classes such as the economy, sports, health, and technology. Also, we intend to extend sentiment classification to more domains by crawling data from numerous Vietnamese websites, such as product reviews, hotel reviews, and book reviews.

References

1. Bhowmick, A., Hazarika, S.: Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends. ArXiv. (2016) https://doi.org/10.1007/978-981-10-4765-7_61
2. J, R.K., G, M., P, S.: Email Spam Detection using Machine Learning Techniques. IARJSET. 8, 189–193 (2020). https://doi.org/10.17148/iarjset.2021.8632.
3. Shafi, S., Qamar, U.: [WiP] Web Services Classification Using an Improved Text Mining Technique, 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA). 210-215 (2018) https://doi.org/10.1109/SOCA.2018.00037.
4. Crasso, M., Zunino, A., Campo, M.: AWSC: An approach to Web service classification based on machine learning techniques. INTELIGENCIA ARTIFICIAL. 12, (2008). https://doi.org/10.4114/ia.v12i37.955.
5. P Gayathri: Texture Classification for Fake Indian Currency Detection. International Journal of Engineering Research and. V9, (2020). https://doi.org/10.17577/ijertv9is060211.
6. Jain, A., Kasbe, A.: Fake News Detection, 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). 1-5 (2018) https://doi.org/10.1109/SCEECS.2018.8546944.
7. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2, 1–135 (2008).
8. Staudemeyer, R., Morris, E.: -Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks. (2019).
9. Nguyen, D., Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese. Association for Computational Linguistics (2020).
10. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf: Support vector machines, in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28 (1998), doi: 10.1109/5254.708428.
11. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2, 1–135 (2008).
12. Zhang, H., Li, D.: Naïve Bayes Text Classifier, 2007 IEEE International Conference on Granular Computing (GRC 2007). 708-708 (2007). https://doi.org/10.1109/GrC.2007.40.
13. Cunningham, P., Delany, S.J.: k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). arXiv:2004.04523 [cs, stat]. (2020).
14. Ziebart, B., Maas, A., Bagnell, J., Dey, A.: Maximum Entropy Inverse Reinforcement Learning. (2008)
15. Zhang, L., Wang, S., Liu, B.: Deep Learning for Sentiment Analysis : A Survey. arXiv:1801.07883 [cs, stat]. (2018).
16. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. Pattern Recognition. 77, 354–377 (2018). https://doi.org/10.1016/j.patcog.2017.10.013.
17. Sherstinsky, A.: Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena. 404, 132306 (2020). https://doi.org/10.1016/j.physd.2019.132306.
18. Munikar, M., Shakya, S., Shrestha, A.: Fine-grained Sentiment Classification using BERT. (2019)
19. Devlin, J., Chang, M.-W., Lee, K., Google, K., Language, A.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019).

20. Gelbukh, A.: Natural language processing. Fifth International Conference on Hybrid Intelligent Systems (HIS'05). (2005) https://doi.org/10.1109/ICHIS.2005.79.
21. Kieu, B.T., Pham, S.B.: Sentiment Analysis for Vietnamese. 2010 Second International Conference on Knowledge and Systems Engineering. (2010)
22. Huynh, T., Hoang, K.: GATE framework based metadata extraction from scientific papers. 2010 International Conference on Education and Management Technology. (2010). https://doi.org/10.1109/ICEMT.2010.5657675.
23. Vo, Q.-H., Nguyen, H.-T., Le, B., Nguyen, M.-L.: Multi-channel LSTM-CNN model for Vietnamese sentiment analysis, 2017 9th International Conference on Knowledge and Systems Engineering (KSE). 24-29. (2017) https://doi.org/10.1109/KSE.2017.8119429.
24. Vu, T., Quoc Nguyen, D., Nguyen, D., Dras, M., Johnson, M.: VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. (2018).
25. Wang, C., Cho, K., Gu, J.: Neural Machine Translation with Byte-Level Subwords. Proceedings of the AAAI Conference on Artificial Intelligence. 34, 9154–9160 (2020). https://doi.org/10.1609/aaai.v34i05.6451.
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., Facebook, L.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. (2019).
27. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in Transformer.
28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., Allen, P.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019).