Using Phobert model for detecting fake news in title of newpapers written in Vietnamese*

1st Anh-Khoa Nguyen
Faculty of Information and Technology
University of Science, VNU-HCM
Ho Chi Minh city, VietNam
0000-0002-0664-8911

2nd Minh-Tri Le
Faculty of Information and Technology
University of Science, VNU-HCM
Ho Chi Minh city, VietNam
20120600@student.hcmus.edu.vn

3rd Trung-Hieu Do
Faculty of Information and Technology
University of Science, VNU-HCM
Ho Chi Minh City, VietNam
20120007@student.hcmus.edu.vn

4th Xuan-Nam Cao
University of Science
University of Science, VNU-HCM
Ho Chi Minh City, VietNam
cxnam@fit.hcmus.edu.vn

5th Minh-Triet Tran
Faculty of Information and Technology)
University of Science, VNU-HCM
Ho Chi Minh City, VietNam
tmtriet@fit.hcmus.edu.vn

6th Hai-Quan Vu
Faculty of Information and Technology
University of Science, VNU-HCM
Ho Chi Minh City, VietNam
vhquan@vnuhcm.edu.vn

***Abstract*—The more the world develops, fake news has become the smoking of our generation. It can make disrupting public order and affect emotional and physical health, it also may be a cause of distortion social. Fake news detection become a real problem. There are many ways to approach this problem: SVM, CNN, ANN,... Many people used to do experiments with many deep learning models: BERT, deBERT, XLNET,. . . in many datasets in English or some other languages. It is an undeniable fact that all model BERT can get a good score and high accuracy. Unfortunately, most models like this do not support Vietnamese, when we force them into this problem, it can get a low accuracy because we need to process to help a word in Vietnamese become a word in English, It will be making a word become more and more unclear. In recent years, VinAI create a new mode called phoBERT("phở" Bert) it works too well in Vietnamese although complicated in Vietnamese. Two factors that support this are nlpVNcore and RoBert model. Finally, when our team use this model to train and predict labels for titles, it brings a good result of more than 95%**

I. Introduction

Fake news detection is a common problem that appears everywhere on The Internet [1]–[3]. You can see fake news from many social network accounts, fanpage more than millions like,... Most of them are often used to get more like, comment, and share, attracting everybody. But fake news not only appears on the social network but also appears in many newspapers - They may even gain credibility. Many online newspapers build strong convictions in people, then they usually post fake news and people cannot detect it. Fake news in newspapers is caused significantly more than fake news in social networks because residents believe them so much. It can affect on physical and mental heal of a person or a group [4]–[11]. For example: In Covid-19, there is much fake news about when people eat eggs that can prevent and help the victim of this disease be stronger [12]. It may be because of social distortion: demonstrate, riot, or even a war. Many online newspapers come from reactionaries appear on the internet, the main purpose is to drag residents and attack peace and the highest community of a country. [13], [14]

It is a fact that detecting fake news is not an easy mission, even for a human. You can see in the real world, that many people were affected by fake news [15]. Fake news Many people can ask our team: if you cannot detect it by your mind, how do you teach it to a computer? It is not a new problem, many people used to solve it. Because computer can remember word and context of them in a sentences. There are many different context between fake news and real news [16]. Many individuals say that fake news is a hot trend among researchers in natural language processing. There are many ways to approach it: SVM, BiLSTM,... can detect fake news, but the accuracy of them is just around 88-90%. After that, Bert appear and it is a new way to detect fake news, it can detect them and easier to get more than 90% accuracy.

Vietnamese is a special language, a word in Vietnamese include more than 2 unit word(we will call it complex-word). Complex-word is has two groups: some complex-word when you separate it into many unit word tokens, they have the same meaning, but for some complex-word when we separate them, the token of them do not have a clear meaning [17]. So, bert is unsuitable to train and predict the label for titles. Our group used to think that we can make it become an English word, but it is so hard for bert can understand and train it when Bert is a model work depending on context. For example, when we delete the symbol of a word, it can make two different words ("bán" -sell and "bạn" - a friend will become "ban"). Our team used to use Bert and many models relative to bert for training and prediction, it was just about 0.4. It means that the accuracy of this experiment is lower than the random choice of humans. After that, we find a new model from vinAI - phoBert("pho" from "phở" - traditional food of VietNam) [18] when there are many special models for other languages: Cul et al, De vries et al, ... It uses a corpus from two sources: Wikipedia and Facebook. But we need how to split a sentence into complex words, luckily, vncore is a tool that helps to do it easier

In the next part of this paper, we will know some related project and many details about our project

II. Related work

There are many machine learning help to classifier data, in this case, we need the classifier to fake title and real title. Support vector machine is a good way to approach, it also deletes stop words, and special character lemmatizes work and Bi-gram to the covert matrix. Lately year, many projects use machine learning to detect fake news. For example, a paper from Su et al used to use machine learning to detect fake news and got a good result. [19]. In the Ankit Kesarwani and partner used to use K-nearest neighbor to detect fake news, [20], it got nearly 80 percent accuracy. Shalini Pandey1 [21] et al also write a paper, they use machine learning classifiers including KNN, SVM, Naive Bayes, Decision tree, and Natural language processing. It must be pre-processed by removing stop words and stemming, converting the word to vector, and visual using TSNE. In their experiment, machine learning classifiers can get an accuracy of nearly 90%. The highest result is from Logistic Regression. The same result from the paper of Fathima Nada [22] and team, they use bag-of-words, n-gram, and TF-IDF to feature selection, then they encode sentences, finally, they classifier them by logistic regression.
Besides traditional machine learning, many researchers have another way to approach the problem. A common way this is use Bi-directional LSTM-RNN , for example: project Pritika Bahad et al [23]. There are many steps from text to result, they must process, and tokenize. Traning and Validation will continue the process: train model -> test with trained data -> Tune hyperparameters using validation data and try to predict for test data. In this paper, accuracy in tests is set at about 91 percent.
Lately year, a new model appear and get an amazing result, this is Bert. It can solve many tasks relative to fake news detection, beside Bert, many other models appear based on the construction of bert. In a nearly paper by Sourya Dipta Das and partner, they used many models relative to bert: deBERT, roBERTa, XLM-Roberta to solve the problem, main purpose of them this is to find fake news on Twitter about covid - 19, the accuracy of them nearly 99 percent, it means that rate of error in their experiment is so low.

Related task:

In other countries, they also try to use a model to detect fake news in their language. In Brazilian Portuguese Language [24], Pat et al researched to find the best way to Analysing and Classifying COVID-19 Fake News, they use four supervised machine learning techniques: SVM, Random Forest, Naive Bayes, and gradient boosting, then they use four deep learning models: LSTM, Bi-LSTM, GRU and Bi-GRU. Pakpoom and Lawakoorn [25] also do the same tasks in the Thai social test, they use ULMFiT, Bert, and GPT to solve this problem. Most paper research in other languages needs a step to preprocess for languages. In the next part, our teams will present detail about our research.

III. Methodology

This section will talk about how to solve this problem. We have approached this task in many ways, but they have the same point as a text classification problem. The experiments were performed on a system with nearly 13GB RAM and 2.2 GHz, Intel(R) Xeon(R) CPU @2.30GHz, along with a Tesla T4 GPU, with a batch size of 32. The maximum input sequence length was fixed at 125
The main task for this problem: We must label a new title into two categories: fake or real. It includes many parts from how to get data to get results.

A. Dataset:

Scrapy [26] is a high-level web crawling and web scraping framework used to crawl data or extract data from the web. It works in python.
Our Group uses a special tool called scrapy to crawl data from many websites. We got fake news data from newspapers written by reactionaries on the internet, its attack figure of Communist Party [27], [28], it also was illegally of some countries. About the real title, we get it from a credible newspaper in Vietnam. Such as: Thanhnien, vtv,.... We choose the main topic of the title about world conditions. Most of the real data we crawl from vietnamnet.vn - a prestige newspaper in Vietnam on the internet.
After that, raw data will be processed before using, it includes many steps: filter trash new, delete stop word, delete accent.

B. pre-process for data

- After crawling data, there are many trash titles. It means that it does not have clear content. For example List of beautiful beaches in the world(It is not a statement, we cannot estimate true or false). All similar titles are removed.
- Stop words are basically a set of commonly used words in any language, they are have not an important role in the content of the title. So, we need to remove them, we use a stopword list [29] from https://github.com/stopwords/vietnamese-stopwords
- Vietnamese is a language that has many complex words. it maybe includes two or more unit words, such as:"đi chợ", "rau muống", "lá bạc hà" so we need a tool to detect complex words in sentences. The inventor of this model recommends we should use NLPvnCore [30] to find and separate complex words in sentences.
- Because in file csv, real titles and fake titles are in two areas, so we need to shuffle to merge them.
- All titles after the process are saved in .csv file, then will be used for training model or testing.

C. Encode

In this step, each title is split into tokens(word) being the training model. We use many ways to encode data depending kind of model because each model has different requirements. We choose the byte-pair encoding [31] to encode from the token. After that, we have a list with data encoding, We also attract labels of them. We have a list of sample data from train data. Algorithm of BPE - a simple compression data from Gage(1994). Instead of merge byte usually appear, we merge characters or group characters usually appear

- We operate a word to the character and add with special character to recognize the end of the word.
- In the next step, we use the agreed function, which will collect characters from a couple of characters has the highest frequency. We got a new character. We continue to collect all characters in a new group. For instance: First, the couple 'p' and 'his by far the most frequently occurring, we got 'ph', in the next time 'ph' and 'ở' is a couple of reasonability. we got 'phở'. When we finish this step, we don't have the character repeated, when we split a word in the first group, we have a token in the last group. Ex: 'phở bò' = 'phở' 'bò'. In encode process, a number is a symbol of a group character.

We create a mask (0,1), it will recognize the padding value. It means that it will mark in value more than zero

D. Base Bert architectures:

BERT [32](Bidirectional Encoder Representations from Transformers) is a Model to solve many problem relative with Natural language process, it proposed by Google Research Member in 2018. It can finish any task in the NLP field with high accuracy, It got an achievement in GLUE task, about 80%, an increase of more than 7.6 %. There are two kinds of BERT: base-Bert and large-Bert. The core of BERT includes Semi-Supervised Learning. Model after train will recognize the sample. The most thing in Bert is it can learn the context of two sides of a word. Before BERT, we can not find a model that can do it, most of them just learn context from one side like ELMo or non-context like word2vec, gLove,...

- Base-bert: 12 layers, 12 attention heads, and 110 million parameters
- Large-bert: 24 layers, 16 attention heads and, 340 million parameters

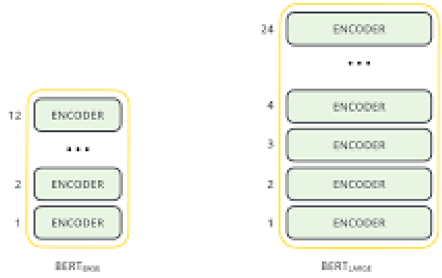

Figure 1. Bert (Source: rupeshgelal)

Bert-base has the same size batch with OpenAI GPT to compare accuracy between the two models.

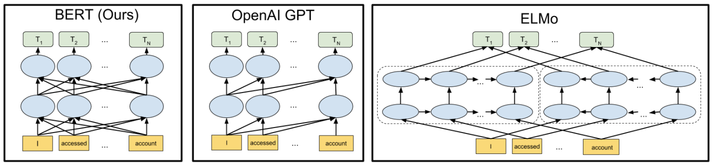

Figure 2. Bert compared (Source: ResearchGate)

A special thing in Bert is a user can easily fine-tunning or mask ML for it, for example, they can add an output layer to support training.

E. phoBERT:

There are two types of phoBERT [18], Base and Large. In this paper, we choose Base-phoBERT, which also corresponds to Base-BERT architectures. Following the composer of phoBERT, it uses roBERT for pre-train. In this period, they use two datasets from Wiki and Facebook to train phoBERT model. This is a different point between roBERT and phoBERT, this is phoBERT use faseBPE to segment a sentences

F. Optimize of phoBERT:

They employ the RoBerta implementation in fairseq. The maximum length of a sentence is 256 tokens. Besides, they also use Adam to improve model

- Base Bert: batch size of 1024 across 4 V100 GPUs (16GB each), peak learning rate 0.0004
- Large Bert: batch size of 512, peak learning rate of large bert is the same of base bert

In this project, we concentrate to fix the learning rate attribute in the Adam algorithm. This algorithm maintains a square of the slope, past vt, and average slope. It is like a heavy ball that has friction, so it can pass the local minimum and get a value of flat minium. This event can appear because Heavy ball with friction effect depend on coefficient $\frac{mt}{\sqrt{vt}}$. To update adam, we use this equation [33]:

$$g_n = \bigtriangledown f(\phi_n - 1)$$

$$m_n = \frac{\beta_1}{1-\beta_1^n} m_{n-1} + \frac{1-\beta_1}{1-\beta_1^n} g_n$$

$$v_n = \frac{\beta_2}{1-\beta_2^n} v_{n-1} + (\frac{1-\beta_2}{1-\beta_2^n}) g_n \odot g_n$$

$$\phi_n = \phi_{n-1} - \alpha \frac{m_n}{\sqrt{v_n} + \epsilon}$$

In our project, we concentrate to alpha, it means learning rate value

G. Use model:

We load mode phoBERT of VinAI, User has two modules to choose from: transformer or fairseq. In this paper, we use Transformer to approach the model. $BertSquenceClassification$ has input is list ids and mask, the output is losing value and figure of a probability distribution. Finally, we use Transformer module to load and train.

IV. Experiement:

A. Evaluation:

For evaluation, we use two parameters: F1 score and accuracy. An equation to know the value of F1, this is:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

- TP: number of true positives
- FP: number of fake positives
- FN: number of fake negatives

B. Experiement:

For evaluation, we use two parameters: F1 score and accuracy. In an equation to know the value of F1, The First Experiment, we train with 60% data crawl, 20% for Val dataset, and 20% for to test and evaluate the quality of the model, epochs = 10, learning rate=10^{-5}, we get a result in this table
Second, we change to epochs=15, learning rate = 10^{-5} value, then we change the learning rate of mode. We can guess the value of the learning rate by this is:

$$lr = lr(0.99)^{epochs\_num}$$

In this equation, epochs num is an epoch when train loss changed little, this equation prevents loss value past extreme locations and improves convergence of loss function.

C. Table Result

First, we run with 10 epochs, learning rate = 10^{-5}. We got a good result. We can see that, just with 10 epochs, accuracy in test about 93.36%

Table I
Performance with 10 epochs

Score	Dataset		
	Train	***Val***	***Test***
accuracy	0.9689	0.9358	0.9336
F1_score	0.9689	0.9314	0.9354

To get a better result, we decide change number of epochs to 15. We run model with learning rate change from 10^{-5} then we continue run with 8.6×10^{-6} and 7.3965×10^{-6}. We can easy to see that, model with learning rate = 8.6×10^{-6} get the highest accuracy and F1_score in all datasets. In 3 case, F1_score and accuracy in all dataset increase, but when we decrase learning from 8.6×10^{-6} to 7.3965×10^{-6}, accuracy and F1_score not increase. In this case, overfitting was appear

Table II
Performance with 15 Epochs

Score	Dataset			
	Learning rate	***Train***	***Val***	***Test***
F1-score	10^{-5}	0.9724	0.9398	0.9477
	8.6×10^{-6}	**0.9769**	**0.9461**	**0.9536**
	7.3965×10^{-6}	0.9719	0.9387	0.9484
accuracy	10^{-5}	0.9742	0.9435	0.9510
	8.6×10^{-6}	**0.9785**	**0.9500**	**0.9570**
	7.3965×10^{-6}	0.9741	0.9440	0.9530

D. Dicussion:

Fake news in a special problem, it depend on training dataset.If number of train value increase, accuracy will be improve. If you want to use it in real world, i think we need to extend train dataset, usually update from official website/newspaper. Prediction label for a title just a temporary, maybe a warning for resident when them acess in a website, then Anti-fake news bureau will labeling for this title. After that, this title will be add in dataset, this is an infinity loop...

E. Future work:

To get a higher score, I think we need to improve processing and tokenizing for sentences, many complex words is not appear when use vncore
Most papers nowadays, just detect fake news from a title but sometimes, it may appear in the main source of the newspaper. But it is not easy to find fake detail in a paragraph. In the future, I hope to find a way to find fake detail in a long paragraph. It is so difficult because to label for a newspaper need so much time.
The next time, maybe our team will be researched by using a special tool to get topic sentences or the main content of a newspaper, then will mix it with the result from the detected title. In this process, many things will appear in the content of newspaper: images, URL, emotion appear,... we need a pre-step to process it to text or a special code

V. Conclusion

In this paper, our team introduced a way to solve an important problem in real life, this is how to detect fake news in a list of titles. It has a special role in our world when people easier approach new information but they do not know about facts after this title. We used to try with many models to find the best way, finally, we use phoBERT from vinAI to finish this task. A model with many advantages: simple, strong, and many supporter tools,... The main point to help this model can easily recognize Vietnamese complex words. It is a model that corresponds to Base-BERT architecture. After collecting data from newspapers, filtering, removing stop words, using vncore to tokenize complex words and encoding it to a vector, we use phoBERT and improve it to get the highest result. There are many ways for this project in the future, our team will continue to research to get a good result for everybody. Finally, we got an accuracy of nearly 0.96, it is not a perfect result but it is a high score.

References

[1] Báo Quân Đội Nhân Dân Việt Nam, Cảnh giác với tin giả trên mạng xã hội, 13/08/2020

[2] Peter Dizikes, MIT News Office, March 8, 2018, Study: On Twitter, false news travels faster than true stories

[3] Vietnamexpress, 31/05/2022, Tin giả ’cảnh báo lừa đảo vaccine’ lan truyền mạng xã hội

[4] Minh Thu, Vietnamplus, 9/11/2021,Quản trị khủng hoảng thông tin trong bối cảnh đại dịch COVID-19 là yêu cầu cấp bách hiện nay bởi ’virus’ tin giả lây lan nhanh và nguy hiểm cũng không kém gì SARS-CoV-2

[5] Nhân Dân, 29-08-2017, Tin đồn, đám đông và những hệ lụy

[6] Nguyễn Quốc, Đại đoàn kết, 14/05/2022, Vụ cô gái bị cắt ghép clip phỏng vấn: Chủ Fanpage có thể bị truy cứu trách nhiệm hình sự?

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Karina Margit Erdelyi, The Psychological Impact of Information Warfare & Fake News

[9] Elianna Lev,13/06/2021,The Dangers of Disinformation and How It Impacts Your Mental Health

[10] cybersmile, CONCERNS GROW THAT FAKE NEWS AND MISINFORMATION ONLINE SURROUNDING COVID19 IS AFFECTING PEOPLES MENTAL HEALTH

[11] Faiz Siddiqui and Susan Svrluga, 5/12/2016, N.C. man told police he went to D.C. pizzeria with gun to investigate conspiracy theory

[12] Báo Công An Nhân Dân,30/03/2020, Bác tin đồn ăn trứng gà để chống dịch COVID-19

[13] KATHERINE OGNYANOVA, 2/6/2020, Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power

[14] VOV, 21/07/2021 Các thế lực thù địch đang lợi dụng dịch Covid-19 để chống phá Đảng, Nhà nước

[15] Mathilde Frot, 29/06/2017, Most People Can't Spot Fake News. Can You?

[16] Maria Temming, 26/06/2018, People are bad at spotting fake news. Can computer programs do better?

[17] Lieu Dang, 28/03/2021, Vietnamese Sentence Structure: Basic Vietnamese Grammar

[18] Dat Quoc Nguyen, Anh Tuan Nguyen, 5/10/2020, PhoBERT: Pre-trained language models for Vietnamese

[19] Z Khanam et al 2021, Fake News Detection Using Machine Learning Approaches

[20] Ankit Kesarwani, Sudakar Singh Chauhan, Anil Ramachandran Nair, Fake News Detection on Social Media using K-Nearest Neighbor Classifier

[21] Shalini Pandey et al 2022, Fake News Detection from Online media using Machine learning Classifiers

[22] Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed, FAKE NEWS DETECTION USING LOGISTIC REGRESSION

[23] Pritika Bahada, Preeti Saxenaa ,Raj Kamal, Fake News Detection using Bi-directional LSTM-Recurrent Neural Network, Volume 165, 2019, Pages 74-82, Procedia Computer Science

[24] Patricia Takako Endo ,* , Guto Leoni Santos , Maria Eduarda de Lima Xavier , Gleyson Rhuan Nascimento Campos , Luciana Conceição de Lima , Ivanovitch Silva , Antonia Egli and Theo Lynn,Illusion of Truth: Analysing and Classifying COVID-19 Fake News in Brazilian Portuguese Language

[25] Pakpoom Mookdarsanit , Lawankorn Mookdarsanit, The COVID-19 fake news detection in Thai social texts

[26] Zyte, Scrapy, version 2.6.1

[27] 02/11/2018, Đại Kỷ Nguyên là một trang phản động, tuyên truyền Pháp Luân Công

[28] Tư Nguyên, báo Nhân dân, 15-08-2018, BBC, VOA, RFA,... lại xuyên tạc để dẫn dắt dư luận xấu

[29] Van-Duyet Le, 2015, vietnamese-stopwords

[30] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras and Mark Johnson, VnCoreNLP: A Vietnamese Natural Language Processing Toolkit

[31] Rico Sennrich and Barry Haddow and Alexandra Birch, Neural Machine Translation of Rare Words with Subword Units, 10 June 2016

[32] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[33] Diederik P. Kingma, Jimmy Lei Ba, ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION, 2015