HNUE JOURNAL OF SCIENCE
Natural Sciences 2024, Volume 69, Issue 3, pp. 67-77
This paper is available online at http://hnuejs.edu.vn/ns
DOI: 10.18173/2354-1059.2024-0036

IDENTIFY STUDENT QUESTIONS ABOUT TRAINING INSTITUTIONS FROM ONLINE MEDIA POSTS

Luong Thai Le[1], Nguyen Quang Duy[2] and Lo Nguyen Thanh Vinh[1]
[1]*Faculty of Information Technology, University of Transport and Communications, Hanoi city, Vietnam*
[2]*Faculty of International Education, University of Transport and Communications, Hanoi city, Vietnam*
[*]Corresponding author: Luong Thai Le, e-mail: luongthaile80@utc.edu.vn

Received September 5, 2024. Revised October 25, 2024. Accepted October 31, 2024.

Abstract. Automatically identifying and understanding students' questions about problems they encounter or issues related to their universities is very important for universities to promptly grasp the aspirations of their students. This enables them to support and satisfy their students and enhance their reputation. Especially as social networks and online media continue to develop, students can easily post their questions and concerns online. This makes it easier for universities to access and address student questions. Although this is not a new problem, it still faces many challenges due to issues in natural language processing. To address this problem, within the scope of this article, we conduct a survey, perform experiments, and propose a model to automatically classify students' questions into 11 areas of interest at the University of Transport and Communications. We conducted careful experiments with a dataset of more than ten thousand posts collected from websites, forums, and school fan pages. Finally, we obtained a model with prediction results that achieved an accuracy of over 85%.

Keywords: intention understanding, opinion mining, deep learning, natural language processing.

1. Introduction

Nowadays, students are much more willing to express their opinions on online social media channels. According to our statistics from forums, fan pages, and websites of the University of Transport and Communications, the number of posts containing opinions and concerns of students on issues related to their universities accounts for 20% of the total number of posts. Although this rate is not high, if through this data source, the university has a tool that can automatically analyze and understand the aspirations and concerns of students in a timely manner, it will not only satisfy students but also enhance

the reputation of the university. In the scope of this article, we are only interested in understanding the intentions from online posts containing questions or opinions of students related to the University of Transport and Communications, for example: "*Guys, if I have bought AIA insurance, do I need to pay for health insurance at school*?".

The problem of identifying student queries through online posts can be classified as a problem of understanding user intent in particular and a problem of natural language processing in general. According to our research, the problem of understanding user intent has been of interest since the early 2000s, and there are two main approaches to solving this problem: (1) Based on user behavior such as clicking on links, saving pages, login history, etc. [1], [2]; (2) Based on semantic, syntactic, and lexical features of users' text posts [3]-[7]. Early research on this problem mostly followed approach (1), typically Lee et al. (2005) [1], Ashkan et al. (2008) [2]. With the appearance of social networks such as Facebook (2004), and Twitter (2006), data in the form of text such as user posts (posts, comments, etc.) began to increase significantly, so research on understanding user intentions from online posts also developed more strongly. Some of the typical studies that follow this way are proposed by Ashkan et al. (2009) [3], and Castellanos et al. (2012) [4]. Unlike a query, a user post on online media is often longer and contains more information. This is an advantage to help understand user intentions more fully and accurately. However, posts contain more noise, that is, there are many sentences in a post that are just greetings or gossip, while the number of sentences containing user intentions and questions is very small. This also causes many difficulties for the system when determining user intentions. Then most of the initial research directions were only aimed at determining whether a post had an intention or not, one of which was the research of V. Gupta et al. (2014) [5]. The authors modeled the problem of determining user intent as binary classification, which is to classify user posts into two classes: PI (Purchase Intention) and non-PI (non-Purchase Intention). Later, a few authors modeled the problem of understanding user intent in multi-class classification, typically the study of Wang et al. (2015) [6]. Realizing that user intent is very rich and diverse, the authors proposed to classify posts with user intent into six corresponding classes: Food & Drink, Travel, Career & Education, Goods & Services, Events & Activities, and Trifle. To build a classification model, the authors used a semi-supervised learning method based on the intention graph with vertices being tweets or IKs (Intent Keywords). With the fast propagation characteristics of the graph and the algorithm for calculating the weights on the edges, the authors built a multi-layer classification model with significantly higher accuracy than other methods at the same time. In a deeper approach, the authors Le Luong Thai et al. (2017) [7] used the Bidirectional Long Short-Term Memory (BiLSTM) deep learning method combined with Conditional Random Fields (CRF) to extract the main content of the user's intention and the supporting information for that main intention.

Then, there are many different approaches to solving the problem of understanding user intent. In this study, to identify and understand student questions from online posts, we propose a classification model using the Attention-based BiLSTM technique combined with some other machine learning models. Our models try to classify student questions and concerns into 11 areas of interest that will be presented in detail below. For example, the post "I have a question: I have registered for health insurance at school. If I get sick, can I go straight to the Transport and Communications hospital for examination or do I need any

other procedures?" will be classified into the "Tuition & Health Insurance" class; while the post "I currently want to join some extracurricular clubs at school but don't know which one to join. Please give me some reviews" will be classified into the "Extracurricular Activities" class. To build an accurate model that reflects reality, we automatically collect online posts from the University of Transport and Communications students. We crawl data from forums, fan pages, and websites such as utc.edu.vn, thongtindaotao.utc.edu.vn, sinhvienhientai.utc.edu.vn, etc. After performing the data preprocessing steps, we carefully conduct experiments with machine learning and deep learning models to find the most suitable model. The model with the best experimental results gives an F1 result of about 85%, which is a fairly positive result and shows that the method we choose is suitable for the problem.

2. Content

2.1. Problem statement

After collecting and studying data, we found that users' questions are very diverse. To understand these questions, we propose to build a model that automatically classifies posts into 11 semantic classes; these semantic classes will be presented specifically in the experimental data section. Therefore, we model the problem as a multi-class classification problem with the input and the output as below.

Input:

- $T = \{t_1, t_2, \ldots, t_n\}$: Set of posts and comments containing questions, opinions, and concerns of students about the University of Transport and Communications.
- M = {Study and Exams, Dormitory, Tuition and Health Insurance, Library, …}: set of k labels corresponding to k semantic classes

Output:

- A multi-class classification model: $f(t): T \rightarrow M$
- Semantic class of the post t_i

2.2. Word embeddings

2.2.1. BERT

BERT is a multi-layer Transformer word encoding model introduced in [8], designed to pre-train word vectors for unlabeled data by combining the left and right context of the word across all layers. With L being the number of Transformer blocks, H being the size of the hidden state, A being the number of self-interested endpoints, and P being the total number of parameters in the model, BERT has two main models:

- B_{BASE}: L = 12, H = 768, A = 12, P = 110M
- B_{LARGE}: L = 24, H = 1024, A = 16, P = 340M

The input of BERT is not a sentence like other models but a contiguous text segment. BERT uses the WordPiece word segmentation tool with 30,000 tokens. The first token is always the [CLS] token, then the input tokens are encoded into encoded tokens, denoted as EEE. To prepare the input vector of the model, each sentence in the input text is first separated by the token [SEP]; this [SEP] token will be encoded together with the data

marking whether the word belongs to sentence A or sentence B, and the word position in the sentence, as described in Figure 1.

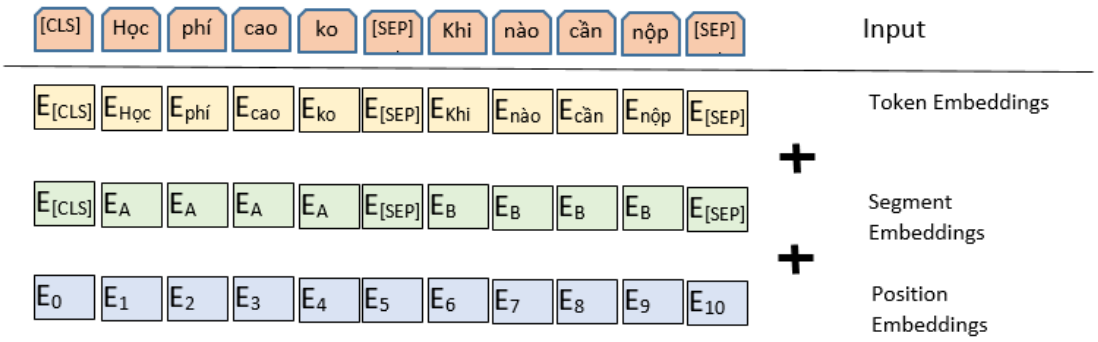


Figure 1. Word Embeddings in BERT

BERT is then trained with two unsupervised machine learning models: Masked LM, and Next Sentence Prediction. Masked LM trains an LM by replacing a random token in the input sentence with a [MASK] token. For example, the sentence "The facilities are very modern" can be masked as "The facilities are very [MASK]." The training step will help the model guess the correct MASK word. Next Sentence Prediction (NSP) is a method to check whether B is the next sentence of A (as in the example), or B is some other sentence. For example, with the input "I am a UTC student. UTC is great," then A is "I am a UTC student," and B is "UTC is great."

2.2.2. PhoBERT

PhoBERT has two versions, $\text{PhoBERT}_{\text{BASE}}$ and $\text{PhoBERT}_{\text{LARGE}}$, corresponding to two versions of BERT [9]. We use $\text{PhoBERT}_{\text{BASE}}$ for our model. To perform sentence segmentation and word segmentation, PhoBERT uses the VnCoreNLP library and the fastBPE tool, respectively, and finally produces a word encoding vector with a dimension of 256. We use the PyTorch framework with the following training parameters: batch size: 1024; learning rate: 0.0004; number of epochs: 40.

2.3. Proposed models

We propose an ensemble model to solve our problem. This model combines three single models: SVMs, MLP, and AB-BiLSTM, which will be described below. We also propose to apply PhoBERT embeddings for all experimental models.

2.3.1. Support vector machines (SVMs)

SVMs are a classification algorithm that can handle both linear and nonlinear data. In the case of nonlinear data, it will use a kernel function to transform the original data set into a new space with a higher dimension to process. The main idea of the SVM method is to build a hyperplane to divide the data into two halves to maximize the margin on both sides of the hyperplane. So the essence of the SVM method is to divide the data into two classes, that is, binary classification. To handle the multi-class classification problem, we use a technique called one-vs-one.

2.3.2. Multi-layer perceptron (MLP)

An MLP is a network of interconnected artificial neurons, organized into multiple layers. The network consists of an input layer, one or more hidden layers, and an output layer. Each layer is fully connected to the previous layer. Each neuron in the network receives an input, applies an activation function to the weighted sum of its inputs, and passes the result as an output to the neurons in the next layer. The activation function can be Tanh, Sigmoid, or ReLU depending on the problem being solved.

2.3.3. Attention-based Bidirectional Long short-term memory (AB-BiLSTM)

We inherit the AB-BiLSTM model proposed by Xianglu Yao [10] and add a word embedding layer using the PhoBERT model as shown in Figure 2.

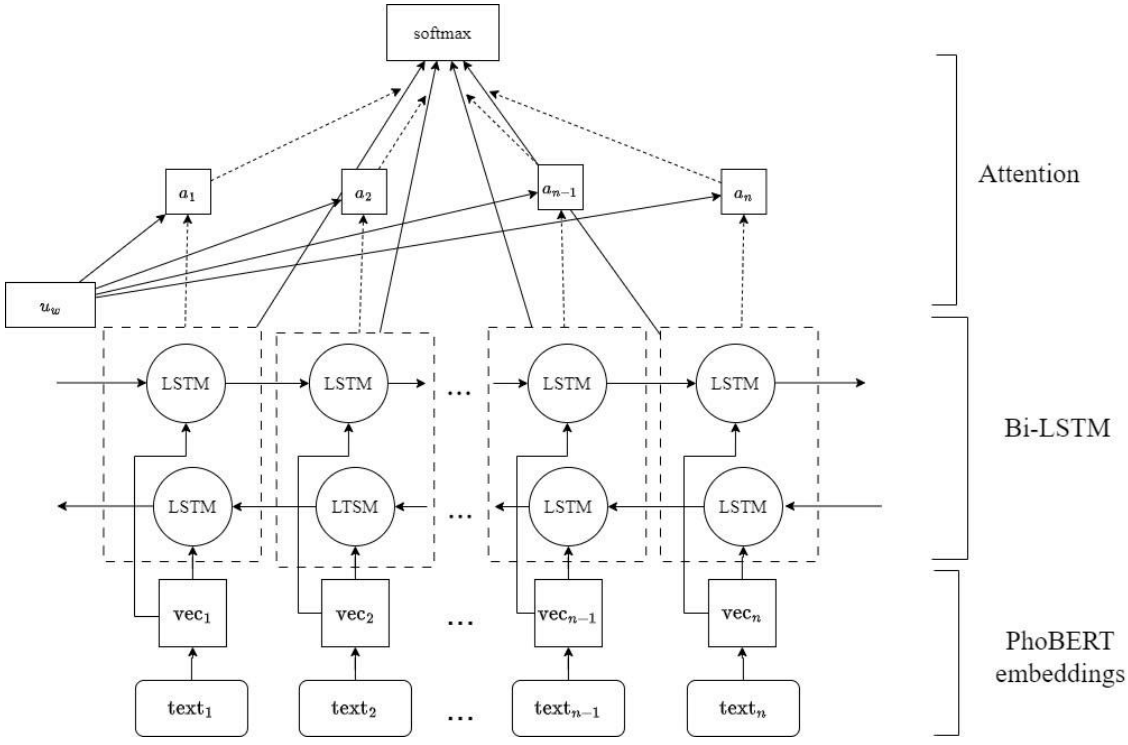


Figure 2. AB-BiLSTM with PhoBERT embeddings model

PhoBERT embeddings: With the PhoBERT word encoding technique presented specifically in section 2.2, each word $\boldsymbol{x_i}$ in the input string $\boldsymbol{x}$ will be encoded by a 256-dimensional vector $\boldsymbol{w_i}$

Bi-LSTM: Bi-LSTM is a deep learning model designed from the combination of two LSTM models: (1) the forward LSTM model calculates and represents the context from the left side of the input object; (2) the backward LSTM model calculates and represents the context from the right side of the input object. Specifically, with an input sequence x, the forward LSTM model will read from the beginning to the end of the sequence x, that

is, from $\boldsymbol{w_1}$ to $\boldsymbol{w_n}$, while the backward LSTM model reads from the end to the beginning of the sequence x, that is, from $\boldsymbol{w_n}$ to $\boldsymbol{w_1}$. Therefore, the backward model will capture the future context of x, which is something that the forward LSTM cannot do. Finally, the output $\boldsymbol{h_t}$ of the Bi-LSTM model at time *t* will be obtained from the concating of the two components $\boldsymbol{fh_t}$ and $\boldsymbol{bh_t}$ from the two forward LSTM and backward LSTM models respectively:

$$h_t = [fh_t, bh_t] \quad (1)$$

Attention: The attention mechanism aims to calculate the context vector for a sentence or a paragraph. The output h_t of the Bi-LSTM model will be passed through a 1-layer MLP model. This MLP learns to transform the representation of h_t into a feature space suitable for calculating the attention value u_t. The value a_t is the attention weight at time t. It indicates the importance level of the word at time t compared to other words in the sentence. The larger the value of a_t, the more the model pays attention to that word. a_t is computed based on the dot product between the vector u_t and the context vector of the word at time t (u_w). The context vector u_w contains information about the semantics of that word. The dot product between the two vectors indicates their level of similarity. If u_t and u_w are more similar, then the value of a_t will be larger. Finally, the attention model will calculate the significance weight value c for the input sequence through a softmax function. Specifically, we have the following functions:

$$u_t = \tanh\left(W_w h_t + b_w\right) \quad (2)$$

$$a_t = \frac{\exp\left(u_t^T u_w\right)}{\sum_t \exp\left(u_t^T u_w\right)} \quad (3)$$

$$c = \sum_t a_t h_t \quad (4)$$

2.3.4. PhoBERT-SMA model

To improve the result, we experimented with a learning model that combines three single models: SVMs, MLP, and AB-BiLSTM; we call this model the PhoBERT-SMA model. The final prediction result of the ensemble model is based on the majority choice from the prediction results of the three single models. In the case that each single model gives a different prediction result, we rely on the model with the highest prediction confidence to choose it as the final prediction.

2.4. Building a partition consisting of semantic classes

After carefully studying the collected data, we proposed a partition consisting of 11 semantic classes to classify the posts. The 11 classes are *Others*, *Study & Examination*, *Dormitory*, *Tuition & Health Insurance*, *Facilities & Infrastructure*, *Training Quality*, *Extracurricular Activities*, *Scientific Research & Cooperation*, *Library*, *Security*, *Scholarship & Policies*. Choosing a reasonable partition is not simple because the partition must have a small number of semantic classes but must cover all the questions of students. To ensure that criterion, we must use a class *Other* to contain all the questions that do not belong to the remaining 10 semantic classes. Table 2 shows examples of posts belonging to the 11 corresponding classes.

Table 1. Examples of posts that belong to 11 semantic classes

No.	Class name	Examples
1	*Others*	I'm a female born in 2005, and I have long wished to get into UTC. Is there any way to increase my chances of entering UTC? =)))
2	*Study & Examination*	I would like to ask if I get a D+ in subjects over 5% of total credits, will my graduation rank be reduced as in section 2?
3	*Dormitory*	Can you guys tell me about the dormitory? - Does the dorm have parking and electric vehicle charging? - Is the dorm gym good and how much does it cost?
4	*Tuition & Health insurance*	I have a question for you: I have registered for health insurance at school. If I get sick, can I go straight to the Transport Hospital or do I need any other procedures?
5	*Facilities & Infrastructure*	Small school but full of sports training and meeting areas. Full of high-end projectors and air conditioning. Tuition is very reasonable, around 6 million to 7 million per semester.
6	*Training quality*	I heard that the construction engineering industry is quite attractive, so I would like to take this opportunity to review some of your school's industry, such as entrance scores, teaching quality, and after graduation...
7	*Extracurricular activities*	I currently want to join some extracurricular clubs at school but don't know which one to join. Please give me some reviews.
8	*Scientific research & Cooperation*	Guys, let me ask you something, how many people can join a student research group at most? Thank you! ❤
9	*Library*	In case I don't have a student card, do I need to prepare any documents or procedures to be able to borrow books from the library?
10	*Security*	Guys, can I ask you, do you often lose things in the dormitory? ☺
11	*Scholarship & Policies*	What is the tuition fee reduction for policy students at UTC?

2.5. Experiments and results

2.5.1. Experimental data

After automatically collecting data from student posts on the online media channels of the University of Transport and Communications, we filtered out posts that were promotional or contained non-standard content, and finally, we obtained 10,090 posts. We had a team of five people (including lecturers and specialists from the University of Transport and Communications) to help label the data. The final label for each instance was decided by majority voting from the group. Figure 3 is a graph showing the distribution of posts belonging to the 11 corresponding classes in the partition we presented in section 2.4. This chart also shows the statistics of students' concerns about universities in general and the University of Transport and Communications in particular. From this chart, it can be seen that the issues that students care about the most are "Study & Examination" with a rate of 19%, followed by "Scientific Research & Cooperation" with 12%; these are also the two areas that the university puts first in its development strategy. In addition, we also noticed an imbalance in this data set because some classes only account for a rate of 6%. Therefore, during the training process, we used weights for each class to reduce the model's excessive attention to classes with a high data rate.

For problems in the field of natural language processing, data preprocessing is especially important. So we carry out the following five preprocessing steps: (1) Word segmentation: we use the VnCoreNLP library to segment Vietnamese words, then complex words in Vietnamese will be segmented according to their true meaning; for example, the phrase "học sinh" will be divided into one segmentation "học_sinh" instead of being split into two words "học" and "sinh"; (2) Remove spaces, symbols, and images expressing emotions; (3) Convert abbreviations and slang words into standard form; (4) Tokenize sentences into smaller units: words and phrases; (5) Remove stop words.

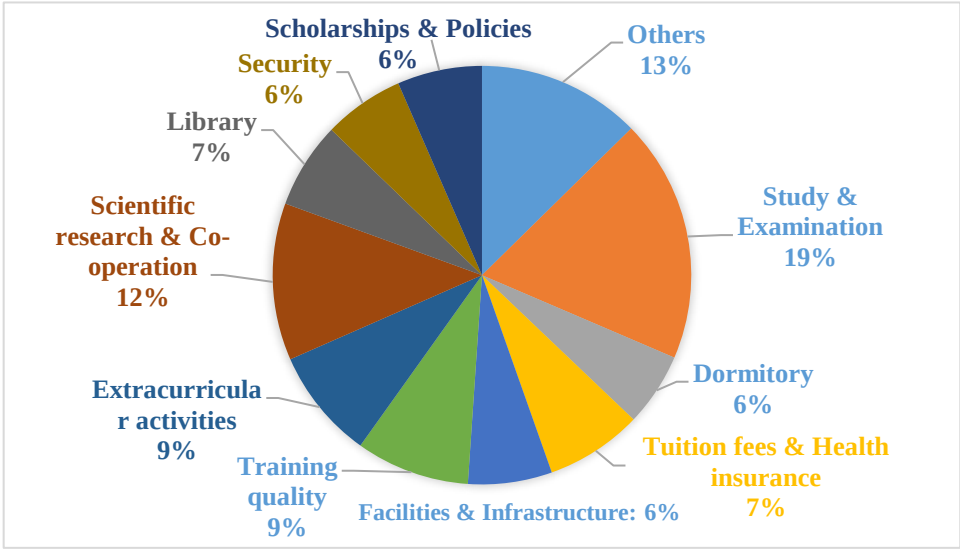


Figure 3. Proportion of data belonging to each semantic class

Then we split the data into three parts in the ratio of 60% : 20% : 20%, corresponding to training, validation, and testing sets, to prepare for training models.

2.5.2. Experimental design

To select the most appropriate model for the problem, we carefully conducted the following experiments:

- *Experiment with the SVM model:* For the SVM model, we used the Gaussian RBF kernel function, and the gamma parameter was set automatically. We carefully conducted experiments with the SVM model many times to choose the best parameter C, and finally, we decided to choose the value for C as 2.5.

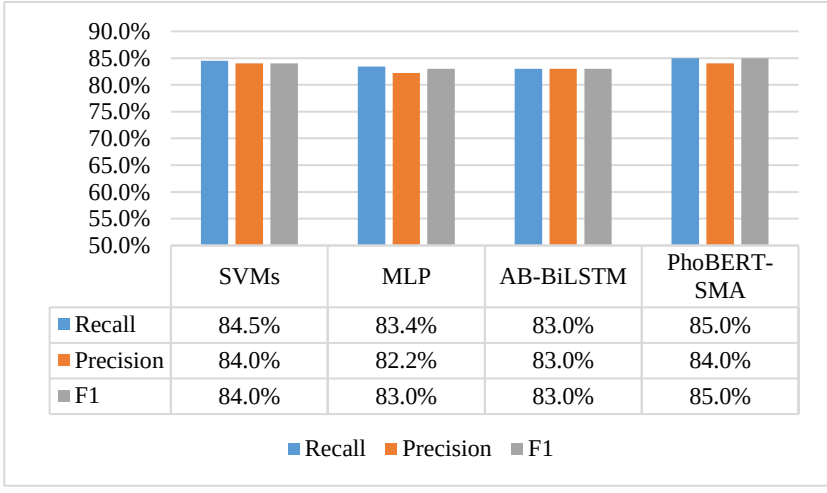
- *Experiment with MLP model:* We used batch size 256. We included a safe epoch parameter in the Keras callback API on_epoch_end(). This means that stopping rules are only applied when the number of epochs exceeds or is equal to the safe epoch parameter. Besides, to ensure that the model was not overfitting, we used a dropout parameter of 0.6; this parameter had been selected after testing the model many times. The total number of parameters used in this model is 99,851.

- *Experiment with AB-BiLSTM model:* Each input example was re-represented as a three-component object with corresponding sizes of (None, 8, 96), and then we used the same batch size and epochs parameters as the MLP model to train the AB-BiLSTM model. Finally, we used the dropout value of 0.4. The total number of parameters used in this model is 84,684.

- *Experiment with the PhoBERT-SMA model*: We use parameters for each single model the same as above.

2.5.3. Results and discussion

To evaluate the models, we use Precision, Recall, and F1 measures. Figure 4 shows the results of the experimental models: SVMs, MLP, AB-BiLSTM, and PhoBERT-SMA. It can be seen that among the three single models, SVMs is the model that achieves the highest result. This can be explained because with a training dataset of only more than 10 thousand sentences divided into 11 classes, the two deep learning models MLP and AB-BiLSTM cannot best demonstrate their self-learning ability. However, with an average F1 score of over 82% and quite even between the two Precision and Recall measures, it can be seen that these deep learning models are suitable for our classification problem.



	SVMs	MLP	AB-BiLSTM	PhoBERT-SMA
Recall	84.5%	83.4%	83.0%	85.0%
Precision	84.0%	82.2%	83.0%	84.0%
F1	84.0%	83.0%	83.0%	85.0%

Figure 4. The result of each experimental model

Furthermore, using the PhoBERT model to train the input word embeddings for the models also helps to improve the results of the models by about 8% compared to not using PhoBERT, as presented in Table 2. This result shows that our choice is reasonable. Besides, careful data preprocessing has helped us have a good, standard dataset, contributing to improving the result of the model.

Table 2. Results comparison between the models with or without PhoBERT

Model	SVMs	AB-BiLSTM	PhoBERT-SMA
F1 with PhoBERT	84,0%	83,0%	85,0%
F1 without PhoBERT	75,5%	74,0%	76,1%

Finally, the PhoBERT-SMA model combining the three single models gives the best prediction results for all three types of metrics. Interestingly, when those single models are combined to build the ensemble model, components based on models with lower F1 scores contribute positively and boost the overall results of the whole ensemble model. This means that each component in the proposed models can support its counterparts to reach better overall performance. This is reasonable because the combined learning model takes advantage of all three single-component models. Moreover, the prediction of the ensemble model is decided based on the majority voting from three single models, so it improves the chances of correct results. Table 3 presents the results for each class when predicting using the PhoBERT-SMA model. It can be seen that the average F1 score of all classes is greater than 74%, which shows that the model we built is suitable for the proposed problem.

Table 3. Results of the PhoBert-SMA model

Class name	Recall	Precision	F1
Others	75.1%	82.2%	78.5%
Study & Examination	81.0%	82.4%	81.7%
Dormitory	89.3%	77.9%	83.2%
Tuition & Health insurance	84.7%	88.5%	86.6%
Facilities & Infrastructure	81.5%	76.5%	78.9%
Training quality	75.8%	72.9%	74.3%
Extracurricular activities	88.1%	90.8%	89.4%
Scientific research & Cooperation	90.9%	87.4%	89.1%
Library	89.1%	89.1%	89.1%
Security	93.6%	93.6%	93.6%
Scholarship & Policies	88.5%	85.9%	87.2%

3. Conclusions

In this paper, we propose a hybrid learning model to solve the problem of understanding and identifying students' questions about the training institutions they are interested in. We model this problem as a multi-class classification problem with 11 semantic classes to predict each student's question into a corresponding class. With a dataset of more than ten thousand student posts on online media, we conduct experiments with machine learning and deep learning models. To improve the accuracy of the model, we use the PhoBERT model to train the input vector and carefully preprocess the data. Finally, we obtain a hybrid learning model combining three single models SVMs, MLP, and AB-BiLSTM, which is the model that gives the highest prediction results among the models we experimented with, with F1 accuracy reaching more than 85%.

Acknowledgements. This research is funded by the University of Transport and Communications (UTC) under grant number T2024-CN-006.

REFERENCES

[1] Lee U, Liu Z & Cho J, (2005). Automatic identification of user goals in the web search. *Proceedings of the 14th International Conference on World Wide Web*.

[2] Ashkan A, Clarke CL, Agichtein E & Guo Q, (2009). Classifying and characterizing query intent. *Proceedings of 31st European Conference on Information Retrieval Research,* 578–586.

[3] Ashkan A & Clarke CL, (2009). Term-based commercial intent analysis. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval,* 800-801.

[4] Castellanos M & et al. (2012). Intention insider: discovering people's intentions in the social channel. *Proceedings of the 15th International Conference on Extending Database Technology*.

[5] Gupta V, Varshney D, Jhamtani H, Kedia D & Karw D, (2014). Identifying purchase intent from social posts. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1).

[6] Wang J, Cong G, Zhao WX & Li X, (2015). Mining user intents in Twitter: a semi-supervised approach to inferring intent categories for tweets. In *Proceedings of Association for the Advancement of Artificial Intelligence*.

[7] Luong TL, Cao MS, Le DT & Phan XH, (2017). Intent Extraction from Social Media Texts Using Sequential Segmentation and Deep Learning Models. *Proceedings of the 9th International Conference on Knowledge and Systems Engineering,* 215-220.

[8] Devlin J, Chang MW, Lee K & Toutanova K, (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805.

[9] Nguyen DQ & Nguyen AT, (2020). PhoBERT: Pre-trained language models for Vietnamese. *Proceeding of Findings of the Association for Computational Linguistics: EMNLP*.

[10] Yao X, (2017). Attention-based BiLSTM neural networks for sentiment classification of short texts. *Proceedings of the International Conference on Information Science and Cloud Computing,* 110-117.