NEU-ESC: A Comprehensive Vietnamese Dataset for Educational Sentiment Analysis and Topic Classification

Nguyen Quang Hung† , Mai Phan Quoc Hung† , Duong Phuong Giang†,
Nguyen Thi Hong Hanh† , Huan Vu† , Tien Cuong Nguyen*, and Thien Van Luong†

†Business AI Lab, Faculty of DS&AI, National Economics University (NEU), Hanoi, Vietnam
*VNPT AI, VNPT Group, Hanoi, Vietnam
Email: thienlv@neu.edu.vn

Abstract—In education, comprehending students' desires through their comments is essential and that demand for Vietnamese language is no exception. While numerous research efforts have focused on developing datasets for general Vietnamese text analysis, such datasets for education areas have been overlooked. Specifically, current datasets on education often lack domain specificity or contain inadequate common slang. This paper addresses these gaps by introducing a new Vietnamese dataset for Educational Sentiment Analysis and Topic Classification called NEU-ESC, specifically curated from university forums. Compared to existing educational datasets, our proposed dataset offer more samples, more sentiment and topic classes, as well as a larger sample length, i.e., the average number of words per sample, and a larger vocabulary size. We then evaluate our dataset by fine-tuning various BERT models as initial benchmarks on both single-task and multitask classification. Our NEU-ESC dataset is published at `https://huggingface.co/datasets/hung20gg/NEU-ESC`.

1. Introduction

Social listening is important respond to the needs of various stakeholders in the digital age. In education, particularly online learning environments, understanding students' desires through their comments is essential. It helps monitor discussion forums, gain valuable insights from students' post comments, and avoid spam or unnecessary drama.

Text classification, a hot topic in Natural Language Processing (NLP), has numerous applications across various domains. In education, it can be beneficial for analyzing and categorizing large volumes of student-generated content, such as forum posts and feedback. However, there is a notable lack of comprehensive datasets about education in Vietnam that offer deep classification of content, particularly from university forums. This gap presents a significant challenge for researchers and educators who aim to leverage NLP techniques to improve the educational experience in the Vietnamese context. A well-curated dataset with detailed content classification could provide valuable insights into student behaviors, preferences, and learning patterns specific to the Vietnamese educational system.

Inspired by these issues, our contributions include two folds as follows. Firstly, we propose a novel Vietnamese dataset called NEU-ESC for Educational Sentiment Analysis and Topic Classification, which offers more samples, more sentiment and topic classes, a larger sample length and a larger vocabulary size than existing Vietnamese education datasets. Secondly, we provide a detailed analysis of the NEU-ESC dataset using a variety of BERT variants for both single-task and multitask classification schemes, which serve as initial benchmarks for our proposed dataset.

2. Related Works

Text classification and sentiment analysis are two common tasks in NLP, involving assigning predefined labels or categories to text documents, sentences, or phrases based on their content. Text classification plays a vital role in educational management for understanding students' desires.

There have been several major contributions to the text classification datasets for Vietnamese NLP. For example, [8] focuses on analyzing hate speech and comments on social media, while [6] proposes two different datasets for sentiment analysis, namely, VLSP 2016 for news and VLSP 2018 for hotels and restaurants. The UIT-VSFC dataset [7] comprises over 16,000 student feedback entries, offering two different label schemes for sentiment-based and topic classification tasks. This dataset is valuable for educational research, providing insights into student experiences and perceptions, and is mainly used as a benchmark for text classification. However, none of these datasets fully meet the requirements for analyzing open education forums on social media platforms. Particularly, while the UIT-VSFC dataset is relatively formal and lacks common slang, the others are not closely related to the education sector.

We leverage the convenience and effectiveness of fine-tuning different BERT models, such as, BERT [3], XLM-Roberta [2], phoBERT [5], vELECTRA [1] and VisoBERT

Corresponding author: Thien Van Luong.
ORCID iDs Nguyen Quang Hung: 0009-0006-8312-064X, Mai Phan Quoc Hung: 0009-0006-9558-7052, Duong Phuong Giang: 0009-0008-9844-1918, Nguyen Thi Hong Hanh: 0009-0008-7365-3594, Huan Vu: 0000-0002-0785-6907, Tien Cuong Nguyen: 0009-0005-0533-5728, Thien Van Luong: 0000-0002-4661-4900

Criteria	UIT-VSFC [7]	Our NEU-ESC
Total samples	16,175	32,966
Sentiment labels	3	4
Topic labels	4	10
Avg. words/sample	14.23	25.09
1-10 words	7,232	13,638
10-20 words	6,061	9,934
20-50 words	2,702	6,169
50+ words	180	3,225
Vocabulary size	2,877	12,759

Table 1: Comparison between UIT-VSFC and NEU-ESC.

[9], the pre-trained models that use the English Wikipedia corpus to perform multiple NLP tasks. By fine-tuning these pre-trained models on our proposed NEU-ESC dataset, we can achieve good performance with task-specific data, making them ideal for our text classification evaluation.

3. Dataset Collection

3.1. Data curation pipeline

The raw data was collected from various sources, mostly on university groups, and community forums on Facebook. For the collection of hate speech, we crawled the comments from unregulated forums such as xamvn and vozer.

As expected from collected data, the use of teen code and typography errors are unavoidable. Hence, after the data scraping step, the data go through the preprocessing step by cleaning, remove redunance character and emoji, ect. During that process, we also notice many acronyms used among the commenters. Many of these are for school subjects and majors. For instance, "qtkd" is commonly used as shorthand for "Quản trị kinh doanh" (Business Administration), and "nlkt" stands for "Nguyên lý kế toán" (Principles of Accounting). Hence, we create a mapping dictionary between these acronyms and their full-length text. After applying such the data curation pipeline, our proposed NEU-ESC dataset finally has nearly 33,000 samples.

3.2. Dataset composition

On the sentiment analysis task, students' comments belong to each of these four categories: *Neutral*, *Positive*, *Negative*, and *Toxic*. For the topic classification task, students' comments are categorized into 9 different topics, namely, *Spam*, *News*, *Academic*, *Other*, *Service*, *Job & Recruitment*, *Personal Affairs*, *Social Affairs*, *Helping & Sharing*, and *Club & Events*. Our dataset expands upon UIT-VSFC [7] by incorporating a broader range of labels. We include 4 sentiment categories and 10 topic classifications, compared to UIT-VSFC's 3 sentiments and 4 topics. Additionally, our dataset covers a significantly wider scope, extending beyond the limited domain of student feedback. Finally, our dataset comprises much more long sentences of more than 50 words than UIT-VSFC. Table 1 shows a detailed statistic comparison between the two datasets.

Label	Sentiment	Count	Percentage	Mean length
0	Neutral	22,773	69.08%	23.21
1	Positive	4,148	12.58%	24.29
2	Negative	5,200	15.77%	34.30
3	Toxic	845	2.56%	22.78

Table 2: Statistic of sentiment analysis labels.

Label	Classification	Count	Percentage	Mean length
0	Spam	405	1.23%	22.71
1	News	902	2.74%	59.55
2	Academic	10,512	31.89%	29.62
3	Other	14,402	43.69%	11.40
4	Service	2,358	7.15%	30.94
5	Job & Recruitment	808	2.45%	55.14
6	Personal Affairs	1,478	4.48%	33.17
7	Social Affairs	769	2.33%	67.11
8	Help & Share	670	2.03%	37.03
9	Club & Events	662	2.01%	68.82

Table 3: Statistic of topic classification labels.

NEU-ESC contains a total of 32,966 samples. The distribution of the number of each sentiment and topic label pair is divided into train, validate, and test sets with a ratio of $7/1/2$, corresponding to $23,048$ for the training set, $3,305$ for the validation set, and $6,613$ for the test set.

3.3. Label distribution

The dataset's sentiment distribution in Table 2) reveals that most comments are neutral (69.08%), implying factual or unemotional feedback. Negative comments (15.77%) highlight dissatisfaction, while positive comments (12.58%) reflect satisfaction. Toxic comments are minimal (2.56%) but require management for a respectful environment. Negative comments are generally more detailed, as shown by their higher mean text length.

The classification distribution in Table 3 shows that the majority of comments fall into the *Other* category with 14,402 samples (43.69%). Academic-related comments are the second largest category with 10,512 samples (31.89%), underscoring the significance of academic discussions among students. Other notable classifications include *Service* with 2,358 samples (7.15%) and *Job & Recruitment* with 808 samples (2.45%). *Club & Events* and *Social Affairs* have the longest mean text lengths, indicating more detailed comments in these categories.

4. Experiment Results

4.1. Single-task classification with BERT

In this research, we focus on several common base versions of BERT for Vietnamese natural language understanding, namely, BERT [3], XLM-Roberta (XLM-R) [2], phoBERT [5], vELECTRA [1] and VisoBERT [9].

With single-task classification models, we focus on three metrics: *Accuracy*, *Macro F1* (mF1), and *Weighted F1* (wF1). We first train the model on the training set, eval-

Model	Sentiment Analysis			Topic Classification		
	Accuracy	mF1	wF1	Accuracy	mF1	wF1
$BERT_{base}$[3]	77.34	70.01	76.68	75.32	57.65	75.06
XLM-R_{base}[2]	81.46	75.91	81.27	77.58	60.87	77.60
vELECTRA [1]	78.47	71.73	78.12	75.40	53.46	74.39
VisoBERT [9]	**82.78**	75.79	**82.17**	77.94	60.65	78.06
$phoBERT_{base-v2}$ [5]	81.75	**77.70**	82.02	**78.65**	**62.73**	**78.36**

Table 4: Performance of BERT models on single-task classification using the proposed NEU-ESC dataset.

Model	Sentiment Analysis			Topic Classification		
	Accuracy	mF1	wF1	Accuracy	mF1	wF1
XLM-R_{base}[2]	81.13	75.83	80.99	77.35	59.93	77.50
VisoBERT [9]	82.17	75.41	82.48	77.86	61.03	78.17
$phoBERT_{base-v2}$ [5]	**82.88**	**77.64**	**82.85**	**79.18**	**63.60**	**79.59**

Table 5: Performance of BERT models on multitask classification using the proposed NEU-ESC dataset.

uate it on the validation set, and select the models with the highest mF1 for testing. Table 4 shows the detailed performance of a single classification task on different models.

As observed in Table 4, $phoBERT_{base-v2}$ emerges as the leading model in both Sentiment Analysis and Topic Classification tasks. For Sentiment Analysis, it achieves the highest scores with 81.75 accuracy, 77.70 mF1, and 82.02 wF1. VisoBERT has a higher accuracy score and wF1 but a significantly lower mF1, and XLM-R_{base} followed closely behind. In Topic Classification, $phoBERT_{base-v2}$ maintains its top position with 78.65 accuracy, 62.73 mF1, and 78.36 wF1, again follows by strong performances from VisoBERT and XLM-R_{base}. In contrast, vELECTRA and $BERT_{base}$ underperform in both tasks. As such, $phoBERT_{base-v2}$'s consistent superiority, especially in wF1 scores, across both tasks, establishes it as the most effective and versatile model. Hence, we will select this model as well as VisoBERT and XLM-R_{base} for implementing multitask classification in the next subsection.

4.2. Multitask classification with BERT

Rather than only fine-tuning on an individual task, one can alternatively make use of multitask learning to update BERT. For example, [4] introduced models that incorporate a framework combining a shared BERT layer with task-specification prediction heads. Therefore, using a shared BERT layer enables the model to leverage cross-task data and generate more generalized representations. Multitask learning effectively increases the sample size for training the model, thus leading to performance improvement by increasing the generalization of the model.

We modify the loss function of the model to follow (1), where $\mathcal{L}_{sentiment}$ and $\mathcal{L}_{topic}$ are simply Cross Entropy Losses for each classification task,

$$\mathcal{L}_{multi} = \mathcal{L}_{sentiment} + \mathcal{L}_{topic}. \quad (1)$$

For multitask classification, we focus on the top three models from our previous single-task evaluations. Evaluating models in a multitask setup is trickier, so we had to be selective. We look for models that show solid results on both tasks in the validation set before our final testing. Table 5 provides a detailed breakdown of these results.

VisoBERT and XLM-R_{base} demonstrate resilience in the multitask environment, with their performance on both tasks only face minor decline compared to single-task results. Interestingly, their F1 macro scores for topic classification improved, increasing by 0.5 and 2 points respectively. The standout performer, however, is $phoBERT_{base-v2}$ by showing significant improvement in the multitask setting. We observe enhancements of 0.5 to 1 point across all benchmarks, which is a notable achievement in a multitask context.

4.3. Error Analysis

For our error analysis, we use the results from the $phoBERT_{base-v2}$ multitask model, which provided the best performance overall. Figure 1 displays the normalized confusion matrix by recall of both classification tasks. With the sentiment analysis task, the model demonstrates high recall score on *Neutral*, indicating the high ability to capture neutral expression. This is because *Neutral* samples are dominant over three other classes in our dataset, as shown via Table 2. However, approximately 28% of *Positive* and *Negative* were misclassified as *Neutral*, suggesting some difficulty in distinguishing clear sentiments from neutral expressions. For the *Toxic* label, although the normalized recall was 0.65, 28% of the misclassified instances were *Negative*, which is within an acceptable range, and just a fraction of others was misclassified as *Toxic*.

In the topic classification tasks, the dominant classes such as *Academic* and *Other* achieve the highest recalls of 0.85 and 0.88, respectively. However, the other classes having less samples prove to be more ambiguous, leading to their low or moderate recalls of ranging from 0.43 to 0.67. The majority of misclassified labels are mainly either *Academic* and *Other*. This observation is understandable due to

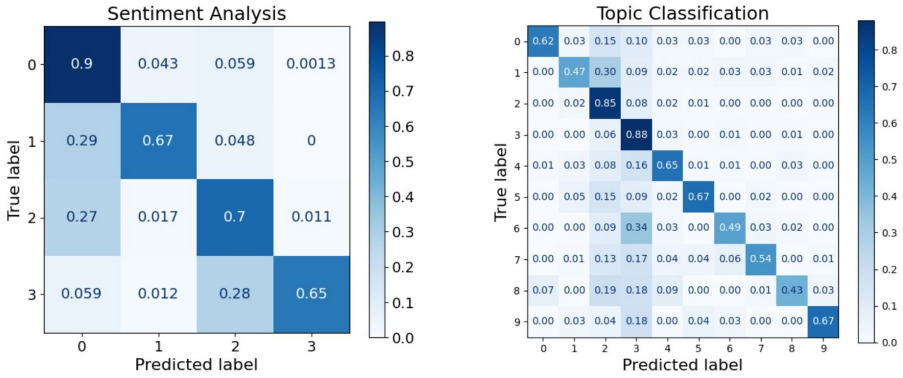

Figure 1: Confusion matrix of both classification tasks evaluated by phoBERT$_{\text{base-v2}}$ on the proposed NEU-ESC dataset.

the imbalance in the dataset across different classes. More specifically, classes with fewer examples are harder for the model to learn effectively, resulting in lower scores.

4.4. Limitations and Future Work

The dataset is currently imbalanced. Most comments focused on topics related to other or academic within the confines of college life. This limited scope presents a significant challenge for developing broadly applicable models. To address this issue, it is crucial to expand the dataset to encompass a wider range of subjects and perspectives.

A more diverse dataset would ideally include comments from various age groups, professional backgrounds, and life experiences beyond the academic sphere. This could involve gathering input on topics such as career development, personal finance, healthcare, technology, social issues, and global events. By broadening the scope of the dataset, we can capture a more accurate representation of real-world conversations that we consider as our future work.

5. Conclusion

This paper introduced a comprehensive dataset specifically designed for applications in social listening on academic forums. The proposed NEU-ESC dataset comprises 32,966 samples collected from university forums, providing a robust foundation for understanding and interpreting online conversations within educational settings. Compared with existing Vietnamese educational datasets, our dataset has more samples, more sentiment and topic classes as well as larger vocabulary size and higher average length. Additionally, we investigated several BERT baseline models for both single-task and multitask classification. Experiment results suggested that multitask models tend to outperform single-task models on our NEU-ESC dataset.

References

[1] The Viet Bui et al. “Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models”. In: *Proceedings of the 34th Pacific Asia Conf. on Language, Information and Computation*. 2020, pp. 13–20.

[2] Alexis Conneau et al. “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 8440–8451.

[3] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 2019, pp. 4171–4186.

[4] Xiaodong Liu et al. “Multi-Task Deep Neural Networks for Natural Language Understanding”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. July 2019, pp. 4487–4496.

[5] Dat Quoc Nguyen et al. “PhoBERT: Pre-trained language models for Vietnamese”. In: *EMNLP*. 2020, pp. 1037–1042.

[6] Huyen T M Nguyen et al. “VLSP SHARED TASK: SENTIMENT ANALYSIS”. In: *Journal of Computer Science and Cybernetics* 34.4 (2019), 295–310.

[7] Kiet Van Nguyen et al. “UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis”. In: *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. 2018, pp. 19–24.

[8] Luan Thanh Nguyen et al. “Constructive and Toxic Speech Detection for Open-Domain Social Media Comments in Vietnamese”. In: *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*. 2021, pp. 572–583.

[9] Nam Nguyen et al. “ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 5191–5207.