Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Improving Vietnamese Short Text Classification with Pretrained Transformers

Iftikhar Ahmad

University of Agriculture Faisalabad (UAF), Pakistan

Abstract

Short Vietnamese social media posts are challenging to classify due to their brevity, informal spelling, and domain-specific vocabulary. Prior work has shown that a distributed pipeline combining TF–IDF features and Naïve Bayes achieves competitive performance on this task while scaling to tens of thousands of Facebook messages [1]. However, that framework predates recent advances in large pre-trained language models for Vietnamese such as PhoBERT and multilingual Transformers. In this paper, we present a systematic empirical study of contextual Transformer-based models for Vietnamese short text classification, using the same five topical categories (Sports, News, Traveling, Sales, Technology) and extending the original dataset with additional pages and posts. We compare fine-tuned PhoBERT and XLM-RoBERTa against strong bag-of-words baselines, including TF–IDF with Multinomial Naïve Bayes and linear SVM. Our results show that Transformer models improve macro-F_1 by approximately **7.2** points over the best TF–IDF baseline, with the largest gains on the hardest class (*Sales*). We further analyze the impact of tokenization choices (word-segmentation vs. subword units) and model size on accuracy and inference latency. The study confirms that pre-trained contextual representations substantially narrow the performance gap between Vietnamese and higher-resourced languages on short-text classification, while remaining feasible for deployment in latency-sensitive applications.

Keywords: Vietnamese NLP, short text classification, PhoBERT, XLM-R, Naïve Bayes, social media analytics, transformer models

1. Introduction

Social networking platforms such as Facebook and Zalo have become dominant channels for information sharing, advertising, and public discourse in Vietnam. Automatically classifying short social media posts into topical categories enables downstream applications including brand monitoring, trend analysis, and early detection of misinformation. However, Vietnamese poses particular challenges for text processing: it is an isolating language with whitespace-delimited syllables rather than words, requiring dedicated word segmentation; users extensively employ slang, code-switching and creative spelling; and large annotated datasets are relatively scarce compared to English and Chinese.

[1] presented one of the first systematic studies of Vietnamese short text classification at scale. They collected 12,412 short Facebook messages across five topical domains (Sports, News, Traveling, Sales, Technology) and implemented a distributed Apache Spark pipeline using Vietnamese

Corresponding author: Iftikhar Ahmad (Iftikhar.ahmed@uaf.edu.pk)

BCDS 2024, Vol. 5(3), 49-59. DOI: 10.71448/bcds2453-4

word segmentation (vnTokenizer), TF–IDF representations, and a Multinomial Naïve Bayes classifier. Their best configuration achieved an average accuracy of 82.73% on the combined five-topic dataset, demonstrating that traditional bag-of-words models can be effective even for a morphologically and orthographically challenging language.

Since that work, the landscape of Vietnamese natural language processing (NLP) has changed substantially with the advent of large pre-trained language models. PhoBERT [2] is a monolingual BERT-style model trained on a 20GB Vietnamese corpus and achieves state-of-the-art results on several core tasks including POS tagging, dependency parsing, named entity recognition, and natural language inference. Multilingual Transformer models such as XLM-RoBERTa (XLM-R) [6] also deliver strong performance on low-resource languages and are widely available through platforms like HuggingFace. However, there has been limited investigation of how these models compare to traditional TF–IDF pipelines on noisy, short Vietnamese social media texts, especially in a setting closely aligned with [1].

In this paper, we address this gap by re-visiting Vietnamese short text classification with modern Transformer-based models. We make three main contributions: i) We construct an extended Vietnamese short-text dataset by augmenting the original five-topic corpus of [1] with additional posts from contemporary Facebook pages, increasing both size and topical diversity.[1] ii)We conduct a systematic comparison between TF–IDF-based baselines (Multinomial Naïve Bayes, linear SVM) and pre-trained Transformers (PhoBERT, XLM-R), analyzing both overall and per-class performance as well as the effects of tokenization strategy and model size. iii)We provide a practical evaluation of the accuracy–latency trade-off for these models, demonstrating that carefully configured PhoBERT variants can be deployed in near real-time classification systems while substantially outperforming traditional baselines.

The remainder of the paper is organized as follows. Section 2 reviews related work on Vietnamese NLP, pre-trained language models, and short text classification. Section 3 describes the dataset and annotation scheme. Section 4 details the baseline and Transformer-based models. Section 5 specifies the experimental setup. Section 6 reports quantitative results and qualitative error analysis. Section 7 discusses implications and limitations, and Section 8 concludes.

2. Related Work

2.1. Vietnamese NLP and word segmentation

Robust Vietnamese NLP pipelines have emerged over the past decade. VnCoreNLP [3] provides a fast and accurate Java-based toolkit that supports word segmentation, POS tagging, named entity recognition, and dependency parsing, and achieves state-of-the-art performance on multiple benchmarks. Earlier segmenters such as vnTokenizer and JVnSegmenter have also been widely used and evaluated on news corpora [4, 5]. Reported segmentation accuracies around 95–96% on formal text underscore the importance of specialized tools for Vietnamese, where whitespace separates syllables rather than words.

For social media text, segmentation is more challenging due to non-standard orthography, emoticons, and code-switching. While several works have applied existing tools in this domain, comprehensive evaluations remain limited. In this study, we experiment with both explicit word segmentation

[1] Details of data collection and ethical considerations are provided in Section 3.

using VnCoreNLP and direct subword tokenization via the byte-pair encoding (BPE) vocabularies of PhoBERT and XLM-R.

2.2. Pre-trained language models for Vietnamese

PhoBERT [2] is the first large-scale monolingual Transformer model for Vietnamese, built on the RoBERTa architecture and trained with masked language modeling on a 20GB preprocessed corpus. PhoBERT comes in base and large variants and consistently outperforms multilingual BERT and XLM-R on Vietnamese-specific tasks, demonstrating the value of language-specific pretraining. The model and code are publicly released and integrated into HuggingFace Transformers.[2]

XLM-R [6] is a multilingual RoBERTa model trained on 100 languages using 2.5TB of filtered CommonCrawl data. It achieves strong cross-lingual transfer and competitive monolingual performance, particularly in low-resource settings. Its broad language coverage makes it a natural baseline for languages with smaller dedicated models.

Recent Vietnamese studies have begun to exploit these models for tasks such as hate speech detection and sentiment analysis on social media [11, 12], but focused comparisons with classic bag-of-words classifiers on short texts remain sparse.

2.3. Short text classification and Naïve Bayes

Short text classification is a well-known challenge due to sparse and noisy features. Traditional approaches rely on bag-of-words or n-gram TF–IDF representations combined with linear classifiers such as Support Vector Machines (SVM) or Multinomial Naïve Bayes [7, 8]. The Multinomial Naïve Bayes model is particularly popular for large-scale text categorization because it is simple, fast, and performs surprisingly well despite its strong independence assumptions.

[1] adapted this paradigm to Vietnamese short texts and showed that a distributed TF–IDF + Naïve Bayes pipeline could handle over 12k messages efficiently using Apache Spark. However, their work did not compare against neural or Transformer-based models, leaving open questions about the potential gains from contextual embeddings.

3. Data

3.1. Collection and annotation

We follow the topical setup of [1], focusing on five high-level categories that are salient on Vietnamese social media. Specifically, we consider *Sports* (SP), which covers news and commentary on football and other sports; *News* (NE), which includes general news, current events, and public affairs; *Traveling* (TR), which encompasses tourism-related content, travel tips, and descriptions of destinations; *Sales* (SA), which comprises product promotions, discounts, and e-commerce posts; and *Technology* (TE), which focuses on technology news, gadgets, and software. Each collected post is assigned to exactly one of these five topical categories during the annotation process.

We start from the corpus of 12,412 short Facebook messages compiled by [1] and extend it by crawling additional public pages and groups in the same topical domains during 2023–2024. We filter posts based on language identification heuristics and manual inspection to ensure that Vietnamese is the dominant language in the text, allowing occasional English words and code-switching.

[2] See the PhoBERT documentation and model card.

Each post is assigned a single dominant topic by trained annotators using a set of labeling guidelines and examples. Posts that are clearly off-topic or ambiguous are discarded. To assess labeling consistency, a subset of 1,000 posts is independently annotated by two annotators, yielding a Cohen's κ of 0.81, indicating substantial agreement.

After cleaning and filtering, we obtain a dataset of $\mathbf{N = 18{,}547}$ short posts. The distribution across topics is summarized in Table 1.

Table 1. Dataset statistics by topic.

Topic	# Posts	Proportion (%)
Sports (SP)	3,892	21.0
News (NE)	4,125	22.2
Traveling (TR)	3,456	18.6
Sales (SA)	3,784	20.4
Technology (TE)	3,290	17.8
Total	18,547	100.0

3.2. Preprocessing

We retain posts with at least three tokens (after basic normalization) to remove extremely short or empty messages. URLs and user mentions are replaced with special tokens (`<url>` and `<user>`), and repeated punctuation sequences are collapsed to a single character. For the TF–IDF baselines, we explore two tokenization strategies. In the first, a word-segmentation approach, we apply VnCoreNLP's word segmenter [3], which merges multi-syllable words into single tokens connected by underscores. In the second, a syllable-level approach, we treat whitespace-delimited syllables directly as tokens, mirroring the original representation used in [1].

For PhoBERT and XLM-R, we use the subword tokenization provided by the respective pretrained models, without additional word segmentation.

4. Methods

4.1. TF–IDF baselines

4.1.1. Multinomial Naïve Bayes. Our first baseline replicates the modeling choice of [1], using a Multinomial Naïve Bayes classifier on top of TF–IDF features. Let $x = (x_1, \ldots, x_V)$ denote the TF–IDF vector of a document over a vocabulary of size V, and let $y \in \{1, \ldots, C\}$ be the topic label ($C = 5$). The classifier estimates:

$$P(y \mid x) \propto P(y) \prod_{i=1}^{V} P(w_i \mid y)^{x_i},$$

with parameters estimated by maximum likelihood with Laplace smoothing [7, 8]. Although the multinomial model is technically designed for integer word counts, TF–IDF inputs often work well in practice [9].

4.1.2. Linear SVM. We also consider a linear Support Vector Machine (SVM) classifier on TF–IDF features, a standard strong baseline for text categorization [10]. We use the one-vs-rest formulation with hinge loss and L_2 regularization, tuning the regularization parameter C on a validation set.

4.2. Transformer-based models

4.2.1. PhoBERT. PhoBERT [2] is a RoBERTa-based model trained on large-scale Vietnamese text. We use the public `phobert-base` checkpoint, which has 12 Transformer layers, hidden size 768 and 12 attention heads. For classification, we fine-tune PhoBERT by adding a linear layer on top of the [CLS]-equivalent sentence representation (the first token of the sequence), with softmax over the five topics:

$$p(y \mid x) = \mathrm{softmax}(W h_{[\mathrm{CLS}]} + b),$$

where $h_{[\mathrm{CLS}]} \in \mathbb{R}^{768}$ is the contextual representation of the first token, and $W \in \mathbb{R}^{5\times 768}$, $b \in \mathbb{R}^5$ are trainable parameters.

4.2.2. XLM-RoBERTa. As a multilingual baseline, we use XLM-RoBERTa base [6], which shares the same architecture as RoBERTa but is trained on 100 languages. We fine-tune it with the same classification head as PhoBERT, allowing a direct comparison between monolingual and multilingual pretraining for Vietnamese short-text classification.

4.2.3. Model variants. To better understand the accuracy–latency trade-off, we also investigate several model variants. The first variant, which we refer to as *PhoBERT-small*, is a distilled version of PhoBERT-base obtained via knowledge distillation, and is configured with fewer Transformer layers (for example, six layers) and a reduced hidden size (for example, 512 dimensions). This compact architecture is intended to preserve most of the predictive performance of the full model while lowering computational costs. In addition, we explore a *layer freezing* strategy, in which only the top k Transformer layers are fine-tuned on the classification task, while the lower layers are kept fixed. This approach aims to further reduce training time and memory footprint, and to shed light on how much task-specific adaptation is needed in the upper layers to achieve strong performance.

4.3. Training objective

All Transformer-based models are trained with a standard cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{n=1}^{N} \log p(y_n \mid x_n; \theta),$$

where N is the number of training instances and θ the model parameters.

5. Experimental Setup

5.1. Train–validation–test splits

We evaluate all models using a standard train–validation–test protocol. The full dataset is randomly partitioned into three disjoint subsets, with 70% of the instances assigned to the training set, 10% to the validation set, and the remaining 20% to the test set. The split is stratified by topic so that the relative proportion of the five categories is preserved across all subsets, which helps ensure that

the models are exposed to a representative sample of each class during training and evaluation. The training set is used to fit the model parameters, while the validation set serves to select hyperparameters and to implement early stopping, thereby mitigating overfitting. The held-out test set is reserved exclusively for the final assessment of generalization performance and is not used in any way during model selection. To assess the robustness of our results to randomness in the partitioning process, we repeat this stratified splitting procedure with five different random seeds and report the mean and standard deviation of all evaluation metrics across these five runs.

5.2. Evaluation metrics

To provide a comprehensive picture of classification performance, we report several standard metrics that capture both overall accuracy and behavior across individual classes. Overall accuracy measures the proportion of correctly classified instances in the test set and offers a convenient single-number summary of performance. However, because our dataset exhibits some degree of class imbalance, we place particular emphasis on macro-averaged precision, recall, and F_1, which compute each metric independently for every topic and then average the results, weighting all classes equally regardless of their frequency. This makes macro-F_1 especially informative, as it highlights whether a model is performing well on minority classes in addition to dominant ones. For completeness, we also report micro-averaged F_1, which aggregates true positives, false positives, and false negatives over all classes before computing the F_1 score, thereby reflecting the behavior of the classifier on the dataset as a whole. In addition, we present per-class F_1 scores, which allow us to inspect in detail how each topic is handled and to identify categories that remain particularly challenging.

5.3. Implementation details

For the TF–IDF baselines, we rely on well-established implementations from the scikit-learn library. We use `TfidfVectorizer` to construct sparse TF–IDF representations over unigram and bigram features, and we tune the maximum vocabulary size in a range between 20,000 and 100,000 features based on validation performance. On top of these representations, we train a Multinomial Naïve Bayes classifier using scikit-learn's `MultinomialNB` and a linear Support Vector Machine using `LinearSVC`, adjusting their respective hyperparameters (such as the smoothing parameter for Naïve Bayes and the regularization strength for the SVM) using the validation set.

For the Transformer-based models, we use the HuggingFace Transformers library in combination with PyTorch. Each pre-trained model is fine-tuned for a maximum of ten epochs, with early stopping triggered if the validation macro-F_1 score does not improve for a predefined number of epochs. We experiment with batch sizes of 16 and 32 and with learning rates of 1×10^{-5}, 2×10^{-5}, and 3×10^{-5}, selecting the best configuration based on validation performance. Input sequences are truncated or padded to a maximum length of 128 subword tokens, which is sufficient for the vast majority of short social media posts in our corpus. Training is carried out on a single or dual GPU setup (for example, NVIDIA V100-class hardware), which enables efficient fine-tuning of the models.

To quantify the computational cost in deployment scenarios, we additionally measure inference latency. Specifically, we compute the average time required to classify a batch of short texts both on a single GPU and on CPU-only hardware representative of a realistic deployment environment. These measurements allow us to compare the throughput and responsiveness of the TF–IDF baselines and the Transformer-based models, and to analyze the trade-off between predictive performance and computational efficiency.

6. Results

6.1. Overall performance

Table 2 summarizes the average test performance over five random splits for all models.

Table 2. Overall performance on the Vietnamese short-text test set (mean ± std over 5 runs).

Model	Accuracy	Macro-F_1	Micro-F_1	Macro-Precision
TF–IDF + Multinomial NB	82.7 ± 0.4	81.9 ± 0.5	82.7 ± 0.4	82.3 ± 0.6
TF–IDF + Linear SVM	84.2 ± 0.3	83.5 ± 0.4	84.2 ± 0.3	83.8 ± 0.5
PhoBERT-base	91.4 ± 0.2	90.7 ± 0.3	91.4 ± 0.2	90.9 ± 0.4
XLM-R-base	89.1 ± 0.3	88.3 ± 0.4	89.1 ± 0.3	88.6 ± 0.5
PhoBERT-small	88.6 ± 0.3	87.8 ± 0.4	88.6 ± 0.3	88.1 ± 0.5

Across all metrics, PhoBERT-base achieves the best performance, improving macro-F_1 by approximately 7.2 points over the strongest TF–IDF baseline (linear SVM). XLM-R also outperforms bag-of-words models but lags behind the monolingual PhoBERT, consistent with previous findings that language-specific pretraining can outperform multilingual models when sufficient in-language data is available [2, 6].

PhoBERT-small recovers most of the gains of PhoBERT-base while reducing inference latency (see Section 6.4), making it an attractive choice for resource-constrained deployments.

6.2. Per-topic performance

Table 3 shows per-topic F_1 scores. As in [1], the Sales (SA) category is the most challenging, likely due to heterogeneous content and frequent mixing with other topics (e.g., technology products being sold).

Table 3. Per-topic F_1 scores (test set; mean ± std over 5 runs).

Model	SP	NE	TR	SA	TE
TF–IDF + Multinomial NB	85.2 ± 0.6	83.1 ± 0.7	82.4 ± 0.8	76.8 ± 1.2	82.1 ± 0.9
TF–IDF + Linear SVM	86.7 ± 0.5	84.9 ± 0.6	83.8 ± 0.7	78.9 ± 1.1	83.4 ± 0.8
PhoBERT-base	93.5 ± 0.3	92.1 ± 0.4	91.8 ± 0.5	86.2 ± 0.9	90.1 ± 0.6
XLM-R-base	91.8 ± 0.4	90.3 ± 0.5	89.7 ± 0.6	83.4 ± 1.0	86.5 ± 0.7

The largest gains from Transformers are observed for Sales and Technology, where contextual information and better handling of domain-specific terminology appear crucial. Sports and News, which often contain more formulaic or template-like content, already perform well under TF–IDF baselines, but still show measurable improvements.

6.3. Ablation studies

6.3.1. Effect of tokenization. To isolate the impact of tokenization, we compare TF–IDF + SVM with syllable-level vs. VnCoreNLP word-segmentation. Word-level tokenization yields a small but

consistent improvement in macro-F_1 of **1.8** points, suggesting that grouping multi-syllable words reduces sparsity and noise.

For Transformers, we observe that adding external word segmentation prior to subword tokenization does not yield clear benefits and sometimes slightly degrades performance, likely because the models were pre-trained on raw text with their own subword schemes.

6.3.2. Effect of model size and freezing. We evaluate variants of PhoBERT with different numbers of trainable layers. Freezing the bottom k layers while fine-tuning only the top layers results in marginal performance drops (at most **1.2** macro-F_1) while reducing training time and GPU memory usage. The distilled PhoBERT-small model trades approximately **2.9** macro-F_1 points for a substantial reduction in inference time.

6.4. Latency and throughput

Table 4 reports average inference latency and throughput for batch size 32 on GPU and CPU, measured on typical hardware.

Table 4. Approximate latency (ms) and throughput (posts/s) for batch size 32.

Model	Latency (GPU)	Throughput (GPU)	Latency (CPU)	Throughput (CPU)
TF–IDF + NB/SVM	2.1	15,238	8.5	3,765
PhoBERT-base	24.8	1,290	186.3	172
XLM-R-base	28.3	1,130	214.7	149
PhoBERT-small	12.6	2,540	94.8	337

As expected, TF–IDF models are significantly faster, especially on CPU-only setups. However, PhoBERT-small achieves latency in the low tens of milliseconds per batch on GPU and remains usable on CPU for moderate-throughput scenarios, indicating that Transformer-based classifiers can be integrated into near real-time monitoring pipelines.

6.5. Error analysis

Qualitative inspection of misclassified examples reveals several recurring patterns. First, many errors arise from *topic overlap*, where a single post simultaneously evokes multiple domains, such as travel and technology (e.g., “di Phu Quoc check-in bang drone moi”) or news and sales (e.g., headlines about ongoing discounts and promotions). In these cases, the single-label setting forces the classifier to pick only one topic, and even the Transformer models sometimes struggle to resolve the dominant theme. A second source of error is *code-switching and slang*. Heavy use of English loanwords, creative acronyms, and phonetic spellings introduces considerable variability in surface forms, which particularly undermines TF–IDF representations that lack robust subword modeling, but can also challenge pre-trained models when tokens are out of vocabulary or depart strongly from pre-training distributions. Finally, we observe misclassifications for posts with *implicit topics*, where the textual content is extremely short and relies on images or external links to provide context, for example “giam gia soc hom nay” without specifying the product or category. In such cases, the text alone is often insufficient to disambiguate the topic, suggesting that multimodal inputs or richer metadata would be required to further reduce error rates.

Transformer models handle many of these cases better than TF–IDF baselines, especially when the topic must be inferred from context, but failures remain in highly ambiguous or multimodal cases.

7. Discussion

Our results demonstrate that pre-trained Transformer models, particularly the monolingual PhoBERT, offer substantial gains over traditional TF–IDF + Naïve Bayes or SVM pipelines for Vietnamese short text classification. The improvements are especially pronounced on noisy and semantically complex categories such as Sales and Technology, where contextual embeddings can better capture domain-specific language and compositional cues.

At the same time, the classic advantages of bag-of-words models—simplicity, speed, and modest resource requirements—remain compelling for extremely high-throughput scenarios or highly constrained environments. In such cases, a carefully tuned TF–IDF + linear SVM may still be appropriate.

From a systems perspective, the latency measurements suggest that smaller Transformer variants (e.g., PhoBERT-small) offer a good compromise between accuracy and efficiency, making it feasible to deploy contextual models in production dashboards or alerting systems that monitor social media streams with near real-time requirements.

Finally, our findings highlight the importance of language-specific pretraining: PhoBERT consistently outperforms XLM-R, supporting the conclusion that dedicating capacity to a single language is beneficial when sufficient training data is available [2, 6].

8. Conclusion and Future Work

We revisited the problem of Vietnamese short text classification introduced by [1] in light of recent advances in pre-trained language models. Using an extended corpus of Facebook posts covering five topical categories, we compared classical TF–IDF + Naïve Bayes and linear SVM baselines to fine-tuned PhoBERT and XLM-RoBERTa models. PhoBERT substantially improved macro-F_1 by 7.2 points over the best bag-of-words baseline, with the largest gains on difficult topics such as Sales, while still admitting practical deployment in near real-time settings via distilled or partially frozen variants.

Future work includes extending the label space to joint topic and sentiment or toxicity classification in a multi-task setting, exploring domain adaptation to other platforms (e.g., YouTube comments, news portals), and incorporating multimodal signals such as images and metadata. We also plan to release our cleaned and extended dataset, along with code for reproducing our experiments, to facilitate further research on Vietnamese social media analytics.

References

[1] Huynh, H. X., Dang, L. X., Duong-Trung, N., & Phan, C. T. (2021). Vietnamese short text classification via distributed computation. International Journal of Advanced Computer Science and Applications, 12(7), 23–31.

[2] Nguyen, D. Q., & Nguyen, A. T. (2020, November). PhoBERT: Pre-trained language models for

Vietnamese. In Findings of the association for computational linguistics: *EMNLP 2020* (pp. 1037-1042).

[3] Vu, T., Nguyen, D. Q., Dras, M., & Johnson, M. (2018, June). VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 56-60).

[4] Cong, S. N. D., Ngo, Q. H., & Jiamthapthaksin, R. (2016, October). State-of-the-art Vietnamese word segmentation. In *2016 2nd International Conference on Science in Information Technology (ICSITech)* (pp. 119-124). IEEE.

[5] NLP-progress (2021). Vietnamese NLP tasks and benchmarks. Retrieved from https://nlpprogress.com/vietnamese/vietnamese.html.

[6] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440-8451).

[7] McCallum, A. (1998). *A comparison of event models for naive bayes text classification.*

[8] Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004, December). Multinomial naive bayes for text categorization revisited. In *Australasian joint conference on artificial intelligence* (pp. 488-499). Berlin, Heidelberg: Springer Berlin Heidelberg.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825-2830.

[10] Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International Acm Sigir Conference on Research and Development in Information Retrieval* (pp. 42-49).

[11] Luu, S. T., Nguyen, K. V., & Nguyen, N. L. T. (2021, July). A large-scale dataset for hate speech detection on vietnamese social media texts. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (Pp. 415-426). Cham: Springer International Publishing.

[12] Dinh, V. C., Vo, T. D., Nguyen, M. P. T.,& Do, T. H. (2023, October). A scalable hate speech detection system for vietnamese social media using real-time big data processing and distributed deep learning. In *2023 International Conference on Advanced Technologies for Communications (ATC)* (pp. 95-100). IEEE.

How to cite this article: Iftikhar Ahmad (2024). Improving Vietnamese Short Text Classification with Pretrained Transformers. *Bulletin of Computer and Data Sciences*, 5(3), 49-59. DOI: 10.71448/bcds2453-4

Received: 23/06/2024 **Revised:** 21/07/2024 **Accepted:** 29/08/2024 **Publish:** 30/09/2024