International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-15 Issue-1, December 2025

Context-Enriched Sentiment Analysis for Short Vietnamese Restaurant Reviews Using Large Language Models

Nguyen Thi Thanh Thuy, Nguyen Ngoc Diep

Abstract: ***Sentiment analysis of short text has posed a significant challenge in natural language processing, particularly for context-rich and low-resource languages such as Vietnamese. User-generated texts are usually brief; therefore, they do not explicitly express their sentiments. Consequently, traditional models struggle to process those reviews. This paper introduces a new approach that leverages the strengths of large language models to address the gap in context scarcity. The method works primarily in two ways: a) by feeding in structured metadata, such as restaurant name and location, directly into the model input, and b) using large language models to automatically generate likely contextual sentences so that short reviews become long informative statements. Results from comprehensive experiments carried out on a newly assembled Vietnamese food review dataset show improved sentiment analysis output based on this kind of context enrichment, beating several strong baselines, including the state-of-the-art monolingual PhoBERT model, particularly when it came to resolving semantic vagueness typical of ultra-short word reviews or even short reviews with implicit subjects. This work offers a strong, flexible approach to addressing the problem of missing context in low-resource languages. This will bring value to both the commercial world and academic study.***

Keywords**:** ***Sentiment Analysis, Large Language Models, Context Enrichment, Vietnamese, Short Text.***

Nomenclature:
LLMs: Large Language Models
SVMs: Support Vector Machines
BoW: Bag-of-Words
RAG: Retrieval-Augmented Generation

I. INTRODUCTION

Over the last 10 years, the rapid growth of the digital economy has made e-commerce and online services the main drivers of growth in Vietnam. This situation is most evident in the food and restaurant sector, where digital platforms such as Foody.vn and Google Maps are widely used to enhance consumer outreach.

Manuscript received on 26 November 2025 | First Revised Manuscript received on 03 December 2025 | Second Revised Manuscript received on 08 December 2025 | Manuscript Accepted on 15 December 2025 | Manuscript published on 30 December 2025.
*Correspondence Author(s)
Dr. Nguyen Thi Thanh Thuy, Department of Information Technology, Posts and Telecommunications Institute of Technology, A2, Hanoi, Vietnam. Email ID: thuyr205@gmail.com, ORCID ID: 0000-0002-0243-0835
Dr. Nguyen Ngoc Diep*, Department of Information Security, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Email ID: diepnguyenngoc@ptit.edu.vn, ORCID ID: 0000-0003-1757-9517

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open-access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

User-generated reviews are now the most powerful driver of consumer behaviour because they reflect customer satisfaction and the real customer experience.

To further enhance response times and brand reputation management in the digital environment, organisations need support applications that break down massive volumes of online feedback into customer satisfaction assessments. One such application is automatic sentiment analysis. However, building such applications efficiently remains one of the most significant challenges, particularly for Vietnamese developers. The reasons are the special linguistic features and limitations in processing constraints [1].

The analysis of sentiment in short text falls far behind the high benchmark performances achieved on longer samples. This is called the "context deficit problem". Much more critically, it invalidates the perceived, explainable polarity in sentiment classification. Practically, typical F&B reviews on Vietnam's platforms are short and ultra-short forms with just a few words like "Ngon" [Delicious], "Tệ" [Bad], or "Tạm được" [Average]. The extreme contextlessness of the text left too much open to interpretation, leading to confusion about the true meaning. The classic bag-of-words-based approach to language processing, as well as neural network architectures, relies heavily on surrounding context to determine sentiment and meaning. In their absence, internal model information is lost; hence, low accuracy in sentiment prediction [2]. PhoBERT [3] and Bert-GCN [4] are state-of-the-art systems that established new performance baselines for Vietnamese, but rely heavily on this type of text. The primary reason for such a performance gap lies in the very nature of contemporary models, since self-attention lacks sufficient data to capture semantic relationships in extremely short text inputs.

It is further made worse by subtle linguistic features peculiar to the Vietnamese language:

A. Lexical Ambiguity: Words in the Vietnamese language mainly express different sentiment polarities based on the contexts in which they are used. For instance, the word "đắt" [expensive] can express negative sentiment about price ("this restaurant is expensive") or positive sentiment about popularity ("this handbag sells really well").

B. Implicit Subjects and Sarcasm: People often leave out the subject when writing their reviews (e.g., simply stating "Delicious" instead of "This dish is delicious"). Moreover, sarcasm and implicit tones are prevalent in user reviews, posing challenges that sentiment analysis models cannot adequately address when relying solely on raw textual input.

C. Event-Dependent Context: The sentiment of the review may

© Copyright: All rights reserved.

significantly depend on temporal or contextual factors, such as major holidays or special events. For example, "Slow service" might be perceived as slightly negative under normal circumstances, but could be seen as neutral or acceptable when made during the high-traffic context of the Tet festival.

Several major problems further complicate the inherent sparsity of information in short reviews. Firstly, this is due to lexical ambiguity in Vietnamese. Other issues include the complex linguistic structure of Vietnamese and the always contextual nature of sentiment shifts. This means that only improvements in model architecture will not be enough to solve the problem. Therefore, it is required to deliberately add external information to provide more context for the input data.

To bridge these gaps, this paper introduces a new approach that uses large language models (LLMs) to augment the context of short texts artificially. The methodology pivots on constructing and refining input representations that guide the behavior of advanced language models [5], enhancing precision and interpretability for Vietnamese sentiment analysis.

In general, the method operates based on two key strategies. The first part of the approach uses structured metadata, such as the restaurant name, dish, and location, as well as temporal factors (e.g., holidays, special events), which are first formatted and then explicitly added to the model input. And the second part involves the LLM-driven creation of synthetic contextual sentences. By utilizing the generative abilities of large language models, concise reviews (e.g., "Ngon" [Delicious]) are automatically extended to complete, context-rich sentences (e.g., "Món phở ở Phở Thìn rất ngon." [The pho at Phở Thìn is very delicious.]).

This work hence tries to make the following contributions:

- It offers a new way with the help of large language models in the preprocessing step that will automatically enrich the contextual information for short Vietnamese reviews.
- It carries out an empirical test of the effectiveness of the method proposed by benchmarking with strong baseline models like PhoBERT.

The rest of this paper is organised as follows. Related work comes next in Section II. The proposed methodology is discussed in Section III. Experiment design and evaluation take place in Section IV. Results and analysis are reported in Section V. Concluding remarks are presented in Section VI.

II. RELATED WORK

Early studies in Vietnamese sentiment analysis can be basically categorized into two primary methodologies. As regards the lexicon-based approach, VietSentiLex [6] is an example. It was designed explicitly for social media in Vietnam to annotate sentiments. Though easy and direct, this method cannot fully capture the complexities of pragmatic nuance, out-of-vocabulary words, and context-dependent ambiguity in text. The second approach employs conventional machine learning algorithms, such as Support Vector Machines (SVMs) and Naive Bayes, along with features such as Bag-of-Words (BoW) [7], [8]. Since these models rely on manual features, they are less expressive in capturing detailed semantic relations and the syntactic order of words in Vietnamese. Therefore, short, highly ambiguous reviews do not yield good results with either lexicon-based or traditional machine-learning approaches. They rely solely on lexical representations and lack strong contextual modelling capabilities. This has forced the research community to move toward deep learning models, which can more effectively capture context.

Deep learning marks a giant leap in the evolution of sentiment analysis. Architectures of recurrent neural networks, including RNNs, LSTMs, and particularly BiLSTMs, have proven valuable for modelling sequences [9]. For example, empirical evidence from the BiLSTM-CNN model's performance on the VSFC dataset indicated a noticeable performance improvement over traditional methods [10].

But it was truly with the emergence of Transformer architecture and pretrained language models that a real revolution in the field took place. For Vietnamese, PhoBERT [3] has been immediately recognised as the state-of-the-art monolingual model. PhoBERT is pre-trained on a large Vietnamese corpus and can effectively capture fine-grained grammatical, semantic, and syntactic distinctions typical of Vietnamese, outperforming multilingual models like mBERT [11]. It shows promising results but has apparent limitations when dealing with short texts. No matter how complex they might be, Transformer-based models typically rely on the internal context of input texts for inference. Therefore, their performance will decrease considerably as contextual clues become less abundant. Examples include one-word reviews or ultra-short reviews. The results have been verified by several studies indicating that deep learning models such as PhoBERT lose accuracy under conditions involving extremely short, even more context-deficient texts [9]. Thus, improving the architecture alone would not be sufficient. The reliance of pre-trained language models on internal contexts calls for novel strategies. In other words, the lack of information in the original text should be made up for by proactively contextualising and enriching it with external sources.

The emergence of Vietnamese large language models, such as PhoGPT [12], marks a turning point in how NLP technology research and development are carried out in Vietnam. These state-of-the-art LLMs outperform previous versions (including PhoBERT). They can carry out complex instructions, exhibit in-context learning, and, most importantly, generate rich, plausible contextual information on their own. At the same time, there are now quite a few multilingual models supporting Vietnamese, such as Gemma [13] and GPT-OSS [14], as well as open-sourced models OpenLLaMA [15] and Mistral [16], which have been tuned to better process Vietnamese. And in turn, that makes applications scalable across sectors such as education, healthcare, and finance by enabling AI systems to interact more effectively and faster with Vietnamese users.

This has led researchers in technological advancement to develop a new research paradigm known as "Context Engineering," which uses few-shot prompting and Retrieval-Augmented Generation (RAG). With this, the AI model will reason with direct external information input into its

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

reasoning process. This approach is promising for sentiment analysis of short Vietnamese reviews. Recent studies demonstrate that enriching LLMs with external information (e.g., restaurant attributes, dish details) significantly mitigates the ambiguity of ultra-short texts [17], [18]. For instance, a brief remark such as "Ngon" (Delicious) is transformed into an expanded, more informative sentence like "The pho at restaurant X is very delicious". With the ongoing development of monolingual Vietnamese LLMs and advances in context engineering techniques, this field is well-positioned to enable robust sentiment analysis solutions for Vietnamese-language applications.

III. PROPOSED METHOD

We introduce an end-to-end solution whose framework encompasses everything from data collection and architectural setup to deployment-level context-enrichment strategies for a context-deficient problem.

A. Data Construction and Preprocessing

i. *Data Sources*

A new corpus was constructed by taking reviews from two big online sites in Vietnam. The primary source is Google Maps, which provides many user reviews across different types of restaurants and places. To make language and context more varied, extra review data were also added from Foody.vn, a well-known food-review site often used in earlier works. This combined dataset ensures full coverage for model training and avoids missing context cues that can occur when only short texts are used. In detail, short reviews, which have no more than 10 words, are kept in the corpus but later expanded with additional context.

This review data has been collected under the terms and conditions of these platforms. Only publicly available user-generated content is included in the dataset. No personally identifiable information is ever stored. Any possible user identifiers in the raw data were eliminated before conducting the analysis.

ii. *Data Structure*

All reviews are keyed as an object in the fields {review_text, restaurant_name, restaurant_address, review_time}, alongside the original review text. This metadata is of great importance since it forms the basic level on which further context-enrichment strategies shall operate.

iii. *Preprocessing Pipeline*

Since user-generated content tends to carry a lot of noise, a rigorous Vietnamese-specific preprocessing procedure was adopted that integrated well-known best practices:

- **Text Normalization:** All texts are converted into lowercase and cleaned by removing emojis, memorable characters, and unnecessary HTML tags.
- **Translation of Foreign Reviews:** For reviews written in all languages except Vietnamese, the content was by default translated into Vietnamese using Google Translate. This produced linguistic homogeneity across the entire corpus, enabling uniform processing.
- **Abbreviation and Slang Expansion:** It is essential for normalization of informal texts. A manual mapping dictionary was prepared to map standard abbreviations, teencode, and misspelt words to their correct forms (e.g., "ko" → "không" /" not", "bt" → "bình thường" / normal, "vs" → "với" / "with").
- Tokenization: The word tokenizer called Underthesea was used for segmenting Vietnamese sentences. (https://github.com/undertheseanlp/underthesea).

iv. *Data Labelling*

To generate a gold-standard data set for model training and evaluation, sentiment (Positive, Negative, Neutral) was manually annotated by experienced native annotators. Two annotators independently assigned labels to each sentence after preprocessing was complete. If the two annotators disagreed on a particular label, then a third annotator reviewed that instance and made the final decision. The Vietnamese data set annotated in this study yielded a Cohen's Kappa score of 0.87, indicating high inter-annotator agreement.

B. Integration of Event Context

A significant extension in this study is the inclusion of the event context in the model. That is very important for the food service review because service quality often changes during special periods, such as holidays, the Lunar New Year, or festivals. If the temporal context is accurately recognised, we can interpret short, ambiguous reviews more accurately. We use the review timestamp (review_time) and manually cross-check against industry reports and market analysts to identify special periods, such as Tet, National Day, or Christmas, during preprocessing. If a review falls into these periods, it will be labelled with an event type (event_type).

Location data helps the system identify contextual fields such as city name (city_name) and regional type (e.g., lowland, highland, coastal) (region_type). Time and location info together let the system determine the seasonal context (season_type) at each spot, which might be four seasons in northern Vietnam or just two seasons in southern provinces (the dry and rainy seasons).

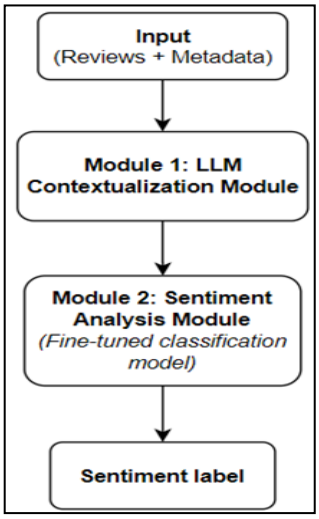
The contextual fields are then passed to the LLM Contextualization Module as auxiliary input variables. Prompts are designed to guide the LLM to generate text that is contextually relevant to the event, season, or region. For example, "Slow service" during Tet can be enriched to "The restaurant was very crowded during Tet, so service was slow." RAG-based approaches leverage real-world event activity and background data to enable the LLM to generate much richer, more accurate context.

Recent work shows that augmenting the model's input with contextual metadata, including event data and retrieval-based knowledge, significantly improves the accuracy of LLM sentiment predictions on short, low-context review datasets [5].

C. System Architecture

The proposed framework, with a sequential architecture, comprises two modules, thereby explicitly separating context enrichment from sentiment classification (Fig. 1).

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



[Fig.1: The Proposed Framework for Context-Enriched Sentiment Analysis Using LLMs. Module 1 Accepts Only Short Review Sequences and Generates Longer Reviews Based on Context Data. Module 2 Receives Input Data, Which Can Be Either the Original Non-Short Reviews or the Output from Module 1, and then Predicts the Sentiment Label]

Module 1 (LLM Contextualization Module) runs on two primary inputs: a brief review and the metadata dictionary. At its core, this module runs on an instruction-tuned LLM that can handle Vietnamese. Models considered include PhoGPT-7B5-Instruct [12] and Gemma [13]. Both models show strong text generation and instruction-following abilities for Vietnamese. The role of this module is to use one of a few context enrichment strategies (as described in Section III.D) to produce an "enriched input" sequence. This output will be a single text sequence in which the original review is placed within a richer contextual frame.

Module 2 (Sentiment Analysis Module) either receives the "enriched input" sequence from Module 1 or directly processes non-short reviews, since these do not require context-based enrichment. The main component of this module is a fine-tuned classification model, namely PhoBERT-base. This model was chosen because it has shown strong results on Vietnamese text in earlier studies. The token's vector from PhoBERT's last layer is passed through a linear layer for classification. The primary function of Module 2 is to classify the input sequence into one of three sentiment classes: Positive, Negative, or Neutral. The output of this module is the final sentiment determination.

D. Contextualization Strategies

This is the central part of the approach, where we break down the two specific ways we tackle things in Module 1 (the LLM Contextualization Module).

i. Strategy A: Direct Metadata Integration

Method: In this strategy, organised metadata and the review text are formatted as a template-based input sequence to explicitly inject structured information into the input, so that the attention of a classification model, e.g., PhoBERT, can directly learn relations between elements of metadata and the content of the review. A possible template is:

Restaurant Name: {restaurant_name}, City: {city_name}, Region={region_type}, Season={season_name}, Event={event_type}. [REVIEW] {review_text}

Borrowed from studies in other fields, it relies on the fact that additional structured information - metadata - can be very useful in boosting the performance of tasks related to text classification and prediction. In most cases, metadata supplies definite "anchor points" for the model to better resolve textual uncertainties.

ii. Strategy B: Generative Context Synthesis

- **Method:** This strategy will leverage LLMs' generative capabilities. It will provide a carefully designed prompt that guides and encourages the LLM to generate a complete, natural sentence from the short review and its metadata. For example:

"You are an assistant who converts short reviews into complete sentences. Based on the context and the following review, write a complete, natural-sounding sentence.

Context: Restaurant={restaurant_name}, City={city_name}, Region={region_type}, Season={season_name}, Event= {event_type}.

Short review: {review_text}.

- **Full sentence:"**

The output that is produced by the LLM (for instance, "The pho at restaurant X is very delicious.") will then be used as input to the Sentiment Analysis Module. This methodology uses zero-shot prompting as well as instruction following - core capabilities of today's LLMs. It takes implicit, vague utterances and converts them into explicit, well-grounded utterances, thereby disambiguating before classification.

These two strategies represent different attitudes toward the philosophy of what context is and how it should be treated: Strategy A treats context as structured, separable data, while Strategy B treats it as unstructured natural language. In this respect, a hybrid approach in which the sentence generated by Strategy B follows the structured metadata from Strategy A may yield optimal results - factually grounded yet linguistically fluent. To prove this assumption, an ablation study is required to assess the effectiveness of individual and combined approaches.

IV. EXPERIMENTAL DESIGN AND EVALUATION

A. Dataset and Baseline Models

Table I summarises the main statistics of the dataset to be analysed in the experiments. It has 30,000 restaurant reviews, with a post-preprocessing vocabulary of 10,148. The average length of these reviews is 19.86 words. However, short reviews, which account for about 40% of the data set, have an average length of only 4.93 words, underscoring how short texts may lack contextual information. The sentiment distribution is highly imbalanced, skewed toward positivity: 76.87% are positive, 16.7% negative, and only 6.29% neutral.

Table I: Key Characteristics of the Dataset

Characteristics of the Dataset	Value
Total reviews	30,000
Short reviews	11,526
Vocabulary size (post-preprocessing)	10,148
Average review length	19.86
Average short review length	4.93
Positive class (%)	76.87%
Negative class (%)	16.7%
Neutral class (%)	6.29%

The newly created and tagged The corpus splits into three subsets: Training set (80%), Validation set (10%), and test

© Copyright: All rights reserved.

set (10%). Stratified sampling ensures class shares for sentiment tags (Positive, Negative, Neutral) remain the same across all three subsets. For a comprehensive and fair comparison, we'll measure how well our new framework performs against a strong benchmark like PhoBERT. Specifically, PhoBERT-base (Context-Free) is applied directly to the preprocessed short review texts. The results of this model set up a reference level of performance that can be achieved without external help for context enrichment.

B. Evaluation Metrics and Implementation Details

i. Metrics

Precision, Recall, and F1-Score will be used in assessing the model. Precision is the rate of correct classification of a class, while recall is the rate of actual correct classification. The F1-score is the average of its components; thus, it facilitates easy comparison among classification models.

$$Precision = \frac{TP}{TP + FP} \quad \dots \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad \dots \quad (2)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \dots \quad (3)$$

Where *TP* (true positives): correct classifications of positive cases; *FP* (false positives): incorrect classifications of positive cases into the negative class; FN (false negatives): incorrect classifications of negative cases into the positive class.

ii. Implementation Details

All experiments shall use the Hugging Face Transformers library to implement PhoBERT and LLM models within the PyTorch deep learning framework. They were carried out on a machine powered by an Intel Core i7 9700K processor, 64 GB of RAM, and GeForce GTX 2080 Ti graphics processing units. PhoGPT-7B5-Instruct sometimes produces mismatched review sentences, and so far, it has not performed as well as Gemma. For example, in some test runs on simple review sentences such as “Good food,” and default context information provided in the prompts, output results from PhoGPT-7B5-Instruct can be “Alright,” “Okay,” or “Hello! I am an AI assistant that helps convert short reviews into complete sentences. Please provide me with the name of the restaurant you would like me to help with to do this.” This is most likely due to the relatively minor, less varied training corpus used compared to other multilingual LLMs. Therefore, Gemma-7B (with Q4 quantisation) will be used for the review-rewriting task in this study. Based on the current setup, it takes about 3 hours to use the Gemma-7B model to generate complete review sentences from short reviews in the dataset.

V. RESULTS AND ANALYSIS

A. Comparative Performance Analysis

The main experimental results are described in Table II. It is a comparison of the models' performance on the test set. These results are clear enough to provide quantitative evidence of the effectiveness of the proposed context-enrichment strategies. Several key observations come from Table II. First, the Pho BERT (Context-Free) baseline achieves an F1 Score of 0.822, thereby setting a strong reference point for context-free sentiment classification. All versions of the proposed context-enrichment framework will have clear improvements over this baseline.

Table II: Overall Performance Comparison of Models

Model	Precision	Recall	F1-Score
PhoBERT (Context-Free)	0.825	0.819	0.822
Proposed Framework (Metadata Only)	0.844	0.861	0.852
Proposed Framework (Generative Context Only)	0.876	0.871	0.873
Proposed Framework (Hybrid)	0.891	0.885	0.888

The metadata-only approach increases the F1-Score to 0.852, showing that even simple structured metadata helps with sound contextual cues. The generative context-only model achieves an F1 Score of 0.873, demonstrating that LLM-generated contextual information provides a richer, more discriminative signal than just metadata. Finally, the hybrid model achieves the best performance, with an F1 score of 0.888. This proves that the combination of both types of contexts, metadata and generative context, brings complementary advantages, thereby essentially boosting performance.

B. Ablation Study: Impact of Contextualization Components

An ablation study was performed to better understand the specific contribution of each element within the architecture. The results highlighted in Table III represent performance gains in F1-score over the baseline PhoBERT model.

Table III: Performance Gains Measured by F1-Score

Configuration	F1-Score (Macro)	Improvement Over Baseline (Δ)
PhoBERT (Baseline - Context-Free)	0.822	-
+ Metadata Integration	0.852	+0.03
+ Generative Context Synthesis	0.873	+0.051
+ Combined (Both Strategies)	0.888	+0.066

Table III proves both contextualization strategies to be positive and significant. It is the metadata-based contextual keywords that yield an F1-Score increment of 0.03. This means that concrete “anchor points,” such as restaurant names and event types, effectively help the model resolve ambiguity. The generative contextualization strategy delivers an even greater lift (+0.051), underscoring that turning inputs into complete, coherent sentences enables the classifier to better leverage its acquired language understanding. Most striking is that merging the two strategies yielded a net gain of +0.066, which exceeds the sum of their separate contributions. The idea here is synergy: the structured metadata provides factual grounding, and the generated sentences add the necessary natural language context. Together, they form the best possible input for the classification model.

C. Qualitative Analysis: Case Studies on Ambiguous Reviews

Beyond quantitative measures, specific examples will go a long way toward explaining the proposed framework's

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

mode of operation and advantages. A case study of reviews that are extremely ambiguous or highly concise will therefore go a long way toward proving that this framework can use enriched contextual input, disparate structured metadata, and synthetically generated sentences to disambiguate sentiment and improve classification accuracy. These qualitative findings add to the empirical results, showing real-world use cases where context enrichment works best.

Table IV shows that short review sentences can cause the model to make incorrect sentiment predictions when used out of context. This table gives three examples in Vietnamese and English: each original short sentence, its modified version with contextual information added, the sentiment predicted without context (short), the correct sentiment when the context is considered, and a brief discussion of why the sentiment changes. The first example, "Ăn như ở Hà Nội" [Eating like in Hanoi], is a plain statement, no explicit sentiment at all. Out of context, the model would probably rate it as Neutral. With added contextual information about a summer beach trip where they dined at Bé Mặn restaurant, it is clear there was slight disappointment, which makes the statement fall under the Negative category.

The second example, "Không tệ." [Not bad.] is minimal and vague. Without any context, the predicted sentiment would be Neutral. However, adding information about Teacher's Day at Bún Chả Hương Liên clearly expresses mild satisfaction, hence a positive sentiment.

The third example, "Ngồi đây mát không chịu được" [Sitting here, it's too chilly], carries a description on its own with a Neutral prediction. When "sitting at Phở Thìn" is added as its context in "Hanoi's winter", the sentence expresses discomfort because of the cold and thus correctly shifts the sentiment to Negative.

Such results underscore the importance of contextual information about time, place, and user expectations for improving the accuracy of sentiment analysis from short reviews.

D. Error Analysis

While the hybrid approach performs best in addressing the context-deficit problem, misclassification still occurs. Several common examples are analysed below.

The first type of error mostly comes from long reviews. These express different aspects with variant sentiments. It becomes rather difficult for the model to synthesize and provide an adequate sentiment label. For instance, "… *có* nhiều món ngon, tuy nhiên hôm nay hơi mặn" [… There are many delicious dishes, but today it's a bit salty. It gives a positive sentiment about the general quality of the food, but expresses a negative sentiment about the specific flavour that day. The model may predict Neutral when the actual label is Negative (because the user's current experience is unsatisfactory).

Another significant error type is due to misspellings and unstandardized abbreviations by users, such as "Tôi khôn thích …" [I don't like] or "Phục_vụ k tôn_trọng khách_hàng …" [Service doesn't respect customers …]. These errors change both word and sentence semantics - for example, "khôn"/ "k" instead of "không" [not]. Thus, the proposed classification model will entirely misread the original intent in the review. To counter this, more powerful LLMs should be used to normalise misspellings and abbreviations during context generation, or special, super-accurate tools for correcting spelling should be integrated into the preprocessing step.

The third type of error is in short review sentences. Even with a context-enrichment strategy that generates longer sentences, short reviews do not always successfully highlight the positive or negative sentiment. For example, "Quán tại Hà Nội vào mùa xuân ăn ngon nhưng lâu" [The restaurant in Hanoi during spring is delicious but slow] has been misclassified as Neutral because the generated context does not offer sufficient emphasis on disappointment (because of waiting time) that the user wants to convey. It reflects an urge for more careful prompt design to enable LLMs to capture better the weight of the negative sentiment implied in the original review.

Table IV: The Impact of Contextual Information on Sentiment Prediction in Short Food Reviews

Review Sentence	Modified Sentence with Context	Wrong Predicted Sentiment (Without Using Context)	Correct Predicted Sentiment (Using Context)	Discussion
(In Vietnamese) Ăn như ở Hà Nội. (Translation) Eating like in Hanoi.	**Context:** Location: Near the beach, Season: Summer, Restaurant: Bé Mặn. (In Vietnamese) Du lịch biển hè này, ăn hải sản ở Bé Mặn gần biển mà tôi lại có cảm giác như đang ăn ở Hà Nội. (Translation) During this summer beach trip, eating seafood at Bé Mặn near the sea, I felt as if I were eating in Hanoi.	Neutral	**Negative**	The short sentence lacks explicit sentiment; the context reveals mild disappointment relative to expectations.
(In Vietnamese) Không tệ. (Translation) Not bad.	**Context:** Location: Hà Nội, Season: Autumn, Event: Teacher's Day 20/11 ở Hà Nội, Restaurant: Bún chả Hương Liên. (In Vietnamese) Trong dịp lễ Nhà giáo Việt Nam 20/11, ăn bún chả ở Hương Liên tôi thấy không tệ. (Translation) During Teacher's Day, 20/11 in Hanoi, I tried the bún chả at Hương Liên and found it not bad.	Ne utral	**Positive**	Minimal sentence; context provides temporal and experiential cues, indicating mild satisfaction.
(In Vietnamese) Ngồi đây mát không chịu được. (Translation) Sitting here, it's too chilly.	**Context:** Location: Hà Nội, Season: Winter, Restaurant: Phở Thìn. (In Vietnamese) Ngồi trong Phở Thìn giữa mùa đông Hà Nội, tôi cảm thấy lạnh đến mức không chịu được. (Translation) Sitting in Phở Thìn during Hanoi's winter, I felt so cold that it was unbearable.	Neutral	**Negative**	The sentence alone is descriptive; context shows discomfort due to the cold, shifting the sentiment to negative.

© Copyright: All rights reserved.

VI. CONCLUSION

This work has shown that contextual enrichment of short reviews substantially improves sentiment classification performance on Vietnamese food reviews. By using structured metadata and synthetically generated contextual sentences, ambiguous or extremely short user inputs could be better understood.

Experimental results show that contextualizing the input sentence boosts model prediction performance. The baseline method, which simply adds metadata containing contextual information, also yields a slight performance increase. A significant performance improvement is observed when LLMs are used to generate a sentence given the existing context and the input sentence. And the best performance comes when both of these strategies are used together. Complementary qualitative analyses reaffirm this framework's ability to reduce linguistic ambiguity, offering practical insights into real-world applications.

These results underscore the crucial role of context-aware preprocessing in short-text sentiment analysis. Next, we plan to apply the model in other domains and continue our investigation into more sophisticated generative techniques.

DECLARATION STATEMENT

As the article's author, I must verify the accuracy of the following information after aggregating input from all authors.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. H. T. M. Nguyen et al., "VLSP SHARED TASK: SENTIMENT ANALYSIS," J. Comput. Sci. Cybern., vol. 34, no. 4, pp. 295–310, Jan. 2019, DOI: https://doi.org/10.15625/1813-9663/34/4/13160.
2. T. K. Tran and T. T. Phan, "Capturing Contextual Factors in Sentiment Classification: An Ensemble Approach," IEEE Access, vol. 8, pp. 116856–116865, 2020, DOI: http://doi.org/10.1109/ACCESS.2020.3004180.
3. D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in Findings of the Association for Computational Linguistics: EMNLP 2020, Online: Association for Computational Linguistics, 2020, pp. 1037–1042. DOI: http://doi.org/10.18653/v1/2020.findings-emnlp.92.
4. K. T. Tran, M. H. Dinh, T. N. N. Tran, V. X. Nguyen, and T. T. B. Thi, "Exploring the potential of graph neural networks for Vietnamese sentiment analysis," HUFLIT Journal of Science, vol. 9, no. 2, pp. 11–11, 2025. url: https://hjs.huflit.edu.vn/index.php/hjs/article/view/266L.
5. Mei et al., "A Survey of Context Engineering for Large Language Models," 2025, arXiv. DOI: http://doi.org/10.48550/ARXIV.2507.13334.
6. H. Q. V. Vo and K. Yamamoto, "VietSentiLex: a sentiment dictionary that considers the polarity of ambiguous sentiment words," in Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, 2018. url: http://doi.org/https://aclanthology.org/Y18-1081.pdf
7. T. Kim Phung, N. A Te, and T. Thi Thu Ha, "A machine learning approach for opinion mining online customer reviews," in 2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter), Ho Chi Minh City, Vietnam: IEEE, Jan. 2021, pp. 243–246. DOI: http://doi.org/10.1109/SNPDWinter52325.2021.00059.
8. T. Nguyen-Thanh and G. T. C. Tran, "Vietnamese Sentiment Analysis for Hotel Review based on Overfitting Training and Ensemble Learning," in Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019, Hanoi, Ha Long Bay, Viet Nam: ACM Press, 2019, pp. 147–153. DOI: http://doi.org/10.1145/3368926.3369675.
9. E. B. Ramezani, "Sentiment analysis applications using deep learning advancements in social networks: A systematic review," Neurocomputing, vol. 634, p. 129862, June 2025, DOI: http://doi.org/10.1016/j.neucom.2025.129862.
10. L. S. Le, D. V. Thin, N. L.-T. Nguyen, and S. Q. Trinh, "A Multi-filter BiLSTM-CNN Architecture for Vietnamese Sentiment Analysis," in Advances in Computational Collective Intelligence, vol. 1287, M. Hernes, K. Wojtkiewicz, and E. Szczerbicki, Eds., in Communications in Computer and Information Science, vol. 1287., Cham: Springer International Publishing, 2020, pp. 752–763. DOI: http://doi.org/10.1007/978-3-030-63119-2_61.
11. D. Van Thin, D. N. Hao, and N. L.-T. Nguyen, "Vietnamese Sentiment Analysis: An Overview and Comparative Study of Fine-tuning Pretrained Language Models," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 22, no. 6, pp. 1–27, June 2023, DOI: http://doi.org/10.1145/3589131.
12. D. Q. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, "PhoGPT: Generative Pre-training for Vietnamese," Mar. 22, 2024, arXiv: arXiv:2311.02945. DOI: http://doi.org/10.48550/arXiv.2311.02945.
13. G. Team et al., "Gemma: Open Models Based on Gemini Research and Technology," Apr. 16, 2024, arXiv: arXiv:2403.08295. DOI: http://doi.org/10.48550/arXiv.2403.08295.
14. OpenAI et al., "gpt-oss-120b & gpt-oss-20b Model Card," Aug. 08, 2025, arXiv: arXiv:2508.10925. DOI: http://doi.org/10.48550/arXiv.2508.10925.
15. H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," Feb. 27, 2023, arXiv: arXiv:2302.13971. DOI: http://doi.org/10.48550/arXiv.2302.13971
16. A. Q. Jiang et al., "Mistral 7B," oct. 10, 2023, arXiv:2310.06825. Doi: http://doi.org/10.48550/arXiv.2310.06825.
17. F. Man et al., "Context-Aware Sentiment Forecasting via LLM-based Multi-Perspective Role-Playing Agents," May 30, 2025, arXiv: arXiv:2505.24331. DOI: http://doi.org/10.48550/arXiv.2505.24331.
18. D. Van Thin, D. N. Hao, and N. L.-T. Nguyen, "Prompt Engineering with Large Language Models for Vietnamese Sentiment Classification," in Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, 2024, pp. 181–192. url: http://doi.org/https://aclanthology.org/2024.paclic-1.17/

AUTHOR'S PROFILE



Dr. Nguyen Thi Thanh Thuy has been on the faculty of the Posts and Telecommunications Institute of Technology (PTIT), Vietnam, for over two decades, teaching and providing academic mentoring in Computer Science. She earned her PhD in 2023 and is a member of the Faculty of Information Technology and the Machine Learning and Applications Laboratory at PTIT. Her interests fall under Artificial Intelligence, Machine Learning, and Natural Language Processing, more specifically, language modelling, semantic representation, and applied AI systems. Dr. Thuy's practice and research approach highlight NLP techniques toward actual problem-solving. Highly recognised for her profound knowledge and continuous commitment to academic advancement, she has consistently inspired students and young researchers to pursue further study in Artificial Intelligence. Machine Learning and Intelligent Information Systems.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



Dr. Nguyen Ngoc Diep is a faculty member at Posts and Telecommunications Institute of Technology (PTIT), Vietnam. He earned his PhD in Information Systems in 2017. Diep has been with PTIT since 2012 as a lecturer and researcher whose interests include Natural Language Processing, Machine Learning, data-centric AI systems—context-aware models, sentiment analysis frameworks, and generative NLP architecture for practical applications—and the integration of core NLP themes with cybersecurity for multilingual text stream attack detection, malicious communication intent classification, and behavioural threat intelligence modelling. Dr. Diep is renowned as an expert in applied machine learning, model fine-tuning, and practical AI methodologies, offering mentorship to students and researchers conducting high-level work in NLP, intelligent systems, or AI-driven information security.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.