ViDia2Std: A Parallel Corpus and Methods for Low-Resource Vietnamese Dialect-to-Standard Translation

Khoa Anh Ta[1, 2], **Nguyen Van Dinh**[1, 2], **Kiet Van Nguyen**[1, 2*]
[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{21522232, 20520657}@gm.uit.edu.vn, kietnv@uit.edu.vn

Abstract

Vietnamese exhibits extensive dialectal variation, posing challenges for NLP systems trained predominantly on standard Vietnamese. Such systems often underperform on dialectal inputs, especially from underrepresented Central and Southern regions. Previous work on dialect normalization has focused narrowly on Central-to-Northern dialect transfer using synthetic data and limited dialectal diversity. These efforts exclude Southern varieties and intra-regional variants within the North. We introduce ViDia2Std, the first manually annotated parallel corpus for dialect-to-standard Vietnamese translation covering all 63 provinces. Unlike prior datasets, ViDia2Std includes diverse dialects from Central, Southern, and non-standard Northern regions often absent from existing resources, making it the most dialectally inclusive corpus to date. The dataset consists of over 13,000 sentence pairs sourced from real-world Facebook comments and annotated by native speakers across all three dialect regions. To assess annotation consistency, we define a semantic mapping agreement metric that accounts for synonymous standard mappings across annotators. Based on this criterion, we report agreement rates of 86% (North), 82% (Central), and 85% (South). We benchmark several sequence-to-sequence models on ViDia2Std. mBART-large-50 achieves the best results (BLEU 0.8166, ROUGE-L 0.9384, METEOR 0.8925), while ViT5-base offers competitive performance with fewer parameters. ViDia2Std demonstrates that dialect normalization substantially improves downstream tasks, highlighting the need for dialect-aware resources in building robust Vietnamese NLP systems.

Code — https://github.com/biuinvincible/ViDia2Std.git
Datasets —
https://huggingface.co/datasets/Biu3010/ViDia2Std

1 Introduction

Vietnamese dialects pose significant challenges for NLP systems due to the mismatch between regionally diverse inputs and the standardized data on which most models are trained. This problem is especially acute in low-resource settings, where dialectal inputs are underrepresented and lack high-quality annotated corpora.

Although we refer to the task as *dialect-to-standard translation* in the title of this paper—to remain consistent with prior literature and emphasize the parallel nature of our dataset—we adopt the term *dialect normalization* throughout the paper. This reflects our interpretation of the task as a preprocessing step that transforms non-standard regional language into standardized Vietnamese, enabling downstream models to process dialectal input more effectively. In the Vietnamese context, where dialects are variations within a single language rather than mutually unintelligible systems, normalization better captures the linguistic nature of this transformation.

As Alam and Anastasopoulos (2025) observe, "NLP models trained on standardized language data often struggle with variations". This is particularly true for Vietnamese, which comprises three major dialect groups—Northern, Central, and Southern—that differ substantially in phonology, vocabulary, and syntax (Le and Luu 2023). Standard Vietnamese is based on the Northern dialect, and most existing Vietnamese NLP models (e.g., PhoBERT, BARTpho) are trained predominantly on this variety. As a result, these models often fail to interpret utterances containing region-specific vocabulary and expressions.

Table 1 illustrates the performance of both commercial translation systems and state-of-the-art language models (e.g., Claude Sonnet 4, Gemini-2.5-Flash, Google Translate, Bing Translator) when handling dialectal Vietnamese inputs. These models exhibit significant errors or misunderstandings, highlighting the urgent need for effective dialect normalization as a preprocessing step to improve translation and understanding.

Several prior efforts have attempted to address this problem. Notably, Le and Luu (2023) constructed a Central-to-Standard Vietnamese parallel corpus. However, their data was collected exclusively from speakers in Ha Tinh province, thus limiting the geographical and lexical diversity represented. Central dialects vary widely—provinces like Thua Thien Hue, Quang Tri, and Thanh Hoa each contribute distinct vocabulary and syntactic patterns. Moreover, regional variation is not confined to the Central region: many provinces in both the North and South exhibit distinct expressions that challenge NLP tools trained solely on the Northern standard.

To overcome these limitations, we introduce **ViDia2Std**,

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

	Dialect	***Normalized***	***Gold Translation***
	Gấy tau nỏ biết mi mô	Vợ tao không biết mày đâu	
Gemini-2.5-flash	I don't know where you are.	My wife doesn't know you.	My wife doesn't know you
Claude Sonnet 4	I don't know what you're saying	My wife doesn't know you	
Google Translate	Cause I don't know you	My wife doesn't know you.	
Bing Translator	I don't know what you are talking about	My wife doesn't know you.	

Table 1: Comparison of Translation Models on Dialectal and Normalized Data.

the first large-scale Vietnamese dialect normalization corpus with nationwide coverage. ViDia2Std consists of over 13,000 manually aligned sentence pairs drawn from real user comments collected from all 63 provinces of Vietnam, ensuring diverse representation across Northern, Central, and Southern dialects. A rigorous annotation protocol involving native speakers from each dialect region ensures high-quality standardization.

In addition to releasing this corpus, we establish strong baselines using modern sequence-to-sequence models such as BARTpho and ViT5, fine-tuned for the dialect normalization task. These models demonstrate effective handling of dialectal variation and offer reference results for future research.

Importantly, we show that dialect normalization leads to substantial improvements on downstream tasks. For example, sentiment classification accuracy on dialectal inputs increases from 51% to 62% after normalization. Similarly, normalized inputs improve the performance of machine translation systems and large language models. These findings confirm that dialect normalization is a crucial preprocessing step for robust Vietnamese NLP.

Our contributions are as follows:

- **ViDia2Std corpus:** We release a high-quality, manually annotated Vietnamese dialect-to-standard parallel corpus covering all major dialect regions. The data—sourced from public Facebook comments nationwide—far exceeds prior resources in both size and dialectal diversity.
- **Baseline normalization models:** We evaluate several neural sequence-to-sequence architectures (e.g., BARTpho, ViT5) on the dialect normalization task, providing reference benchmarks for future work.
- **Downstream evaluation:** We demonstrate that dialect normalization improves core NLP tasks. For instance, our normalization models raise sentiment-analysis F1 (Weighted) from 0.52 to 0.63 and significantly enhance translation quality, confirming the benefit of integrating normalization into Vietnamese NLP pipelines.

2 Related Work

Text normalization has been extensively studied in historical text restoration (Bollmann 2019), progressing from rule-based systems to neural seq2seq models. In Vietnamese, ViLexNorm (Nguyen, Le, and Nguyen 2024) introduced a 10K-sentence corpus for standardizing informal online text, improving downstream tasks like POS tagging. However, dialect normalization is more complex, involving systematic variation in vocabulary, syntax, and phonology. Globally, recent work treats dialect normalization as character- or byte-level transduction (Kuparinen, Miletić, and Scherrer 2023), or multi-dialect machine translation (Abe et al. 2018), with growing use of adapters for robustness in low-resource dialects (Held, Ziems, and Yang 2023). Ibn Alam et al. (Alam and Anastasopoulos 2025) show that fine-tuning open LLMs with small dialect corpora can significantly boost BLEU scores. In Vietnamese, Le and Luu (2023) proposed a parallel corpus for Central-to-North normalization and showed BARTpho outperforms multilingual models, but their dataset is limited in size and regional diversity.

Our Contribution. We introduce **ViDia2Std**, the first nationwide, manually annotated Vietnamese dialect-to-standard corpus covering all 63 provinces and three major dialect zones. Our work significantly scales previous efforts and evaluates multiple state-of-the-art normalization models. Beyond intrinsic evaluation, we conduct extrinsic tests on sentiment analysis and translation, confirming that normalization leads to large performance gains in real-world Vietnamese NLP applications

3 Dataset Construction

In this section, we describe the process of constructing our Vietnamese dialect-to-standard parallel corpus. The overall data creation pipeline is illustrated in Figure 1.

3.1 Data Collection

We construct a parallel corpus (dialect–standard) by harvesting user-generated content from Facebook. Our collection strategy involved identifying potential news fanpages for all 63 provinces of Vietnam. **We prioritized fanpages managed by individuals over those managed by local authorities, as the former exhibited a significantly higher rate of user interaction using local dialects**. Using Python scripts with the Selenium library, we then collected data. The target of 350-500 posts per province was a guideline, not a rigid quota, as fanpage activity varied greatly. **If a province yielded too few comments, we actively searched for other relevant local groups and fanpages to ensure fair representation and avoid bias toward any single province**.

Facebook is ideal because it is widely used across Vietnam (over 70% of the population on social media in 2022 (Nguyen, Le, and Nguyen 2024)) and hosts open groups where users naturally write in local dialects. In line

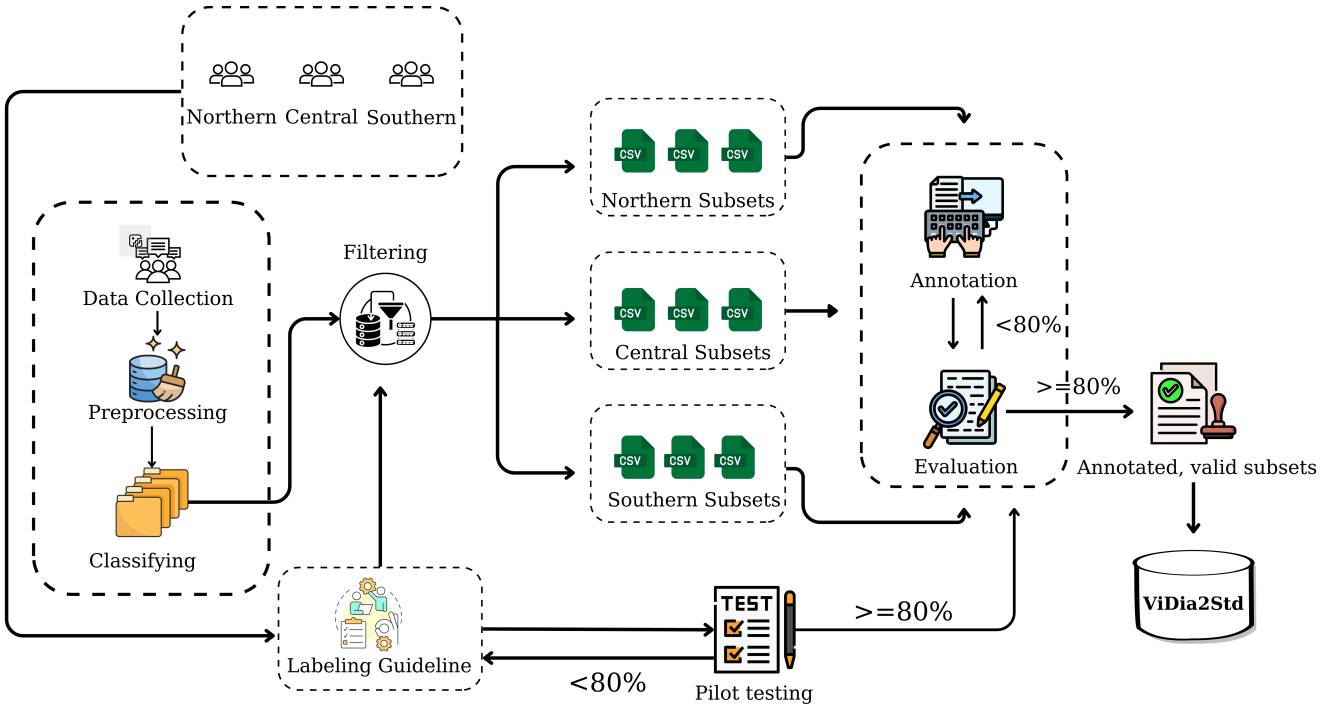

Figure 1: Overview of our dataset creation pipeline, from raw social media data to the final dialect-standard parallel corpus.

with prior work in other languages, we follow a "web-as-corpus" approach to dialectology (Burghardt, Granvogl, and Wolff 2016; Alshutayri and Atwell 2019). For example, Alshutayri and Atwell built a multi-dialect Arabic corpus from Facebook and Twitter (Alshutayri and Atwell 2019), and Burghardt et al. used Bavarian Facebook posts to compile a dialect lexicon (Burghardt, Granvogl, and Wolff 2016). Inspired by (Le and Luu 2023), who prepared scripted dialogues to elicit Central/Northern Vietnamese variants, we instead use authentic Facebook comments as prompts. Native speakers take each comment and restate it in an accurate, cleaned dialectal form (see Annotation below), then map it to standard Vietnamese. In this way we leverage real multilingual user data without incurring the cost of eliciting responses for all provinces.

3.2 Data Preprocessing and Dialect Filtering

The raw social media data was highly unstructured and noisy, containing emojis, URLs, non-standard text, and metadata. To extract clean dialectal content suitable for annotation, we applied a **two-stage pipeline**:

Stage 1: Automatic Preliminary Preprocessing. We first applied an automatic script to de-noise the raw comments. This involved several key steps:

- Converting all text to lowercase.
- Removing non-linguistic content (e.g., empty comments, emojis, stickers, repeated characters).
- Stripping metadata (e.g., URLs, @mentions, #hashtags).
- Normalizing common social media language and *teencode* using a manually curated dictionary (e.g., "ko" –> "khong", "ae" –> "anh em").

This stage was designed to reduce noise and ease the subsequent human annotation burden.

Stage 2: Annotator-Driven Dialect Filtering. The cleaned data from Stage 1 was then presented to our annotator team for dialect filtering. This step was crucial for isolating dialect-rich sentences from standard Vietnamese. Annotators used a set of **region-specific keyword lists**, prepared by our team based on linguistic analysis, to identify and extract comments with a high probability of containing dialect. This two-stage pipeline yielded a cleaner, dialect-rich subset suitable for the main annotation task (described in the next subsection).

3.3 Annotation Protocol

We recruited nine native Vietnamese annotators, purposefully selected to represent the three major dialect regions: North (2), Central (4), and South (3). The Central region was assigned more annotators due to its higher internal lexical diversity. All annotators received basic linguistic training and underwent cross-dialectal familiarization via dialect blogs, glossaries, and curated social media content.

Dialect normalization is treated as a lexical-semantic mapping task rather than free-form paraphrase. Each sentence is annotated in three stages:

1. **Dialect Cleaning:** Annotators first fix orthographic issues (e.g., typos, abbreviations, missing punctuation) while retaining all dialectal lexical and syntactic features.
2. **Dialect-to-Standard Mapping:** Dialectal tokens are translated into natural standard Vietnamese, preserving

meaning, tone, and sentence structure.

3. **Ambiguity Flagging:** Any cases that are idiomatic, ambiguous, or difficult to align are flagged for collaborative team discussion.

To assist with rare or ambiguous cases, annotators are provided with dialect dictionaries, regional glossaries, and prior annotated corpora. This hybrid approach—combining native intuition with structured resources—helps ensure semantic fidelity across diverse dialectal input. Annotators were compensated at a rate of $0.038 (1,000 VND) per sentence pair, broken down as $0.019 for the manual normalization stage and $0.019 for the dialect-to-standard labeling stage.

3.4 Annotator Training and Quality Control

To maximize regional coverage, annotators from each dialect zone were exposed to lexical variation across neighboring provinces. For instance, Central annotators from Nghe An were introduced to features from Quang Tri and Hue to enable broader dialectal understanding. This intra-regional calibration improves coverage of localized vocabulary (e.g., "bui" for "vui" in Quang Tri) and enhances consistency within each macro-region.

Prior to full-scale annotation, we conducted a pilot round in which all annotators collectively annotated a shared set of 100 dialectal sentences. No single annotator or gold reference was used for comparison. Instead, we adopt a strict group-level semantic consistency criterion:

A sentence is marked as "agreed" if and only if all annotators produce semantically equivalent mappings for every dialectal token. Semantic equivalence is defined using curated synonym sets and reviewed during collaborative sessions.

Formally, let N be the total number of pilot sentences, and K be the number of annotators. For sentence i, each dialectal token b_j is mapped to a normalized form $c_j^{(k)}$ by annotator k. The sentence is considered semantically agreed upon if:

$$\text{Agree}(i) = \begin{cases} 1, & \text{if } \forall j,\ \forall k_1, k_2,\ c_j^{(k_1)} \sim c_j^{(k_2)} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The overall **Strict Semantic Group Agreement (SSGA)** score is then computed as:

$$\text{SSGA} = \frac{1}{N} \sum_{i=1}^{N} \text{Agree}(i) \quad (2)$$

Annotators were only allowed to proceed with full annotation if the group achieved an SSGA of at least 80% on the pilot set. If this threshold was not met, *all annotators were retrained collectively*, with clarification of disagreements and refinement of dialectal interpretation rules.

In the final corpus, group-level agreement scores reached 86% (North), 82% (Central), and 85% (South), indicating high internal consistency. The strict nature of SSGA makes it more demanding than pairwise or majority-vote metrics but better reflects full-group semantic convergence.

3.5 Corpus Summary and Release

The resulting corpus comprises 13,657 dialect–standard sentence pairs, each aligned at the sentence level. One column contains the cleaned dialectal sentence, and the other its semantically equivalent form in standard Vietnamese. Examples of these sentence pairs are illustrated in Table 2.

The dataset will be released for non-commercial research purposes. We hope this resource will support future work in dialect-aware modeling, low-resource machine translation, and inclusive Vietnamese NLP.

Dialect: "Thất nghiệp bơ về chơ ăn Tết mô giờ ni."
Standard: "Thất nghiệp nên về chứ ăn Tết đâu giờ này."
Dialect: "Hồi nhỏ miềng đi chự bò thì gặp."
Standard: "Hồi nhỏ mình đi giữ bò thì gặp."
Dialect: "Chị muốn dìa là dìa chớ ước gì chèn?"
Standard: "Chị muốn về là về chứ ước gì trời?"

Table 2: Sample sentence pairs: dialect vs. standard

Region	Number of Sentences	Avg. Words	
		Dialect	Standard
Central	9,033	10.55	10.82
Northern	1,054	9.31	9.28
Southern	3,570	10.58	11.02
Total	**13,657**	**10.46**	**10.77**

Table 3: Basic Dataset Statistics

4 Intrinsic Evaluation

4.1 Task Definition

Following the approach of Le and Luu (2023), we define Vietnamese dialect normalization as a conditional sequence generation task. Given an input sentence $x = [x_1, x_2, \ldots, x_n]$ in dialectal Vietnamese, the model generates an equivalent sentence $y = [y_1, y_2, \ldots, y_m]$ in standard Vietnamese by optimizing the conditional probability $P(y|x)$:

$$L = -\sum_{t=1}^{m} \log P(y_t \mid y_{<t}, x; \theta) \quad (3)$$

where θ denotes model parameters. The objective is to preserve both meaning and pragmatic intent (e.g., interrogatives, imperatives), rather than perform simple word-level substitution.

4.2 Experimental Setup

This subsection outlines the foundational elements of the experimental methodology, including the selection of models, the metrics employed for evaluation, and the specific training configurations.

Models for Dialect Normalization We benchmark five sequence-to-sequence models fine-tuned for Vietnamese text generation:

- **BARTpho-word-base** (Nguyen Luong Tran and Duong Le and Dat Quoc Nguyen 2022): A Vietnamese BART model with word-level tokenization, capturing high-level grammatical structures.
- **BARTpho-syllable-base** (Nguyen Luong Tran and Duong Le and Dat Quoc Nguyen 2022): A variant using syllable-level tokenization, better suited for Vietnamese's monosyllabic nature and dialectal variations.
- **ViT5-base** (Phan et al. 2022): A Vietnamese version of T5, treating all tasks as text-to-text generation; pretrained on diverse NLP tasks for strong generalization.
- **Vietnamese-correction-v2** (bmd1905 2024): A BARTpho-syllable model trained for spelling correction; included to explore overlap with normalization tasks.
- **mBART-large-50** (Liu et al. 2020; Tang et al. 2020): A multilingual model trained on 50 languages via denoising objectives, offering strong robustness to diverse input forms including dialects.

Evaluation Metrics We evaluate model outputs using BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), WER, and CER—capturing both surface-form accuracy and semantic preservation. This diverse metric set ensures reliable assessment for the nuanced task of dialect normalization.

Training Configuration All models were trained using HuggingFace Transformers (Wolf et al. 2020). The ViDia2Std corpus was split into 10870 samples for training, 1184 for development, and 1603 for testing.

To ensure a rigorous and fair comparison, **we applied an identical hyperparameter configuration across all models**. We used `AutoTokenizer` with a maximum sequence length of 50 tokens—sufficient given the average sentence length ($<$11 tokens). Training utilized the AdamW optimizer with a batch size of 32, a learning rate of 2e-5, and a maximum of 10 epochs, using mixed precision (`fp16=True`) on a single A100 GPU.

Models were evaluated and checkpointed each epoch, with early stopping (`patience=1`) based on BLEU improvements $\geq$0.01. Crucially, to ensure strict reproducibility, **all models in this camera-ready version were retrained with the exact same fixed random seed of 42**. Consequently, the reported metrics may exhibit minor variations compared to preliminary runs, reflecting a more stable and reproducible convergence. **Importantly, we verified that these slight variations did not affect the results or conclusions of the subsequent experiments.**

Results and Discussion Table 4 summarizes the performance of the five sequence-to-sequence models on the dialect normalization task using the test set. All models demonstrate good performance, but notable differences in effectiveness are observed.

As shown in Table 4, **mBART-large-50** consistently outperforms all baselines across lexical and semantic metrics, confirming its strength for dialect normalization viewed as an intra-lingual translation task.

ViT5-base delivers comparable performance while using less than half the parameters, highlighting its parameter efficiency and suitability for real-world deployment.

BARTpho variants and **Vietnamese-correction-v2** yield moderate results, with the latter showing that pretraining on general correction tasks still transfers well.

Overall, **mBART-large-50** provides the best performance, while **ViT5-base** offers a strong trade-off between accuracy and model size.

5 Extrinsic Evaluation

Beyond intrinsic metrics, extrinsic evaluations were conducted to demonstrate the practical utility of dialect normalization as a preprocessing step for real-world NLP tasks. These experiments assess whether normalization improves language processing at a semantic level, especially when dealing with non-standard Vietnamese inputs.

5.1 Machine Translation

This section evaluates whether normalizing dialectal Vietnamese before translation improves English MT quality, under the hypothesis that standardization benefits systems trained on standard language.

Experimental Design Two parallel translation pipelines were designed for a dataset of 600 dialectal sentences, equally distributed across Northern, Central, and Southern regions (200 sentences per region), manually selected from ViDia2Std's test set. Each dialectal sentence had a human-translated English reference for evaluation.

- **Direct Translation**: Dialectal sentences were translated directly into English using an MT system.
- **Translation after Normalization**: Dialectal sentences were first normalized to standard Vietnamese using the trained normalization model, then translated by the MT system.

Evaluation Tool and Metric Traditional metrics (e.g., BLEU, ROUGE, METEOR) often fail to capture meaning preservation, especially in cases of paraphrasing or dialectal variation. We therefore adopt an *LLM-as-a-Judge* approach, using Gemini 2.5 Flash to evaluate translations on two criteria: *semantic completeness* and *pragmatic accuracy*, producing a binary label: `ACCEPTED` or `UNACCEPTED`.

This approach is supported by recent work: Sun et al. (2025) find that BLEU underrepresents improvements in discourse coherence; Kocmi and Federmann (2023) show that GPT-based metrics (e.g., GEMBA) better align with human judgments than COMET; and Fernandes et al. (2025)

Model	ROUGE-L	BLEU	METEOR	WER	CER	Parameters
BARTpho-word-base	0.9167	0.7601	0.8625	0.1527	0.1049	150M
ViT5-base	0.9300	0.7934	0.8802	0.1340	0.0876	310M
BARTpho-syllable-base	0.9218	0.7597	0.8627	0.1513	0.1045	132M
Vietnamese-correction-v2	0.9257	0.7723	0.8746	0.1416	0.0988	396M
mBART-large-50	**0.9384**	**0.8166**	**0.8925**	**0.1226**	**0.0754**	611M

Table 4: Intrinsic Evaluation Results

propose LLM-based QA to assess semantic retention directly.

These studies underscore a key limitation of traditional metrics—their poor handling of non-literal or creative translations (Li et al. 2025; Patil, Tao, and Jadon 2025). In contrast, LLMs offer contextual reasoning closer to human evaluators.

We report the **Acceptance Rate** as our primary metric:

$$\text{Acceptance Rate} = \frac{|\text{Accepted Samples}|}{|\text{Total Samples}|} \times 100\% \quad (4)$$

Results and Discussion As presented in Table 5, the machine translation evaluation results confirm that normalization improves translation quality across all systems. The most substantial gain was observed for Kimi-K2-Instruct (+12.84 percentage points), while even the top-performing system, Gemini 2.0 Flash, saw its acceptance rate increase from 61.83% to 67.00%. This demonstrates that normalization is a universally beneficial preprocessing step, reducing ambiguity for both commercial APIs and large language models.

Cross-System Consistency Analysis To provide a more robust evaluation, this section analyzes the consistency of normalizationś impact across all six translation systems. The trend in Table 6 is clear: the positive impact of normalization is more consistent than its negative effects. As the agreement threshold increases, the ratio of improved-to-worsened sentences grows significantly, from 1.70 for at least one system to 32:1 for at least four systems. Most notably, 3 sentences were unanimously improved by all six systems, while none were unanimously worsened. This provides strong evidence that the benefits are systematic and not system-specific artifacts, underscoring the reliable value of the normalization process.

Per-System Impact Analysis As detailed in Table 7, the improvement-to-worsening ratio varies across systems. Traditional services like Microsoft Azure AI Translator show a high ratio (4.00), indicating strong reliance on standard input. In contrast, the top-performing Gemini 2.0 Flash has the lowest ratio (1.66), suggesting it is more robust to dialectal variations but can still be adversely affected by suboptimal normalizations. DeepSeek-V3 shows a performance comparable to Googleś translation service, with an improvement-to-worsening ratio of 2.04.

In-depth Analysis of Translation Regressions While the aggregate results show a clear net benefit, a granular analysis of the regression cases—where a translation was downgraded from `ACCEPTED` to `UNACCEPTED` post-normalization—is essential. We manually analyzed all 252 instances of regression across the six translation systems to categorize the root cause of the failure. The results, summarized in Table 8, reveal that the regressions are not primarily caused by the normalization process itself, but by limitations in the downstream components of the pipeline.

Our analysis identifies three primary sources of error, with their new ranking:

1. **MT Model Fragility (46.8%):** The largest source of regression involves failures of the downstream MT system. Even when provided with a perfectly standard Vietnamese sentence, the MT system sometimes produced a translation that was significantly worse than the one generated from the original dialect. These failures suggest that MT systems may be overfitted to certain linguistic patterns and lack the robustness to handle valid, albeit less common, phrasings.
2. **LLM Evaluator Noise (35.7%):** The second-largest category is the inherent stochasticity and phrasal sensitivity of the LLM-as-a-Judge. In numerous cases, the direct and normalized translations were semantically identical, yet received different verdicts. This highlights a key challenge in automated evaluation: the judgeś preference for a specific phrasing can be misinterpreted as a difference in quality.
3. **Normalization Model Error (17.5%):** The normalization model itself was the least frequent source of regression. These errors typically occurred when the model failed to preserve nuanced semantics.

Crucially, the fact that over 82% of regressions are attributable to MT model fragility and evaluator noise provides strong evidence for the reliability of our normalization approach. These regressions are not systematic flaws but rather isolated artifacts of the current evaluation and translation pipeline. This reinforces the conclusion from our consistency analysis (Table 6): the benefits of normalization are systematic, while the regressions are largely system-specific and non-systematic.

5.2 Sentiment Analysis

To assess the impact of dialect normalization on sentiment analysis performance, a pre-trained Vietnamese sentiment analysis model was evaluated before and after applying dialect normalization. This design aimed to quantify the improvement in the model's ability to correctly understand the emotional tone of dialectal utterances.

Method	Total Samples	Accepted	Unaccepted	Acceptance Rate (%)
Microsoft Azure AI Translator	600	68	532	11.33
Microsoft Azure AI Translator + Normalized	600	134	466	22.33
Google Cloud Translation – Basic	600	230	370	38.33
Google Cloud Translation – Basic + Normalized	600	275	325	45.83
DeepSeek-V3	600	327	273	54.50
DeepSeek-V3 + Normalized	600	379	221	63.17
Kimi-K2-Instruct	600	308	292	51.33
Kimi-K2-Instruct + Normalized	600	385	215	64.17
Gemini 1.5 Flash	600	282	318	47.00
Gemini 1.5 Flash + Normalized	600	342	258	57.00
Gemini 2.0 Flash	600	371	229	61.83
Gemini 2.0 Flash + Normalized	600	402	198	**67.00**

Table 5: Machine Translation Evaluation Results (Overall)

Threshold ($\geq$)	Improved	Worsened	Ratio
1	316	186	1.70
2	152	48	3.17
3	69	17	4.06
4	32	1	32.00
5	11	0	∞
6	3	0	∞

Table 6: Consistency of Normalization Impact Across 6 Systems

System	Improved	Worsened
Microsoft Azure AI Translator	88	22
Kimi-K2-Instruct	125	48
Gemini 1.5 Flash	102	42
Google Cloud Translation	88	43
DeepSeek-V3	102	50
Gemini 2.0 Flash	78	47

Table 7: Per-System Sentence Improvement vs. Worsening

Error Category	Count (%)	Description
MT Model Fragility	118 (46.8%)	The MT system fails to translate a correctly normalized sentence.
LLM Evaluator Noise	90 (35.7%)	The evaluator provides inconsistent ratings for translations of equivalent quality.
Normalization Model Error	44 (17.5%)	The normalization model incorrectly alters the meaning of the source sentence.
Total	**252 (100%)**	

Table 8: Categorization of Translation Regression Causes Across All Systems (percentages shown in parentheses next to counts)

Experimental Design A test set of dialectal sentences was used, which were automatically labeled for sentiment (Positive, Negative, Neutral) using Gemini 2.0 Flash and deepseek-ai/DeepSeek-V3. The automatic labeling achieved a Cohen's kappa of 0.7082, which is considered acceptable and indicative of substantial agreement on a 300-sample consensus set. A pre-trained sentiment analysis model (5CD-AI/Vietnamese-Sentiment-visobert (5CD-AI 2024)) was then applied to classify sentiment under two scenarios: (1) direct analysis on original dialectal input, and (2) analysis on input normalized by the model.

Evaluation Metrics Standard classification metrics were used: Accuracy, Precision, Recall, and F1-score.

Results and Discussion Table 9 demonstrates that dialect normalization substantially improves sentiment analysis performance. Accuracy increased from 50.59% to 62.13%, and macro F1 rose from 0.48 to 0.58.

Performance improved across all sentiment classes. F1-score for NEGATIVE rose from 0.59 to 0.72, NEUTRAL from 0.37 to 0.45, and POSITIVE from 0.47 to 0.58. These gains indicate that dialect normalization helps disambiguate regional lexical forms that often confuse standard models, especially for nuanced sentiment expressions.

Normalization corrected a significant number of prediction errors, with the majority in the NEGATIVE and NEUTRAL categories. This confirms that dialectal ambiguity most affects less polarized sentiments.

	Before			After		
Class	P	R	F1	P	R	F1
NEGATIVE	0.89	0.45	0.59	**0.86**↓	**0.62**↑	**0.72**↑
NEUTRAL	0.47	0.30	0.37	**0.48**↑	**0.42**↑	**0.45**↑
POSITIVE	0.32	0.87	0.47	**0.44**↑	**0.82**↓	**0.58**↑
Accuracy		0.51			**0.62**↑	
Macro Avg	0.56	0.54	0.48	**0.59**↑	**0.62**↑	**0.58**↑
Weighted Avg	0.68	0.51	0.52	**0.69**↑	**0.62**↑	**0.63**↑

Table 9: Sentiment analysis before vs. after dialect normalization. Bold↑ indicates improvement; ↓ indicates decline.

Analysis of Changes To further understand the impact of dialect normalization, we analyzed how the sentiment model's predictions changed for each sentence. A total of 1603 sentences were evaluated. The analysis reveals that the normalization process had a net positive effect, correcting more errors than it introduced.

As summarized in Table 10, out of the 1603 sentences, a significant number of prediction errors were corrected. Specifically:

- **Improvement (Corrected Errors)**: 265 sentences (16.53% of the dataset) saw their incorrect predictions become correct after normalization. This indicates that the model was better able to understand the intended sentiment of the dialectal text.
- **Regression (New Errors)**: Only 80 sentences (4.99% of the dataset) were misclassified after normalization, having been correctly classified before.

This results in a net improvement of 185 sentences, leading to an overall accuracy increase of 11.54% ($0.6213 - 0.5059$).

Analyzing these changes by sentiment class provides deeper insights. The normalization process was highly effective for the **NEGATIVE** class, with a net improvement of 160 sentences. The **NEUTRAL** class also saw a positive net effect of 41 sentences. However, for the **POSITIVE** class, the normalization had a slightly negative net effect, introducing 28 new errors while only correcting 12. This suggests that the dialectal forms for positive sentiment may be more complex or ambiguous, and further refinement is needed for this specific class.

The overall success rate of the normalization, calculated as the ratio of corrected errors to total changed predictions, is high at 76.81% ($\frac{265}{265+80}$). This strong performance confirms that dialect normalization is a highly effective preprocessing step for improving sentiment analysis of Vietnamese dialectal text.

6 Limitations

While the ViDia2Std corpus and our sequence-to-sequence models establish robust benchmarks for Vietnamese dialect normalization, the work also reveals several inherent challenges. A primary challenge is the tendency toward **over-normalization**, where models sometimes "over-normalize" expressions, resulting in the loss of critical pragmatic or stylistic cues embedded in the original dialect. This failure manifests as the inaccurate substitution of formal terms or specialized terminology with common standard equivalents. This issue is confirmed through our in-depth analysis of machine translation regressions (Table 8), where *Normalization Model Error* (failure to preserve nuanced semantics) was the root cause in 17.5% of the analyzed regression instances.

Type of Change	**Count**	**Percentage (%)**
Incorrect → Correct	265	16.53%
Correct → Incorrect	80	4.99%
Correct → Correct	731	45.60%
Incorrect → Incorrect	527	32.88%
Total	**1603**	**100.00%**

Table 10: Comparison: Improvement vs. Regression after Normalization

A second limitation, as noted by reviewers, is the reliance on a **single LLM-as-a-Judge** for the extrinsic evaluation. This approach, while scalable, carries an inherent risk of evaluator bias and stochasticity. This concern is quantitatively supported by our own manual analysis in Table 8. We found that *LLM Evaluator Noise* was the second-largest cause of regression, accounting for a significant 35.7% of all cases where a translation was downgraded. This finding confirms that the LLM judge's stochasticity is a valid concern. Future work should incorporate a parallel human evaluation study to establish a clearer correlation and quantify the precise agreement between the LLM judge and human assessments.

7 Conclusion and Future Work

We introduced ViDia2Std, the first large-scale, manually annotated Vietnamese dialect-to-standard corpus spanning all 63 provinces. With over 13,000 sentence pairs from authentic social media, it offers a new benchmark for dialectal diversity. Experiments show that sequence-to-sequence models—especially mBART-large-50—can effectively normalize dialectal input, leading to significant improvements in downstream tasks like machine translation and sentiment analysis. These findings highlight the value of dialect-aware preprocessing in Vietnamese NLP.

Future work includes (1) expanding the corpus to cover more sources (e.g., spoken transcripts, forums) and genres, and (2) addressing over-normalization, where pragmatic or stylistic cues are lost. We aim to develop context-sensitive models that can distinguish between dialectal terms needing normalization and colloquialisms that should be preserved. This ensures clarity without erasing linguistic nuance.

Ethics Statement

The data for the ViDia2Std corpus was collected exclusively from public Facebook news fanpages. No private user data was accessed. To protect user privacy, all collected data was anonymized during the preprocessing pipeline. All metadata, such as usernames, user IDs, and mentions, was stripped from the comments. The dataset is intended solely for non-commercial linguistic research. This work was completed before Vietnam's June 12, 2025 reorganization (Reso-

lution No. 202/2025/QH15), which reduced provincial-level units from 63 to 34 (effective July 1, 2025); all maps, indexing and analysis use the 63-province structure current at project design.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund. We would like to thank the anonymous reviewers, the Senior Program Committee (SPC), and the Area Chair (AC) of AAAI-26. Their insightful comments and constructive feedback significantly improved the quality of this paper.

References

5CD-AI. 2024. Vietnamese-Sentiment-visobert. https://huggingface.co/5CD-AI/Vietnamese-Sentiment-visobert.

Abe, K.; Matsubayashi, Y.; Okazaki, N.; and Inui, K. 2018. Multi-dialect Neural Machine Translation and Dialectometry. In Politzer-Ahles, S.; Hsu, Y.-Y.; Huang, C.-R.; and Yao, Y., eds., *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics.

Alam, M. M. I.; and Anastasopoulos, A. 2025. Large Language Models as a Normalizer for Transliteration and Dialectal Translation. In Scherrer, Y.; Jauhiainen, T.; Ljubešić, N.; Nakov, P.; Tiedemann, J.; and Zampieri, M., eds., *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, 39–67. Abu Dhabi, UAE: Association for Computational Linguistics.

Alshutayri, A.; and Atwell, E. 2019. A Social Media Corpus of Arabic Dialect Text. In Stemle, E.; and Wigham, C. R., eds., *Computer-Mediated Communication: Building Corpora for Sociolinguistic Analysis*, Cahiers du Laboratoire de Recherche sur le Langage, 1–23. Clermont-Ferrand, France: Presses universitaires Blaise Pascal.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

bmd1905. 2024. vietnamese-correction-v2. https://huggingface.co/bmd1905/vietnamese-correction-v2.

Bollmann, M. 2019. A Large-Scale Comparison of Historical Text Normalization Systems. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3885–3898. Minneapolis, Minnesota: Association for Computational Linguistics.

Burghardt, M.; Granvogl, D.; and Wolff, C. 2016. Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing. In Calzolari, N.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2029–2033. Portorož, Slovenia: European Language Resources Association (ELRA).

Fernandes, P.; Agrawal, S.; Zaranis, E.; Martins, A. F. T.; and Neubig, G. 2025. Do LLMs Understand Your Translations? Evaluating Paragraph-level MT with Question Answering. arXiv:2504.07583.

Held, W.; Ziems, C.; and Yang, D. 2023. TADA : Task Agnostic Dialect Adapters for English. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 813–824. Toronto, Canada: Association for Computational Linguistics.

Kocmi, T.; and Federmann, C. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Nurminen, M.; Brenner, J.; Koponen, M.; Latomaa, S.; Mikhailov, M.; Schierl, F.; Ranasinghe, T.; Vanmassenhove, E.; Vidal, S. A.; Aranberri, N.; Nunziatini, M.; Escartín, C. P.; Forcada, M.; Popovic, M.; Scarton, C.; and Moniz, H., eds., *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 193–203. Tampere, Finland: European Association for Machine Translation.

Kuparinen, O.; Miletić, A.; and Scherrer, Y. 2023. Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13814–13828. Singapore: Association for Computational Linguistics.

Le, T.; and Luu, A. 2023. A Parallel Corpus for Vietnamese Central-Northern Dialect Text Transfer. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13839–13855. Singapore: Association for Computational Linguistics.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2757–2791. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Nguyen, T.-N.; Le, T.-P.; and Nguyen, K. 2024. ViLexNorm: A Lexical Normalization Corpus for Vietnamese Social Media Text. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European*

Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 1421–1437. St. Julian's, Malta: Association for Computational Linguistics.

Nguyen Luong Tran and Duong Le and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Interspeech 2022*, 1751–1755.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Patil, A.; Tao, S.; and Jadon, A. 2025. English Please: Evaluating Machine Translation with Large Language Models for Multilingual Bug Reports. arXiv:2502.14338.

Phan, L.; Tran, H.; Nguyen, H.; and Trinh, T. H. 2022. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation. In Ippolito, D.; Li, L. H.; Pacheco, M. L.; Chen, D.; and Xue, N., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 136–142. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics.

Sun, Y.; Zhu, D.; Chen, Y.; Xiao, E.; Chen, X.; and Shen, X. 2025. Fine-Grained and Multi-Dimensional Metrics for Document-Level Machine Translation. In Ebrahimi, A.; Haider, S.; Liu, E.; Haider, S.; Leonor Pacheco, M.; and Wein, S., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, 1–17. Albuquerque, USA: Association for Computational Linguistics. ISBN 979-8-89176-192-6.

Tang, Y.; Tran, C.; Li, X.; Chen, P.-J.; Goyal, N.; Chaudhary, V.; Gu, J.; and Fan, A. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. arXiv:2008.00401.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.