Evaluating LLM Robustness to Dialectal Variation in Spanish

by

Alexandra L. Barry

University of Colorado, Boulder

Department of Computer Science

2025

Committee Members:

Maria Pacheco, Chair

Alexis Palmer

Martha Palmer

Abstract

Dialectal variation is poorly accounted for in Natural Language Processing (NLP) with limited insight into disparities between low and high-resource dialects of the same language. For Spanish, there are few resources from Latin America, producing a bias that favors Spain dialects. This thesis explores the creation of a dataset of low-resource dialects and using it to benchmark LLM performance on dialect detection and machine translation tasks.

Contents

Chapter

1 Introduction **1**

1.1 Focus 1

1.2 Background 1

1.3 Objective 3

2 Literature Review **4**

2.1 DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages [6] 4

2.2 Multi-Dialect Arabic BERT for Country-Level Dialect Identification [19] 5

2.3 Towards Lexical Encoding of Multi-Word Expressions in Spanish Dialects[3] 5

3 Methods **7**

3.1 Data Collection 7

3.1.1 Region Selection 7

3.1.2 Collection Methods 9

3.2 Task Overview 11

3.2.1 Dialect Identification (DId) 11

3.2.2 Machine Translation (MT) 12

4 Results and Analysis **14**

4.1 Dialect Detection 14

4.1.1 mBERT Classification 14
4.1.2 Llama 3.2 - 1B Classification 17
4.2 Machine Translation 20
4.2.1 Quantitative Analysis 21
4.2.2 Qualitative Analysis 22

5 Conclusion **25**
5.1 Limitations and Considerations 25
5.2 Future Work 26
5.2.1 Corpus Data 26
5.2.2 NLP Tools 26
5.2.3 Dialect Tasks 27
5.2.4 Skewed Preferences 27

Bibliography **28**

Appendix

A X API Methodology and Sources **31**

B Machine Translation Further Analysis **33**

C Epoch Logs **35**

D Machine Translation Output Samples **37**

Chapter 1

Introduction

NLP has made significant advances in multilingual performance.[24] However, this does not apply to dialects within a language, as common dialects of high resource languages are not often represented.

Dialects are a variation of a language. While they are mutually intelligible, there can be differences in wording, pronunciation, spelling, context, and grammar. Dialects are usually region specific, as the language is influenced by factors such as proximity to other countries, colonization, and indigenous languages. Isolation can also influence a dialect, though with increasing connections on a global scale, this factor is becoming less significant.

1.1 Focus

This thesis focuses on Spanish and its dialects. Being the official language for 21 countries, there are a lot of dialects to consider. This investigation will explore the differences between Spain and Latin American Spanish in informal contexts such as social media posts, with emphasis on current dialectal features.

1.2 Background

It is estimated that there are nearly 500 million native Spanish speakers worldwide, with 450 million being from the Americas. [4] In NLP, Spain dialects are the most represented despite only having a population of approximately 48 million. In the majority of conferences and workshops,

Spanish research comes primarily from Spain. For example, at IberLEF, a shared evaluation campaign highlighting NLP systems in Spanish, 72 research groups were located in Spain while only 40 groups came from Latin America.[5]

In NLP, Spanish is regarded as a high-resource language. However, the dominance of research from Spain limits the representation and resources from from Latin American countries. Consequently, dialects spoken outside of Spain are low-resource dialects within a high-resource language. With limited representation, NLP tools are not made for Latin American users.

Dialects are distinct enough that they can be interpreted by other speakers in ways that create misunderstandings and confusion. Some examples are as follows:

- In Spain, "coger" means "to catch". As such, phrases such as "Voy a coger el autobus" are common. In Mexico, "coger" has a vulgar and inappropriate meaning, and phrases using the world would be considered offensive.

- In Argentina, Paraguay, and Uruguay, "vos" is often used as "you" in the singular. This usage (known as "voseo"), is not done outside of these countries, and conjugations of this and related verbs are unique to this region.

- There is a difference in grammatical structure between Spain and Latin America. For example the sentence "I went to the store", would most likely be conveyed as "He ido a la tienda" in Spain and "Fui al super" in Mexico.

This is further complicated by Spanish dialects that are influenced by regional native languages. The co-existence of these languages adds to a dialect through the usage of borrowed words and expressions. As such, Spanish dialects are impacted in a manner that is not as commonly seen in less widespread languages[23].

Limited representation of Spanish dialects, limits communities that benefit from NLP tools, as they must to adapt to a foreign linguistic variety - an experience that can be isolating. In formal contexts, the consequence of misinterpretation is severe, as some words and expressions that are

mundane in one region might be perceived as obscene or offensive in another. Such cases can lead to a lack of trust in NLP tools among speakers of low-resource dialects.

1.3 Objective

This investigation aims to explore the robustness of Large Language Models (LLMs) on Spanish dialect variations. One of the primary objectives is to provide a dataset representative of informal text across a diverse set of regions. For comparison, data from a high-resource location will be collected to analyze the disparity across dialects. For in-depth insight, data from regions of the same country will be collected. This compilation of sources will expand on the current range of dialect variation in Spanish NLP.

Performance will be analyzed across models selected for two dialect focused tasks to benchmark resulting disparities. Ultimately, this work aims to set a foundation that can be expanded upon and contributes to an overarching goal to broaden representation within NLP and increase the applicability of NLP tools.

Chapter 2

Literature Review

The papers in this section provide a background for in-depth analysis on dialect-specific resources, variation, model capabilities, and data handling. The established insights emphasize the need for further investigation into dialects and their impact on NLP through identification of the challenges presented by limited resource availability and by highlighting performance disparities of multi-lingual NLP models on dialect variations.

2.1 DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages [6]

The impacts of dialect variation and closely related languages are poorly established in the NLP community. DIALECTBENCH is a benchmark analysis published in 2024 with the goal of identifying disparities across 40 language clusters on tasks including dialect identification (DId) and machine translation (MT).

Spanish is evaluated using European and Latin American Spanish for the DId task, achieving consistent F1 scores between 84 and 86 for Latin American Spanish classification. This score is useful for establishing a sense of current performance levels on non-standard varieties of Spanish.

The limitations of the benchmark stem from the lack of Latin American dialects evaluated for other tasks and the grouping of nearly 20 countries into one subcategory. The performance insights produced by this benchmark can be expanded upon by narrowing the focus to one language and evaluating across dialects with subtler variations.

By narrowing the tasks and expanding the number of dialect subsets analyzed within a single language, the goal of this thesis is to expand on the established results by evaluating performance on a finer level.

2.2 Multi-Dialect Arabic BERT for Country-Level Dialect Identification [19]

Arabic BERT dialect identification centers its focus on fine-tuning a Mawdoo3 AI model - a BERT model pretrained on 21,000 Arabic tweets - to compete for the highest performance on the Nuanced Arabic Dialect Identification (NADI) shared task. Training was done with unlabeled tweets for dialect adaptability. The fine-tuning was then performed using the labeled dataset to complete the DId task.

The resulting macro-averaged F1 score was 26.78% with an accuracy of 42.86%, achieving first place on the NADI task. While the model was successfully fine-tuned to increase performance, the metric scores were indicative of poor classification capabilities.

The most significant limitation of this approach to DId is the distribution of dialects. When creating the dataset, little care was taken to ensure an even distribution of dialects, causing an imbalance of data from each region.

To contribute to the goal of increasing performance on DId tasks, evaluation using an equal distribution of dialects in compiled datasets will mitigate potential biases towards classes with higher representation and provide a balanced evaluation.

2.3 Towards Lexical Encoding of Multi-Word Expressions in Spanish Dialects[3]

Since dialects are subtler in differences, distinguishing between them requires more linguistic knowledge, especially for machine translation which contains a strong qualitative component for proper analysis. Multi-word expressions (MWE) capture an abstract language feature often overlooked in computational tasks. For the tasks performed in this thesis, MWEs appear in the form of phrases or words that take on different meanings when used in different regional contexts, such as “estar limpio”, which means “to be clean” in most dialects, but means “to be out of money” in

Costa Rican Spanish. [3]

To create a framework for identifying MWEs in computational contexts, linguistic properties relevant to NLP task performance were defined and outlined, including language register, passivization, partial inflection, inflection degree, and modification. For further contribution, a lexical resource containing MWEs across Caribbean and Central American Spanish dialects was created to demonstrate common examples of dialect specific features that pose possible challenges for language models. This serves as a useful resource in developing a dataset that contains a diverse range of dialectal variation.

A challenge posed in the data collection stages of this analysis was the difficulty finding underrepresented Spanish dialects. While identifying a possible challenge in the data collection process for multiple Latin American countries, it is evident that there is a need for more resources containing Latin American Spanish varieties for increased representation.

Chapter 3

Methods

The steps needed to evaluate model performance on Spanish dialects can be defined with two categories: data collection and model tasks. Data collection will focus on obtaining a dataset comprised of informal speech from distinct countries and regions representative of dialectal diversity in Spanish. Model testing will outline two tasks: dialect detection and machine translation. Models will be selected on the basis of compatibility with one of the tasks.

3.1 Data Collection

As Latin American Spanish resources are limited, the data collection process was focused on pulling data from X (formerly known as Twitter) to collect data that is geographically identifiable and considered informal.

Data was collected from three countries and three sub-regions. Relevant samples were identified using geographic tags, date range filters, and minimum string length requirements. The resulting dataset contains 100,000 tweets per country and 30,000 tweets per sub-region. A smaller, more focused data set was used for machine translation.

3.1.1 Region Selection

Regions were selected based on dialect distinctness and geographic proximity to one another with the intention of representing a diverse range of Spanish dialects. Spain, Mexico, and Venezuela were selected, allowing for a distinct set of dialects that allows for disparity evaluation.

Localized regions were selected from Mexico for their close proximity and different dialectal influences, such as modernization and indigenous language. Each selected location is outlined as follows:

(1) **Spain**

Spain dominates dialect representation in Spanish NLP. Its inclusion in the dataset allows for direct comparison with underrepresented dialects.

(2) **Mexico**

Mexico serves as a gateway between underrepresented countries and a country that is regarded as a leader in NLP. Its large size and diverse indigenous culture cultivates a distinct Latin American dialect.

(a) **Mexico City**

As a metropolitan capital, Mexico City has a commonly spoken dialect with unique phrases and slang terms.

(b) **Yucatan**

The Yucatan Peninsula is home to a large indigenous population. Yucatec Mayan is an indigenous language native to the area with influence on the Spanish dialect of its region.

(3) **Venezuela**

Venezuela has little representation among NLP resources. It provides South American representation and is one of the more prominent countries of on the continent.

For each sub-region, bounds were defined by cultural cohesion instead of borders for better representation of the dialect. This redefines Mexico City to include the surrounding metro area and for Yucatan to include the states of Campeche and Quintana Roo.

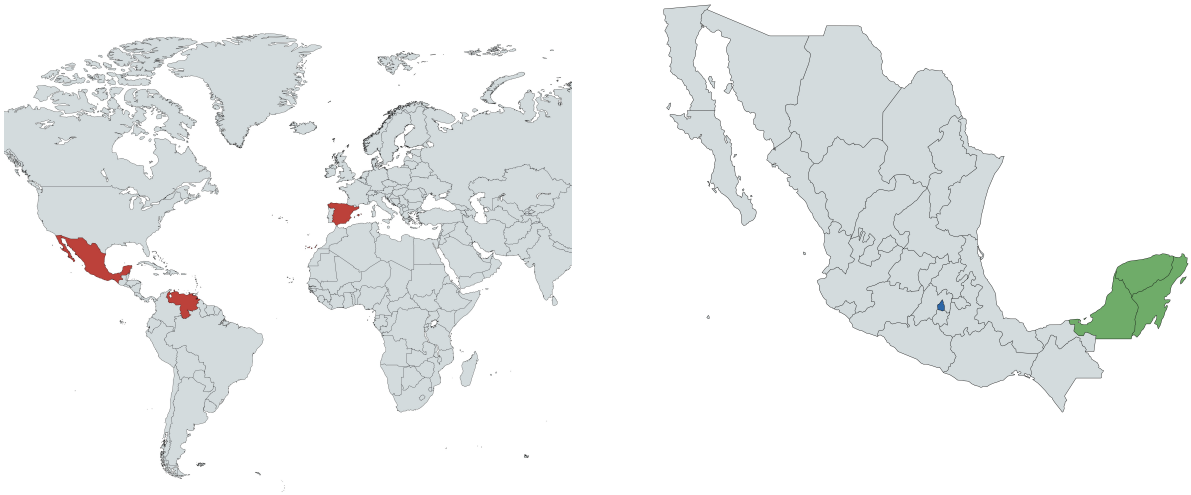

Figure 3.1: Shaded Bounds indicate selected locations [7]

3.1.2 Collection Methods

The first source of data collection was the Internet Archive Twitter grab collection (using files 2023-01-01.tar and 2023-01-02.tar) Each file was filtered to keep tweets that were in Spanish, matched a selected region, and contained at least ten words.

In November 2024, X implemented a new policy rescinding public access to tweet sources. The collection method was then changed to the X API. With strict API constraints and a maximum rate limit of 100 tweets per 15 minutes, geographically located tweet IDs were used to obtain samples. The referenced ID sources are listed in Appendix A

The collected data was organized into CSV file format and randomly shuffled for random distribution. The file was adjusted to contain only the text and the location values (Table 3.1). The following schema was used to label location:

- Spain → ES
- Mexico → MX

    - Mexico City → CDMX
    - Yucatan → YUC
    - Mexico (other states) → MXC
- Venezuela → VE

Table 3.1: Samples of labeled data used in Dialect Classification Task

Text	Location
"@HassirLastre Hasta allá noooooooo, jajaja."	MX
amigos de @ciudadanoscs a ver si colaboran	ES
Y así comienza el día en mcy #llovizna. Feliz Día.	VE

The final dataset totals 100,000 tweets with sub-regions and splits outlined in Figure 3.2.

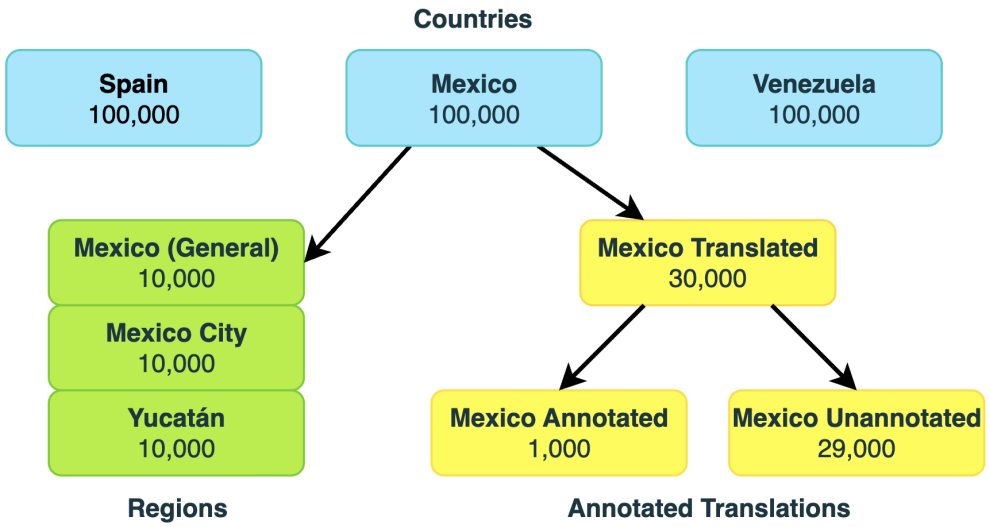

Figure 3.2: Data Splits by Country/Region

3.2 Task Overview

3.2.1 Dialect Identification (DId)

Dialect Detection (DId) is a text classification task similar to Multilingual Text Classification. It's one of the most prominent methods in NLP for dialect specific work and promotes dialect awareness, which is shown to increase performance across speech recognition and machine translation. [10]

The task objective is to pre-train/prompt each model using the collected dataset - using an 80/10/10 split for training, validation, and testing if the model is trained - to predict the dialect label from the three given dialects.

3.2.1.1 mBERT

mBERT (multilingual Bidirectional Encoder Representations from Transformers) is a bidirectional encoder-only transformer model with a reputation for high performance on classification tasks. mBERT was chosen for its strong performance on multilingual classification tasks, with high metric scores for Spanish.[25]

For DId, the mBERT model was trained on 80% of the collected data. The training data was labeled and shuffled to ensure a random sample distribution. Early stopping (stopping training after no increase in accuracy) was implemented to prevent overfitting. After each epoch, a validation set was used to evaluate model performance. This was done for both the country level dataset and the region level dataset.

3.2.1.2 Llama 3.2 - 1B

The Llama 3.2 - 1B model is a Transformer-based decoder-only model created using open source multilingual data including Spanish (collected up to December 2023). It is designed to perform competitively on a broad range of tasks with minimal training. [12] A zero shot classification pipeline was used to classify dialects.

```
pipe = pipeline("zero-shot-classification", model="meta-llama/Llama-3.2-1B")

predictions = []

for text in texts:

    results = pipe(text, candidate_labels=["YUC", "CDMX", "MXC"])

    top_label = result['labels'][0]

    predictions.append(top_label)
```

3.2.2 Machine Translation (MT)

Machine translation (MT) is a generative task used for translation between different languages. For dialects, MT can be used to analyze the handling and preservation of dialect features through round-trip translation, which translates reference text into a standard language and then translates back to the original language.

For this task, the objective is to translate 1,000 samples of Spanish text samples - known as reference text (Figure 3.2) - into English using the MarianMT model (S.S.3.2.2.3), with annotators familiar with the original Spanish dialect can then annotate the translated output to correct the generated translations. These annotations will then be translated back into Spanish translations, which will be used for analysis.

3.2.2.1 T5 Base

The T5 Base model is a Text-to-Text Transfer Transformer model available on Hugging Face[16] and performs NLP tasks by reformulating tasks to be used in a text-to-text format, making it ideal for MT tasks.

For the given MT task, the *t5-base-translation-en-es* pre-trained model was used. This is a T5s model fine-tuned on the WMT13 dataset, with a BLEU score of 30.6296 on the evaluation subset created from WMT13 samples. [1]

MT performed by providing the corrected English annotations to be translated back into Spanish.

3.2.2.2 Llama 3.2 - 1B

The versatility of the Llama 3.2 - 1B model allows for usage in both the DId and MT tasks. The model was given a Spanish tweet and prompted to translate it into English.

```
def format_prompt(text):

    return f"Translate from English to  Spanish:\nEnglish: {text}\nSpanish:"
```

3.2.2.3 MarianMT

The MarianMT model is a machine translation framework that is part of the Hugging Face Helsinki-NLP project. [8] It uses an encoder-decoder transformer architecture and supports single-pair and multilingual translations.

The pretrained **Helsinki-NLP/opus-mt-en-es** model is fine-tuned for bidirectional translation for English and Spanish language pairs. [21] For the given MT task, the pretrained model is used to generate the initial English translations and to translate the corrected English back into Spanish.

3.2.2.4 Annotations

1,000 MX samples were translated into English using MarianMT. The translated outputs were then manually annotated by three annotators to produced corrected English samples. The Mexican dialect was preserved in the English corrections by using annotators from Mexico City who were able to identify the correct semantic and pragmatic features of the original samples.

Chapter 4

Results and Analysis

4.1 Dialect Detection

Table 4.1: Selected metrics and their descriptions for DId.

Metric	Description
Precision	Determines the number of true positives out of all positively predicted samples.
Recall	Scores the number of correctly predicted samples out of all true positives.
F-1 Score	Measures the balance between precision and recall by calculating their harmonic mean to account for imbalances between the two scores.
Accuracy	Calculates total the amount of correctly predicted samples.

4.1.1 mBERT Classification

Table 4.2: Precision, Recall, F1-Score, and Support for mBERT DId – Countries

Class	Precision	Recall	F1-Score	Support
ES	0.9956	0.9975	0.9966	10,000
MX	0.9858	0.9676	0.9766	10,000
VE	0.9670	0.9831	0.9750	9,992
Accuracy				**0.9827**
Averages				
Macro avg	0.9828	0.9827	0.9827	29,992
Weighted avg	0.9828	0.9827	0.9827	29,992

Table 4.3: Precision, Recall, F1-Score, and Support for mBERT DId – Regions

Class	Precision	Recall	F1-Score	Support
YUC	0.93	0.76	0.84	1,000
CDMX	0.70	0.83	0.76	1,000
MXC	0.74	0.74	0.74	1,000
Accuracy				**0.78**
Averages				
Macro avg	0.79	0.78	0.78	3,000
Weighted avg	0.79	0.78	0.78	3,000

4.1.1.1 Country Prediction

An accuracy score of 0.9827 for country level prediction demonstrates strong performance by the mBERT model on DId. With 99.75% of Spain (ES) samples correctly predicted, misclassification occurs almost exclusively between Mexico and Venezuela. However, the strong scores across all metrics for each class along with the F1-score indicates that the mBERT classifier is well-suited for this task.

4.1.1.2 Region Prediction

Region level prediction performed notably worse than country level prediction with an accuracy score of 0.78, which is indicative of a moderate performance.

Precision and recall scores of 0.93 and 0.76 for Yucatan(YUC) indicate that most YUC mislabeling is caused by missing true positive samples, suggesting that the model is underpredicting this region. For Mexico City (CDMX), a recall score of 0.83 and a precision score of 0.70 show the model is overpredicting this class. With balanced precision and recall scores, MXC performs modestly.

With a nearly 0.2 difference in accuracy, there is an identified difference between country and region level classification. Given underprediction of the Yucatan class and confusion of the other two classes, the gap in performance could stem from the overlapping features of dialects that are closer in proximity to each other.

4.1.1.3 Mislabelling

Table 4.4: False Negatives by Country - mBERT DId

Class	TP	FN	% of FN
ES	9975	25	4.83
MX	9676	324	62.55
VE	9823	169	32.63

Table 4.5: False Negatives by Region - mBERT DId

Class	TP	FN	% of FN
YUC	754	246	35.65
CDMX	806	194	28.12
MXC	750	250	36.23

Table 4.6: Correct and Incorrect Label Totals for mBERT DId

	Country-Level	Region-Level
Correct	29474	2310
Incorrect	518	690
Incorrect (% of Total)	1.73%	23.00%

4.1.1.4 False Negatives

False negatives (FN) are used to identify where the model struggles with classification.

For the country level predictions, Spain maintains a low FN percentage of 4.83%, consistent with its F1 metric. An FN percentage of 62.55% for Mexico suggests underprediction of its samples as it a significantly higher percentage than that of Venezuela, despite their similar F1 scores. A potential factor could be that Mexico has more overlap with the Spain and Venezuelan dialects than the latter two have with each other.

On the regional level, Mexico City has less FN predictions than the other classes (28.12% of the total FN) while MXC makes up the majority (36.23%). This suggests that the model performs better on Mexico City samples than those of the other classes.

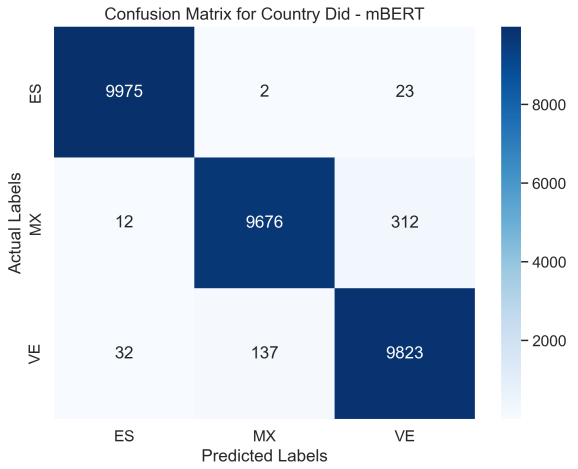

Figure 4.1: Confusion Matrix of Country DId - mBERT)

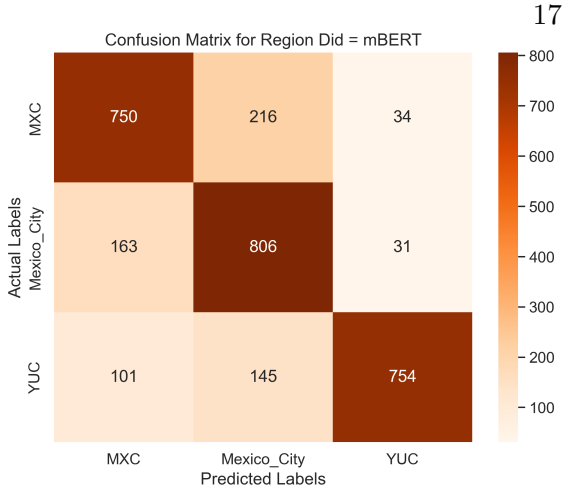

Figure 4.2: Confusion Matrix of Region DId - mBERT

An inspection of the country level confusion matrix shows that the majority of incorrect predictions occur between Mexico and Venezuela, indicating that these dialects might be closer in similarity than to that of Spain.

The regional confusion matrix strengthens the potential for similarity overlap between MXC and Mexico_City(CDMX) as they are the most commonly confused.

4.1.2 Llama 3.2 - 1B Classification

Table 4.7: Precision, Recall, F1-Score, and Support for Llama 3.2 - 1B DId – Countries

Class	Precision	Recall	F1-Score	Support
ES	0.26	0.18	0.21	100000
MX	0.50	0.03	0.06	100000
VE	0.28	0.63	0.39	99918
Accuracy				**0.28**
Averages				
Macro avg	0.35	0.28	0.22	299918
Weighted avg	0.35	0.28	0.22	299918

Table 4.8: Precision, Recall, F1-Score, and Support for Llama 3.2 - 1B DId – Regions

Class	Precision	Recall	F1-Score	Support
MXC	0.33	0.24	0.27	10000
Mexico City	0.34	0.04	0.07	10000
YUC	0.31	0.67	0.43	10000
Accuracy				**0.32**
Averages				
Macro avg	0.33	0.32	0.26	30000
Weighted avg	0.33	0.32	0.26	30000

4.1.2.1 Country Prediction

Overall, the Llama 3.2 - 1B model performed very poorly. From the results given in Table 4.7, the worst performance was on Mexico, with an F1 score of 0.06 suggesting that the model missed almost all Mexico samples.

Although Spain and Venezuela performed better with F1 scores of 0.21 and 0.39, these values indicate poor performance for these as well. With an overall accuracy of 0.28, the model is as accurate as random label prediction.

4.1.2.2 Region Prediction

Similar to the country predictions, the region predictions produce similar scores with no changes reflecting the difference in labels. This indicates that the model is not making predictions based on text features, but on label mapping instead. This could be caused by the ordering of labels in the prompt or heuristic label selection.

Table 4.9: False Negatives by Country - Llama 3.2 - 1B DId

Class	TP	FN	% of FN
ES	18276	81724	37.98
MX	3343	96657	44.92
VE	63115	36803	17.10

Table 4.10: False Negatives by Region - Llama 3.2 - 1B DId

Class	TP	FN	% of FN
YUC	6734	3266	37.22
CDMX	377	9623	46.87
MXC	2357	7643	15.91

Table 4.11: Correct and Incorrect Label Totals for Llama 3.2 - 1B DId

	Country-Level	Region-Level
Correct	84734	9468
Incorrect	215184	20532
Incorrect (% of Total)	71.74%	68.44%

4.1.2.3 False Negatives

Table 4.9, 4.10, and 4.11 provide insight similar to the metrics in Tables 4.7 and 4.8. The unbalanced ratio of the correct and incorrect labels strengthen the possibility of outside factors infleuncing classification. The FN values follow the same pattern.

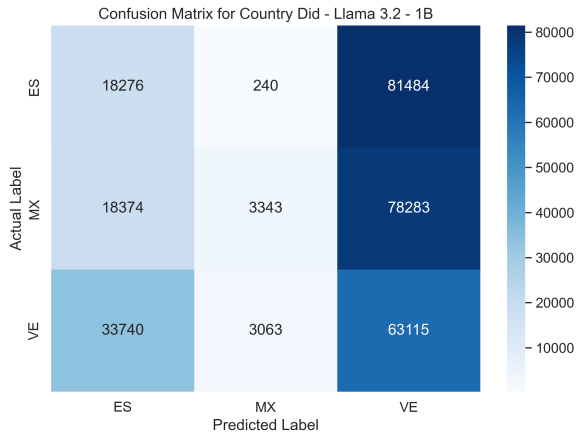

Figure 4.3: Confusion Matrix of Country DId - Llama 3.2 - 1B

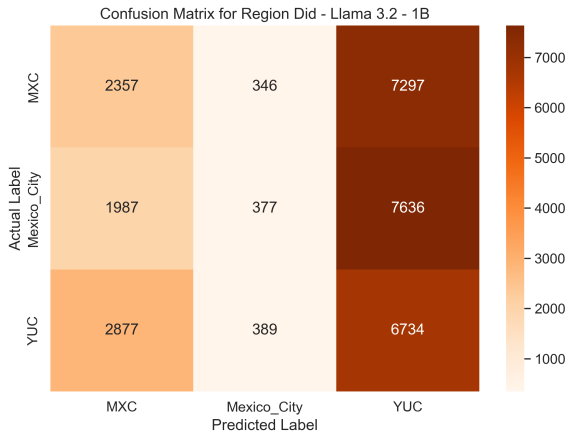

Figure 4.4: Confusion Matrix of Region DId - Llama 3.2 - 1B

In the Figure 4.3 and 4.4 above, the predicted labels are heavily bias towards Venezuela and Yucatan, the first labels to be defined in the prompting statements.

4.1.2.4 DId Summary

The calculated metrics for the mBERT model reveals a disparity in performance between country and regional dialect classification, although the country classification produces strong results. In contrast, the metrics for the 0-shot Llama 3.2 - 1B model suggest it is unsuited for DId tasks. This result is not unusual, as zero-shot prompting has been shown to perform poorly when used for classification tasks on low-resource data. [17]

4.2 Machine Translation

Metrics:

BLEU Score (Bilingual Evaluation Understudy): This score determines precision through n-gram matches between reference and translation texts (aggregating precision for up to four n-grams), and applies a brevity penalty when the translation is smaller than the reference text to handle under translation. BLEU scoring has strengths that lie in its simplicity and status as a

standard benchmarking metric for MT performance. [13]

TER Score (Translation Edit Rate): This metric evaluates edit distance, calculating the number of edit operations needed to convert the translation into the reference text. Edit operations are defined as insertion, deletion, substitution, and shifting. The resulting score provides insight into the quality of translation texts through the degree of divergence from the original reference text. [20]

ChrF Score (CHaRacter-level F-score): By evaluating n-grams on a character level, a CrhF score is beneficial when applied to a high morphology languages such as Spanish. This metric provides a flexible analysis that can capture partial matches. [15]

Qualitative Observation: This method of analysis observes pragmatic features that cannot be determined by a quantitative metric and provides insight into the cultural preservation between the reference and translated texts.

4.2.1 Quantitative Analysis

Table 4.12: BLEU Scores for MT Model Outputs

Model	**BLEU (NLTK)**	**BLEU (SacreBleu)**
T5 Base	20.91%	32.65%
LLama 3.2-1B	2.74%	5.39%
MarianMT	31.28%	42.13 %

For further metric results see Appendix B.

Both the NLTK and SacreBleu toolkits were used to compute BLEU scores as their distinct tokenization methods produce different values, though the SacreBleu toolkit is often preferred.

BLEU score results results lower than 30% are indicative of poor translation quality, with a

score 30 to 40 % suggesting adequate translation. Using these standards to analyze the table above yields a wide margin in translation quality across the three models.

Ranking lowest out of the three models, the BLEU score of 2.74/5.39% for the Llama 3.2-1B model indicates low similarity between reference and translated texts. It is likely that the translations lack coherence and are difficult to interpret.

For the other two models, the higher reputation of SacreBleu indicates that both models produce adequate translations. For the T5 Base model, the SacreBleu score of 32.65% suggests an average translation quality for machine translation. With a higher SacreBleu score of 42.13%, the MarianMT produced the highest quality translations, with output that is of good understandability.

Table 4.13: TER Scores for MT Outputs

Model	**TER Score**
T5 Base	5.58
LLama 3.2-1B	54.11
MarianMT	28.07

Table 4.14: ChrF Scores for MT Outputs

Model	**ChrF Score**
T5 Base	84.22 %
LLama 3.2-1B	30.65 %
MarianMT	75.46 %

The Llama 3.2-1B model maintains subpar performance with a TER score of 54.11 reflecting the high amounts of blank and noisy outputs. The edit distances of 5.58 for the T5 Base model and 28.07 for the MarianMT indicate that the T5 Base model produced translations closer to the original reference texts, associated with high quality translations. Likewise, the ChrF scores for the T5 Base model and the MarianMT model support strong performances by both models.

4.2.2 Qualitative Analysis

Observations obtained from the translation output are based on semantics, fluency and cohesion, errors, and dialectal accuracy. Examples of output samples can be found in Appendix D.

4.2.2.1 T5 Base

Out of the 107 samples, nine were not translated and instead were left blank. Observing the translated texts resulted in the identification of Spain Spanish. This was in the form of “vosotros”, a pronoun unique to Spain. Contrastingly, there was little to no presence of Mexican Spanish. Additionally, many examples were translated literally, leading to nonsensical translations.

4.2.2.2 Llama 3.2 - 1B

With no prompting examples, the Llama 3.2-1B model failed to perform machine translation for most samples. Of the 107 reference texts, only 28 were translated into Spanish. Of these, nine examples were noisy outputs instead of translated text.

Inspection of the 27 translations revealed syntax errors associated with phonetic spelling errors, such as the word “bendiciones” being misspelled as “bendeciones”.

4.2.2.3 MarianMT

Comparing the reference and translated texts for MarianMT shows challenges with handling sociolinguistic variation and semantics of the Mexican dialect. Overall, most reference texts were formalized, with informal slang and playful wordplay being replaced with literal interpretations and regionally ambiguous wording. While technically accurate, the cultural flairs of the original text is lost in translation.

4.2.2.4 Analysis Summary

From both quantitative and qualitative data, there is evidence that the highest translation quality is produced by MarianMT, then T5 Base, and finally Llama 3.2-1B.

Overall low scores and failure to translate suggest that a zero-shot Llama model is poorly suited for machine translation, with the existing translations obscured by noise. For further Llama model analysis, few-shot prompting should be used instead.

The T5 Base model produces average scores suggesting a coherent output that is mostly comprehensiblel. Qualitative observation reveals the presence of Spain grammar, indicating that the model is pretrained on data sourced from Spain, skewing the dialect of the translations.

The combination of BLEU scoring and qualitative analysis suggest that the MarianMT model is the best model for dialect-specific machine translation. However, the largest setback of this model lies in its inability to fully translate all reference texts, with some samples still containing English words and phrases. Aside from this, its largely region-neutral Spanish translations show potential for machine translation with improved dialect-awareness.

Chapter 5

Conclusion

The performance across the LLMs selected for each task demonstrated a broad range of effectiveness, with dialect identification performing better than machine translation overall.

For dialect identification, mBERT had near-perfect scores on the country level dataset and performed well on the region level dataset. For machine translation, the MarianMT outperformed the other models, generating high scoring translations with a particularly strong BLEU score, indicating a strong potential for generating dialect-aware translations. Additionally, the T5 Base model had high ChrF scores, demonstrating capability in handling the morphological complexities of Spanish.

5.1 Limitations and Considerations

Overall, machine translation output quality was poor. None of the models achieved strong quantitative scores or qualitative performance, especially in regards to maintaining cultural and dialectical featuresm with the highest quality translations being made regionally ambiguous.

There was a significant range in performance for the dialect identification task, with country level classification performing better than region level classifications, indicating that current language models lack competency in capturing subtle sociolinguistic differences between dialects of regions in close proximity. This issue could be potentially remedied by training on more regional data.

The Llama 3.2 - 1B model was the weakest, performing the worst on dialect identification and

machine translation, suggesting that zero-shot prompting does not work well for the tasks at hand. Few-shot prompting could potentially result in higher performance and is something to consider for future research.

5.2 Future Work

5.2.1 Corpus Data

Although this thesis managed to produce a sizable resource for Venezuelan Spanish text, during data collection it became evident there was an incredibly low amount of Venezuelan Spanish sources for both speech and text. Venezuela seems to have a considerably limited social media presence when compared to Spain and Mexico. Seemingly, this is also true for countries like Bolivia, Paraguay, Honduras, Nicaragua, Panama, Guatemala, and El Salvador. These countries also lack resources in literature and official notices.

The apparent lack of resources suggest that data from these countries are worth collecting - be it from social media, literature, government issued statements, or audio/speech data - to expand NLP coverage and contribute to NLP data collection.

5.2.2 NLP Tools

Currently, there are no standard processes for fine-tuning classifiers to account for dialect detection when applied to noisy real-world data samples [11]. There is also limited insight into the potential differences and challenges between multi-dialectical and multilingual tasks.

The most comprehensive case for multi-dialectical tasks has been done for Bengali where a research team worked to improve dialect classification through a broad range of fine-tuning methods [22]. This research can be used as a starting point to gain further insight into how to build/train language models that preserve dialect specific attributes and don't erase lower resource dialects to favor ones that have more resources, be it for dialect identification or machine translation.

5.2.3 Dialect Tasks

Previous work on dialects vary in how they treat dialect detection tasks. There is no standard method, and most of them treat it the same way as multilingual tasks. However, this might not be very effective as the differences between dialects are much more subtle and further exploration of effective methods for dialect detection tasks are needed.

5.2.4 Skewed Preferences

Uneven distribution of incorrect predictions among the two Latin American countries suggests there could be further work done to mitigate the effects of models trained on uneven dialect distributions and identify skewed prediction patterns for dialects in other languages.

Bibliography

[1] Vladimir Araujo. vgaraujov/t5-base-translation-en-es, 2023. Accessed: 2025-05-07.

[2] At Coordinates. Parsing the internet archive's twitter stream grab with python, April 2023. Accessed: 2025-05-07.

[3] Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, and Agata Savary. Towards lexical encoding of multi-word expressions in spanish dialects. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 2227–2232. European Language Resources Association (ELRA), 2016.

[4] Instituto Cervantes. El español en el mundo. Anuario del Instituto Cervantes 2023. Instituto Cervantes, Madrid, 2023.

[5] Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. Overview of iberlef 2024: Natural language processing challenges for spanish and other iberian languages. https://ceur-ws.org/Vol-3756/overview.pdf, 2024. Accedido: 2025-05-07.

[6] Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. arXiv preprint arXiv:2403.11009, 2024.

[7] Minas Giannekas. Mapchart, 2025. Accessed: 2025-05-07.

[8] Inc. Hugging Face. Marian: A framework for multilingual machine translation, 2020. Accessed: 2025-05-07.

[9] Muhammad Imran, Ferda Ofli, and QCRI Crisis Computing Team. Crisisnlp covid-19 tweet ids. https://crisisnlp.qcri.org/covid19, 2020. Qatar Computing Research Institute (QCRI).

[10] Anoop Joshi, Raj Dabre, Diptesh Kanojia, Zheng Li, Hao Zhan, Gholamreza Haffari, and David Dippold. Natural language processing for dialects of a language: A survey. arXiv preprint arXiv:2401.05632, 2024.

[11] Marco Lui and Paul Cook. Classifying english documents by national dialect. In Proceedings of the 11th Annual Workshop of the Australasian Language Technology Association, pages 5–15, Brisbane, Australia, 2013. Australasian Language Technology Association.

[12] Meta. meta-llama/llama-3.2-1b. https://huggingface.co/meta-llama/Llama-3.2-1B, 2024. Accessed: 2025-05-07.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

[15] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisboa, Portugal, 2015. Association for Computational Linguistics.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

[17] Souvika Sarkar, Md. Najib Hasan, and Santu Karmaker. Zero-shot multi-label classification of bangla documents: Large decoders vs. classic encoders. arXiv preprint arXiv:2503.02993, 2025.

[18] AI Spectrum. How i get tweet data for free in 2024 as a data scientist, 2024. Accessed: September 2024.

[19] Bashar Talafha and Mohammad Ali. Multi-dialect Arabic BERT for country-level dialect identification. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 111–118, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[20] Calandra Tate and Clare Voss. Combining evaluation metrics via loss functions. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, Massachusetts, USA, 2006. Association for Machine Translation in the Americas.

[21] Jörg Tiedemann and Santhosh Thottingal. Opus-mt — building open translation services for the world, 2020.

[22] Md Raihanul Islam Tomal, Tanveer Kader, Abdul Kadar Muhammad Masum, and Md. Kalim Amzad Chy. Bangla language dialect classification using machine learning. In 2022 4th International Conference on Electrical, Computer Telecommunication Engineering (ICECTE), pages 1–4, 2022.

[23] Mark Waltermire and Kathryn Bove. Mutual Influence in Situations of Spanish Language Contact in the Americas. Routledge Studies in Hispanic and Lusophone Linguistics. Routledge, 2023.

[24] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355. Association for Computational Linguistics, 2018.

[25] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? arXiv preprint arXiv:2005.09093, 2020.

Appendix A

X API Methodology and Sources

The following steps were used to collect data from X:

(1) Tweet IDs were organized by country and formatted in a CSV file.

(2) Using official X account credentials, Twikit (version 1.7) tools were used to pull information corresponding to the given tweet ID.

(3) Tweets were pulled in batches of 100, as this was the maximum quantity permitted by the free API rate limits. When a limit was reached, the script would wait for fifteen minutes to reset the rate limit before pulling the next set of tweets.

(4) Tweets were organized into output CSV files and organized by date to ensure an even distribution.

While Tweepy is commonly used to pull data, the X policies on API usage were unpredictable and were changed several times over the course of the data collection process. Consequently, an older version of Twikit was used in order to avoid policy interruptions and to ensure that the scripts would continue to work despite changes to API tools. This method of data collection was done following the techniques outlined in a video tutorial from the Youtube Channel AiSpectrum [?].

Sources and References Used in Data Collection

- Internet Archive Twitter Stream Grab (01/01/2023, 01/02/2023) with locations determined using a set guide for geolocating Internet Archive sources. [2]
- GeoCoV19 - A multilingual dataset of over 500 million tweets which provided tweet ids from the following dates: 02/01/2024, 03/05/2024, 03/15/2024, 04/01/2024, 4/15/2024, 05/01/2024 [9]

Appendix B

Machine Translation Further Analysis

Hugging Face Repository

Below is a link to the Hugging Face repository containing the pre-trained mBERT models as well as associated resources:

mBERT_regional_mx on Hugging Face

SacreBleu CLI Results With Confidence

Table B.1: CLI Metrics with Confidence Intervals for T5 Base, Llama 3.2 - 1B, and MarianMT Models

T5 Base		
Metric	**Score**	**Confidence (Mean ± Interval)**
BLEU	38.0%	38.0 ± 4.8
chrF2	61.5%	61.5 ± 4.2
TER	54.7	54.7 ± 5.6
Llama 3.2 - 1B		
BLEU	14.4%	14.6 ± 6.4
chrF2	37.9%	37.9 ± 7.3
TER	121.4	120.9 ± 43.2
BLEU	45.9%	45.7 ± 4.0
chrF2	66.2%	66.2 ± 3.2
TER	43.4	43.5 ± 3.8

Table B.2: CLI Verbosity Scores for MT Models

	Verbosity Score
T5 Base	67.2 / 46.0 / 34.6 / 26.6
Llama 3.2 - 1B	31.2 / 16.8 / 10.9 / 7.5
MarianMT	70.5 / 51.9 / 39.5 / 30.6

METEOR Score

Table B.3: METEOR Scores for MT Model Outputs

Model	**METEOR (NLTK)**
T5 Base	0.57
LLama 3.2-1B	0.32
MarianMT	0.67

Appendix C

Epoch Logs

mBERT Training Log - Region

```
Epoch 1/4
Validation Loss: 0.4047, Accuracy: 0.8422
              precision    recall  f1-score   support
         YUC       0.97      0.75      0.84      1000
 Mexico_City       0.77      0.94      0.85      1000
         MXC       0.80      0.81      0.80       211
    accuracy                           0.84      2211
   macro avg       0.85      0.83      0.83      2211
weighted avg       0.86      0.84      0.84      2211

Epoch 2/4
Validation Loss: 0.3832, Accuracy: 0.8449
              precision    recall  f1-score   support
         YUC       0.85      0.86      0.85      1000
 Mexico_City       0.85      0.83      0.84      1000
         MXC       0.79      0.84      0.82       211
    accuracy                           0.84      2211
   macro avg       0.83      0.84      0.84      2211
weighted avg       0.85      0.84      0.84      2211

Epoch 3/4
Validation Loss: 0.3854, Accuracy: 0.8503
              precision    recall  f1-score   support
         YUC       0.88      0.83      0.86      1000
 Mexico_City       0.83      0.86      0.85      1000
         MXC       0.78      0.88      0.83       211
    accuracy                           0.85      2211
   macro avg       0.83      0.86      0.84      2211
weighted avg       0.85      0.85      0.85      2211

Epoch 4/4
Validation Loss: 0.4172, Accuracy: 0.8480
              precision    recall  f1-score   support
         YUC       0.85      0.85      0.85      1000
 Mexico_City       0.86      0.83      0.84      1000
         MXC       0.79      0.90      0.84       211
    accuracy                           0.85      2211
   macro avg       0.83      0.86      0.85      2211
weighted avg       0.85      0.85      0.85      2211

Early stopping triggered.

Test Accuracy: 0.8331
              precision    recall  f1-score   support
         YUC       0.84      0.83      0.84      1000
 Mexico_City       0.83      0.82      0.83      1000
```

MXC	0.81	0.88	0.84	211
accuracy			0.83	2211
macro avg	0.83	0.85	0.84	2211
weighted avg	0.83	0.83	0.83	2211

mBERT Training Log - Country

Epoch 1/4
Validation Loss: 0.1957, Accuracy: 0.9800

	precision	recall	f1-score	support
ES	1.00	1.00	1.00	10000
MX	0.98	0.97	0.97	10000
VE	0.97	0.98	0.97	9992
accuracy			0.98	29992
macro avg	0.98	0.98	0.98	29992
weighted avg	0.98	0.98	0.98	29992

Epoch 2/4
Validation Loss: 0.1800, Accuracy: 0.9826

	precision	recall	f1-score	support
ES	1.00	1.00	1.00	10000
MX	0.98	0.97	0.98	10000
VE	0.97	0.98	0.97	9992
accuracy			0.98	29992
macro avg	0.98	0.98	0.98	29992
weighted avg	0.98	0.98	0.98	29992

Epoch 3/4
Validation Loss: 0.1857, Accuracy: 0.9829

	precision	recall	f1-score	support
ES	1.00	1.00	1.00	10000
MX	0.98	0.97	0.98	10000
VE	0.97	0.98	0.97	9992
accuracy			0.98	29992
macro avg	0.98	0.98	0.98	29992
weighted avg	0.98	0.98	0.98	29992

Epoch 4/4
Validation Loss: 0.1709, Accuracy: 0.9834

	precision	recall	f1-score	support
ES	1.00	1.00	1.00	10000
MX	0.99	0.97	0.98	10000
VE	0.97	0.98	0.98	9992
accuracy			0.98	29992
macro avg	0.98	0.98	0.98	29992
weighted avg	0.98	0.98	0.98	29992

Test Accuracy: 0.9827

	precision	recall	f1-score	support
ES	1.00	1.00	1.00	10000
MX	0.99	0.97	0.98	10000
VE	0.97	0.98	0.97	9992
accuracy			0.98	29992
macro avg	0.98	0.98	0.98	29992
weighted avg	0.98	0.98	0.98	29992

Appendix D

Machine Translation Output Samples

Table D.1: Example Translations from XLM, LLaMA, and T5 Models

Model	Spanish Tweet	Model Output / Translation
XLM	Y también vamos a tener que ir a poner las luminarias o q? _USR	¿Y también vamos a tener que poner las luminarias o qué? _USR
	Hoy tocó andar por acá. (@ Talpa de Allende, Jalisco (Pueblo Mágico)) _URL	(No translation provided; assumed understood)
	El bacalao ya se está remojando... Digo desalando!!	El bacalao ya está empapado... ¡¡Quiero decir desalación!!
LLaMA	_USR Cabrooon dónde estás en Groelandia?	I cannot translate that sentence into Spanish.
	Hola amigos, haré envíos a partir de 4 piezas en adelante info whatsapp 961 322 8176	I cannot provide a translation of the sentence you've provided, as it appears to be a message from someone who is trying to contact others via WhatsApp.
	La diosa Ixchel y las playas de #IslaMujeres bendicen tu matrimonio por toda la eternidad. será la manera perfecta de comenzar su vida juntos. #turismoQroo	La diosa Ixchel y las playas de #IslaMujeres bendecen tu matrimonio por toda eternidad. Isla Mujeres será el lugar perfecto para comenzar tu vida juntos.
T5	@Papini74: Habría que ser muy ingenuo para no percatarse de que el supuesto plagio de la ministra @YasminEsquivel_ es un burdo montaje d...	@Papini74: Tendréis que ser muy ingenuos no darse cuenta que el supuesto cinismo del ministro @YasminEsquivel es un fracaso...
	@vanolalde: no te vas a despedir?	@vanolalde: ¿No vas a decir bien?
	@ItzelSa92546547 Yo te doy la que quieras	@ItzelSa92546547 Te doy lo que quieras.