Common Ground, Diverse Roots: The Difficulty of Classifying Common Examples in Spanish Varieties

Javier Alejandro Lopetegui Gonzalez, Arij Riabi, Djamé Seddah, Djamé Seddah

Common Ground, Diverse Roots: The Difficulty of Classifying Common Examples in Spanish Varieties

Javier A. Lopetegui* **Arij Riabi*** **Djamé Seddah**
INRIA Paris, France
firstname.lastname@inria.fr

Abstract

Variations in languages across geographic regions or cultures are crucial to address to avoid biases in NLP systems designed for culturally sensitive tasks, such as hate speech detection or dialog with conversational agents. In languages such as Spanish, where varieties can significantly overlap, many examples can be valid across them, which we refer to as common examples. Ignoring these examples may cause misclassifications, reducing model accuracy and fairness. Therefore, accounting for these common examples is essential to improve the robustness and representativeness of NLP systems trained on such data. In this work, we address this problem in the context of Spanish varieties. We use training dynamics to automatically detect common examples or errors in existing Spanish datasets. We demonstrate the efficacy of using predicted label confidence for our Datamaps (Swayamdipta et al., 2020) implementation for the identification of hard-to-classify examples, especially common examples, enhancing model performance in variety identification tasks. Additionally, we introduce a Cuban Spanish Variety Identification dataset with common examples annotations developed to facilitate more accurate detection of Cuban and Caribbean Spanish varieties. To our knowledge, this is the first dataset focused on identifying the Cuban, or any other Caribbean, Spanish variety.

1 Introduction

Language reflects culture and identity, while also capturing subtle variations that shape communication. In Natural Language Processing (NLP), it is crucial to account for these nuances, especially in language variety identification, where small shifts in meaning, often tied to cultural interpretations, can impact sensitive tasks like hate speech detection. Expressions that may be benign in one dialect can be offensive in another, making accurate variety identification essential to prevent misclassifications and ensure culturally appropriate responses (Nozza, 2021; Hershcovich et al., 2022). In such tasks,

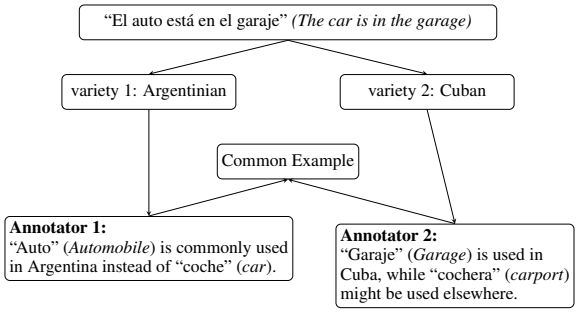


Figure 1: Common Example Identification in Language Variety Classification

cross-lingual models often struggle with these subtle cultural and linguistic distinctions, as the same formulation may carry vastly different meanings across varieties. Language-specific models tend to perform better in such cases, as they are more sensitive to regional variations (Nozza, 2021; Vaidya et al., 2024; Arango et al., 2021; Montariol et al., 2022; Castillo-lópez et al., 2023). However, distinguishing between closely related languages, dialects, and regional varieties of the same language is a key and difficult task in language identification (Tiedemann and Ljubešić, 2012; Lui and Cook, 2013; Zampieri and Nakov, 2021; España-Bonet and Barrón-Cedeño, 2024). Adding to this complexity is the issue of common examples —valid phrases across multiple dialects or varieties. Overlooking these examples can result in biased classifications, especially in languages like Spanish, where variety overlap is frequent. [1] Despite this, many current datasets treat the identification of the language variety as a single label classification

*These authors contributed equally.

[1]Following Hudson (1996), we use the terms varieties of Spanish: "a variety is a set of linguistic items with similar social (including geographical and cultural) distribution."

task, which overlooks this crucial aspect (Zampieri et al., 2024). Current datasets for language variety identification often rely on manual annotations or automated methods such as geographic information (Zampieri et al., 2019; Abdul-Mageed et al., 2020, 2022; Aepli et al., 2022) or keyword-based classification (Althobaiti, 2022). However, both approaches have limitations, and manually checking large datasets for common examples is challenging and costly (Keleg and Magdy, 2023; Bernier-colborne et al., 2023). Datamaps based on training dynamics (Swayamdipta et al., 2020; Weber-Genzel et al., 2024), which track how the confidence of the model changes over epochs, have been used successfully to detect annotation errors and human label variation. These methods highlight which examples are consistently easy or difficult for the model, with hard examples often pointing to ambiguity or errors. We propose using training dynamics to detect common examples in language variety identification tasks. In 1 we show an example of these common examples. These are expected to be among the hard examples the model struggles with during training. By tracking the model's confidence in its predicted labels over multiple training epochs, rather than using gold labels, we aim to detect ambiguous instances that are hard for the model to classify consistently. Our research addresses the following questions:

- **RQ1:** Can training dynamics help detect common examples that are hard for the model to classify during the training?
- **RQ2:** Can we use the model's confidence over predicted labels to detect common examples?
- **RQ3:** Can this approach work effectively across different domains, such as news articles and user-generated content?

To investigate these questions, we use two datasets: the Spanish subset of DSL-TL dataset (Zampieri et al., 2024), which contains texts extracted from news articles, and a new dataset of Cuban Spanish varieties we collected from Twitter. We adapt the Datamaps technique by changing the way confidence and variability are calculated, allowing us to identify common examples. Our results demonstrate the efficiency of this approach in detecting common examples in both datasets.

Our main contributions are as follows:

1. We propose a modified Datamaps model that calculates confidence and variability based on the predicted label's probability, providing a more accurate reflection of model uncertainty. Our model can help accelerate the re-annotation of existing datasets.
2. Using both frequency-based methods and SHAP analysis (Lundberg and Lee, 2017), we provide a thorough error analysis that demonstrates the usefulness of our approach to capture annotation errors and shows how the model predictions are topic-dependent.
3. We present and publicly share a novel Cuban Spanish variety identification dataset, consisting of 1,762 manually annotated tweets by three native speakers, with labels assigned based on agreement and covering Cuban, non-Cuban varieties, and common examples.

2 Related Work

Common Examples. The challenge of handling common examples that can be valid across multiple language varieties has been a recurring issue in language variety identification. Traditional single-label classification often struggles to assign unique labels to common examples(Althobaiti, 2020; Bernier-colborne et al., 2023). Addressing this challenge, Zampieri et al. (2024) introduced a third class specifically for common instances in their DSL-TL dataset for language variety identification. This dataset allowed the exploration of the impact of these ambiguous cases on model performance. The authors found that the models had difficulty distinguishing between common and dialect-specific examples. Then, their results served as a baseline for the DSL-TL shared task at VarDial 2023 (Aepli et al., 2023). In the scope of this shared task, Vaidya and Kane (2023) introduced a two-stage multilingual dialect detection system. Their approach first identifies the macro-language, followed by applying dialect-specific models to refine the classification. Although this system performed well overall, it struggled with the common examples class, where it frequently misclassified examples due to the lack of clear dialect-specific markers. The Spanish language, with its rich array of varieties, provides a particularly challenging landscape for variety identification due to the high similarity between varieties. Zampieri et al. (2024) noted that the prevalence of common examples in

Spanish is especially high. Given the significant lexical and syntactical overlap among Spanish varieties, sentences that can belong to more than one variety are frequent, making traditional classification approaches less reliable. The misclassification of these common instances not only introduces noise into the datasets but also impacts the overall performance of the models, as evidenced by the poor handling of Argentine examples in Vaidya and Kane (2023).

Multi-class Approaches for Variety Identification. In light of these challenges that affect many different languages, several works have proposed moving away from single-label classification towards multi-class or multi-label approaches for variety identification. For example, Keleg and Magdy (2023) demonstrated that many sentences could validly belong to multiple Arabic dialects, arguing for including multiple labels per instance. They introduced the Expected Maximal Accuracy (EMA) metric to measure the upper-bound accuracy in scenarios where common instances occur frequently. Their results indicated that the majority of false positives in traditional single-label classifiers were, in fact, not errors, but cases where multiple dialects could be correct. Bernier-colborne et al. (2023) took this further by employing similarity metrics to identify duplicate or nearly duplicate examples and assigning multiple labels to ambiguous sentences. Their work, focusing on French varieties, showed that this multi-class approach significantly improved F1-macro scores for ambiguous examples. They argued that applying a multi-class framework can improve the accuracy of variety identification and better handle the inherent ambiguity found in multilingual datasets.

3 Task Definition: Automatic Common Examples Detection

Our main task is to identify common examples across similar language varieties. Our proposed pipeline can be separated into two main stages:

- Fine-tune a Transformer-based model on the Variety Identification datasets for single-label classification of varieties (binary).
- Assign a score to each example using a scorer model, expecting higher values for common examples, and rank them with the highest scores at the top.

3.1 Scorer Models

Datamaps Swayamdipta et al. (2020) proposed Datamaps (DM) using Training Dynamics, which is the behavior of a model as training progresses, for detecting annotation errors in datasets. Their approach focused on tracking the confidence and variability on the gold label during training. Specifically, examples consistently showing low confidence for this label across epochs were flagged as potential annotation errors or ambiguous cases. This technique has also been adapted to identify the variation of human labels, where examples can legitimately belong to more than one category (Weber-Genzel et al., 2024). We use this technique to identify common examples for the Variety Identification task.

Datamaps using predicted label probability We adapt the Datamaps metrics to our use case. Unlike Swayamdipta et al. (2020), who focus on the gold labels, and Weber-Genzel et al. (2024), who prioritize re-annotating erroneous labels, our goal is to detect instances that the model struggles to classify consistently. Therefore, we calculate confidence and variability differently: rather than focusing on the correctness of assigned labels or identifying annotation errors, we calculate these metrics based on the maximum predicted probability for each instance at each epoch, aiming to detect instances that exhibit inconsistent predictions or low confidence and, therefore, could belong to both classes or an unobserved third class. For common examples, which can be associated with more than one label, it would be more natural to describe the uncertainty in terms of the model's confidence in its predictions. The *confidence* is defined as:

$$\text{DM}_{\text{mean-pred}} = -\frac{1}{E}\sum_{e=1}^{E}\max_{j}(p_{i,j,e}) \quad (1)$$

where $p_{i,j,e}$ is the probability assigned to the i'th instance for the label j in epoch e. Then, the lowest confidences correspond to a higher score because of the negative sign. The idea is that examples with small probabilities associated with the predicted label across the epochs are likely challenging examples.

The *variability* is defined as:

$$\text{DM}_{\text{std-pred}} = \sqrt{\frac{1}{E}\left(\sum_{e=1}^{E}\max_{j}(p_{i,j,e}) + \text{DM}_{\text{mean-pred}}\right)^2} \quad (2)$$

The high *variability* indicates that the model's confidence changes significantly across epochs, suggesting the model is uncertain about the instance. This can point to an instance that is hard to classify or potentially common.

Random baseline We use a random model as a scorer, which assigns uniformly random scores between 0 and 1 to each example as a baseline.

Language Model For the Variety Identification module we use the model BETO, a monolingual Spanish BERT model version (Canete et al., 2020) for our experiments; it has proven effective in several downstream tasks for this language. This model was trained on all Wikipedia and all Spanish data from the OPUS project (Tiedemann, 2012). In the case of Spanish Wikipedia, by 2017, around 39.2% of edits came from Spain (Spanish Wikipedia, 2021), which can negatively impact the model performance in varieties not from Spain.

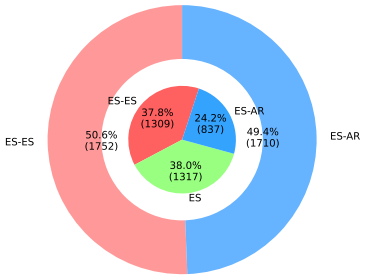
Evaluation The first metric considered for evaluation is the Average Precision Score in the Common Examples Identification Task. In addition, we evaluate precision and recall by considering the top N instances, ranked by their score values, with N ranging from 10 to the size of each dataset.

4 Datasets

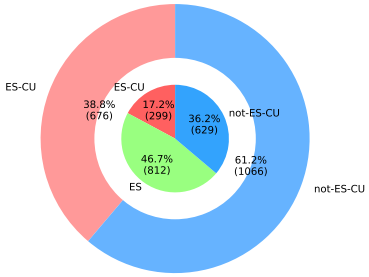
In this section, we describe the datasets used for our analysis. We use an existing dataset DSL-TL and propose a new dataset CUBANSPVARIETY focused on the Cuban Spanish variety.

4.1 DSL-TL

The Discriminating Similar Language - True Labels (DSL-TL) dataset (Zampieri et al., 2024) was employed in a shared task at the VarDial 2023 workshop[2]. This dataset contains examples from Portuguese, Spanish, and English varieties, but our focus is solely on the Spanish subset. The Spanish subset is derived from the DSLCC dataset (Tan et al., 2014) and includes sentences extracted from various Argentinian and Spanish newspapers, with each example annotated based on the country associated with the news source. However, annotating the examples with a single label proved difficult, even for human annotators (Goutte et al., 2016). Specifically, Spanish annotators achieved an average accuracy of only 54.90%. To address these challenges, Zampieri et al. (2024) randomly sampled the Spanish, Portuguese, and English subsets and conducted a new round of human annotations. In addition to the original binary labels, a third label—*both or neither*—was introduced. This additional label was assigned when annotators were unable to identify the characteristics of the different varieties. For our experiments with the DSL-TL dataset, we use the newly introduced labels from the DSL-TL dataset and the original labels from the DSL-2014 corpus. It allowed us to simulate a scenario where new annotations would be unavailable. We only use the training set to analyze the training dynamics.



(a) DSL-TL dataset distribution.



(b) CUBANSPVARIETY dataset distribution.

Figure 2: Datasets distributions.

4.2 CUBANSPVARIETY

To our knowledge, the dataset is the first dataset for Cuban or any Caribbean Spanish variety identification. The dataset contains manually annotated tweets with variety information. We collected the data from the publicly available archive *The Twitter Stream Grab* in the website archive.org. We

[2] VarDial 2023 website.

worked particularly with data from July 2021. [3]

Data Annotation. We randomly sampled 10000 tweets from July 11th and July 12th. Among those, we finally annotated 1762 examples. We considered this time frame because of the high Twitter activity in Cuba after July 11th protest in 2021 with trending hashtags such as *#SOSCuba* or *#SOSMatanzas*. [4] Each tweet was annotated across five columns: *cuban_variety*, *not_cuban_variety*, *specific_variety*, *not_able_to_identify*, and *irrelevant*. Annotators marked *cuban_variety* if the tweet belonged to the Cuban Spanish variety and *not_cuban_variety* if it did not (cf. Section B). In case of identifying a different Spanish variety (e.g., from Spain or Chile), they were asked to annotate it in the *specific_variety* column for future work. When uncertain about the variety, they marked *not_able_to_identify*. Tweets deemed noisy or non-Spanish were marked as *irrelevant*.

We focused on three labels for analysis: *ES-CU* (Cuban variety), *not-ES-CU* (non-Cuban), and *ES* (common examples). Tweets with *cuban_variety* marked True were labeled *ES-CU*, those with *not_cuban_variety* marked True were labeled *not-ES-CU*, and tweets marked only as *not_able_to_identify* were labeled *ES*, aligning with the DSL-TL dataset. Three volunteers, native Cuban Spanish speakers with a Master's degree in Cuba, performed the annotations. Their familiarity with other Spanish varieties helped them recognize common examples. Labels were assigned when at least two annotators agreed and tweets marked as irrelevant by any annotator were discarded. Full agreement was reached for 984 examples (55.8%), partial agreement for 776 (43.5%), with disagreement in just 12 cases (0.7%). We use the full dataset for training dynamics analysis. In this case, we only have the annotations with the common examples information (i.e. not single label approach). Then, to simulate a real-world scenario with single labels, we randomly assigned each common example a label of either *ES-CU* or *not-ES-CU*. Figure 2b shows the final dataset distribution. The internal circle represents the original distribution (cf. Table 2 for an overview of lexical properties).

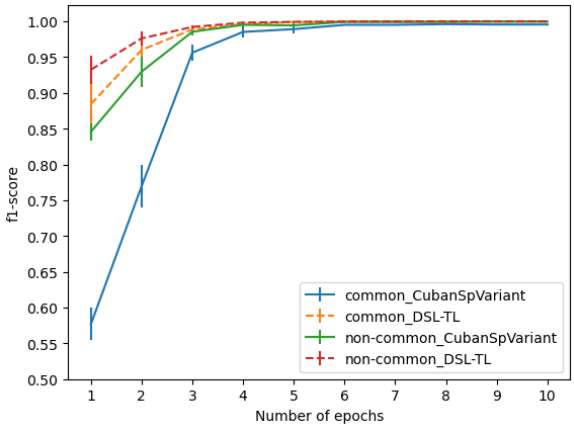


Figure 3: F1-score during training for common and non-common examples on both datasets.

5 Results

5.1 Variety Identification

We investigate the learning behavior of BETO-based Variety Identification model by analyzing the F1 scores across both datasets. Figure 3 presents the F1-score evaluation for Language Variety Classification over 10 training epochs, with separate curves for common examples and the rest of the data in both datasets. As shown in the figure, the performance gap between common and non-common examples is substantial during the early stages of training. Furthermore, the error bars indicate greater variability in the F1-scores for common examples than the rest. This gap is particularly pronounced in the CUBANSPVARIETY dataset, which exhibits lower F1 scores, likely due to the additional challenges of social media content, unlike DSL-TL, which contains sentences from newspaper articles. These observations suggest that the model finds it more challenging to learn common examples, supporting the idea that their characteristics can be identified through training dynamics.

5.2 Common Examples Identification

We present in Table 1 the results for both the DSL-TL and CUBANSPVARIETY datasets, comparing $DM_{\text{mean-pred}}$, $DM_{\text{std-pred}}$ and the random baseline. Across both datasets, the two Datamaps models significantly outperform the baseline, indicating that both capture relevant information about common examples. In addition, $DM_{\text{mean-pred}}$, which leverages the confidence in predicted labels, consistently outperforms $DM_{\text{std-pred}}$. This suggests

[3] Link to available data for July 2021.
[4] New York Times (July 11th, 2021).

Model	APS	Prec-500	Recall-500	Prec-1000	Recall-1000
			DSL-TL		
Random	39.45 ± 2.54	38.71 ± 1.49	14.98 ± 0.57	37.80 ± 1.16	28.99 ± 0.89
$DM_{mean-pred}$	**54.75 ± 1.8**	**62.78 ± 2.47**	**24.31 ± 0.95**	**57.76 ± 1.58**	**44.29 ± 1.21**
$DM_{std-pred}$	52.88 ± 3.00	58.70 ± 3.05	22.73 ± 1.18	56.03 ± 2.59	42.97 ± 1.98
			CUBANSPVARIETY		
Random	46.42 ± 1.20	46.39 ± 2.32	29.10 ± 1.46	46.83 ± 0.52	58.17 ± 0.65
$DM_{mean-pred}$	**63.51 ± 2.56**	**66.19 ± 3.43**	**41.52 ± 2.15**	**59.16 ± 1.25**	**73.50 ± 1.55**
$DM_{std-pred}$	61.97 ± 2.60	64.86 ± 3.59	40.68 ± 2.25	58.15 ± 1.07	72.25 ± 1.33

Table 1: Evaluation metrics for Automatic Common Examples on DSL-TL and CUBANSPVARIETY datasets. We present the Average Precision Score, equivalent to the area under the precision-recall curve, and the precision and recall for Top-500 and Top-1000 instances. All the metrics are expressed in percentages.

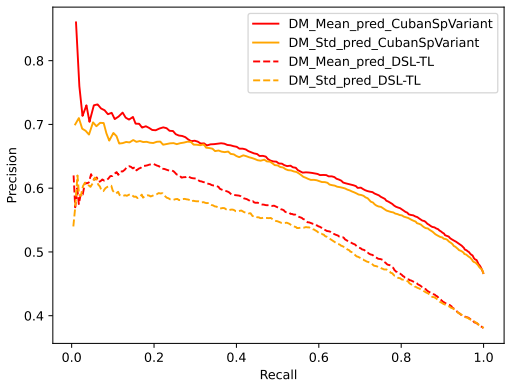


Figure 4: Precision versus recall curve

that the model's average confidence offers a more reliable signal for identifying common examples, while the variability-based approach ($DM_{\text{std-pred}}$) tracks changes that do not always correspond with common examples. We observe that the difference in performance between the two datasets follows a similar pattern across all models, including the random baselines. This is likely due to the proportion of common examples in each dataset. In DSL-TL, where 38% of the examples are common, the random baseline precision is close to 38%. Similarly, in CUBANSPVARIETY, with 46% common examples, the baseline precision is near 46%. This suggests that the metrics' ranges are closely tied to each dataset's distribution of common examples.

Figure 4 shows both datasets' precision versus recall curves. In both cases, precision remains relatively stable in the early ranking stages and begins to converge toward the common examples' proportion as recall increases. The performance difference between $DM_{\text{mean-pred}}$ and $DM_{\text{std-pred}}$ is more pronounced for smaller values of N, particularly in precision. However, the recall curves show a steeper slope at earlier ranking stages, which gradually decreases as N increases, consistent with expected behavior.

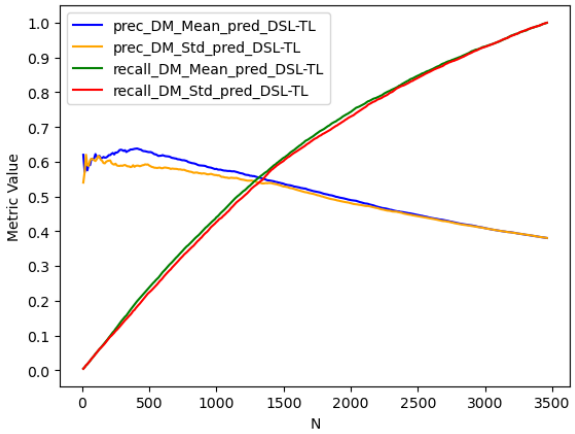


Figure 5: Precision and Recall versus Top-N instances DSL-TL dataset

Figure 5 highlights that in the DSL-TL dataset, which contains clean, edited content unlike our Twitter-based Cuban dataset, $DM_{\text{mean-pred}}$ identifies common examples early in the ranking. This is likely because we had access to the original labels for common examples in this dataset, reducing noise. Furthermore, the clear class boundaries distinguishing Spanish varieties from Spain and Argentina likely contributed to the model's more stable performance, while $DM_{\text{std-pred}}$ is less effective in this context. In Figure 6, we observe that for the CUBANSPVARIETY dataset, which contains more dynamic and informal language from user-generated content, the performance gap between $DM_{\text{mean-pred}}$ and $DM_{\text{std-pred}}$ becomes smaller. This indicates that variability has a more significant impact on identifying common examples in user-generated content. In this dataset, common examples were identified in the first round and randomly assigned to Cuban or non-Cuban classes,

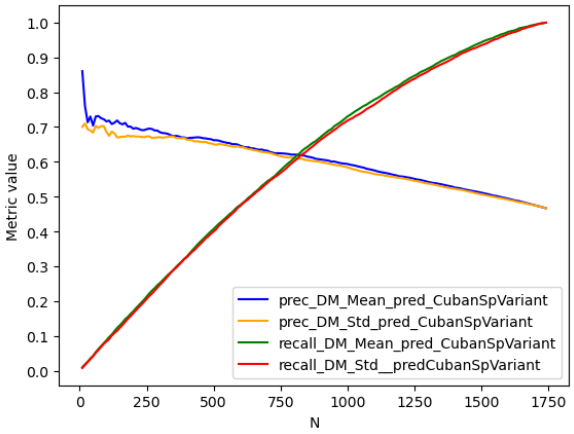


Figure 6: Precision and Recall versus Top-N instances CUBANSPVARIETY dataset

increasing ambiguity. It is worth noting that, beyond the differences in the nature of the dataset (newswire text vs. Twitter user-generated content), the collection period dates vary over six years between both datasets, likely affecting model performance since languages evolve and are shaped by social dynamics. Furthermore, the Cuban dataset includes tweets from July 11th and 12th, during large protests in Cuba that were trending among Spanish-speaking countries. This may introduce biases into the dataset and influence the variety identification.

6 Error Analysis

To better understand our models' performance, we analyzed the errors for each dataset by counting the most frequent words in the Top-500 non-common instances predicted by the $DM_{mean-pred}$ model (prediction errors). After removing stopwords and special tokens, we found that in the CUBANSPVARIETY dataset, the most frequent words were *Cuba* and *SOSCuba*, directly tied to the Cuban variety in this context. In contrast, the DSL-TL dataset showed common words like *ha*, *pero*, *fue*, and *también*, which do not indicate a specific variety. The topic bias in the Cuban dataset can influence the model predictions, mainly when the examples contain keywords specific to the variety. This also explains why $DM_{std-pred}$ performs better for CUBANSPVARIETY, as these keywords in both classes make variability more significant than in DSL-TL.

In the CUBANSPVARIETY dataset, Figure 7 shows that about 67% of the Top-500 non-common examples and 54% of the Top-1000 non-common examples contained the word *Cuba*, suggesting a strong influence on model behavior, given that only 33% of the total examples contain this word. Additionally, we found that 63.31% of the Top-500 errors in CUBANSPVARIETY were cases where only two annotators agreed on the label, and for the Top-1000, this number was 57%. Across all non-common instances, full agreement (three annotators) occurred in 57% of cases, indicating a clear link between annotation difficulty and model errors as shown in Figure 8.

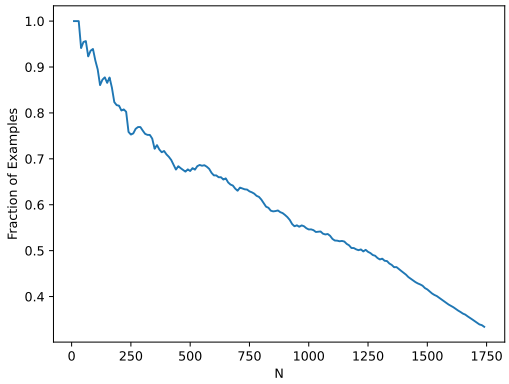


Figure 7: Fraction of error instances containing the word *Cuba* in Top-N instances using DM_mean score metric.

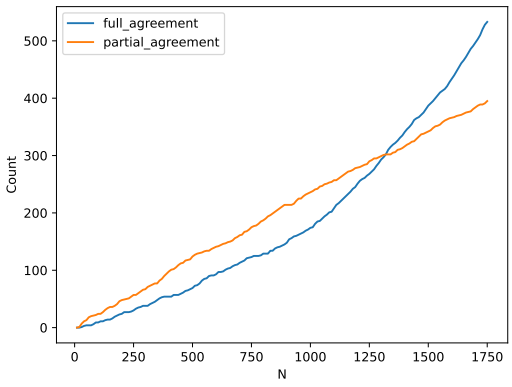
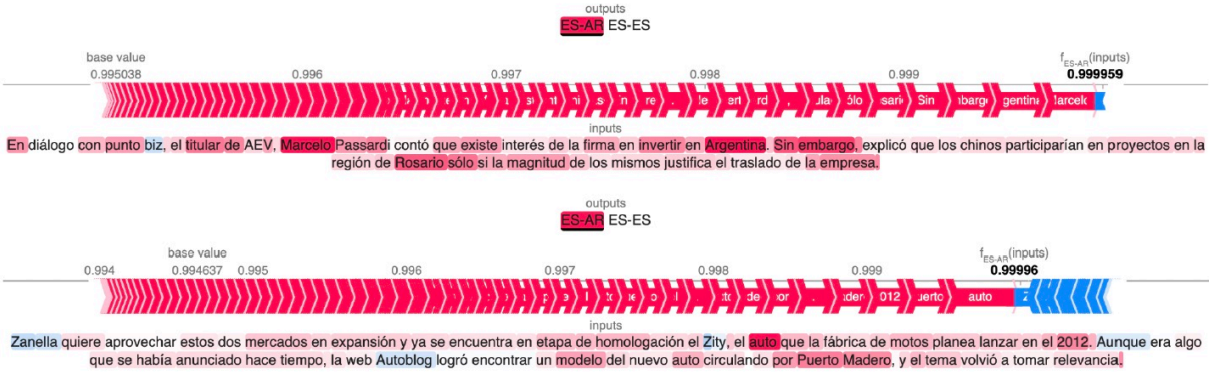
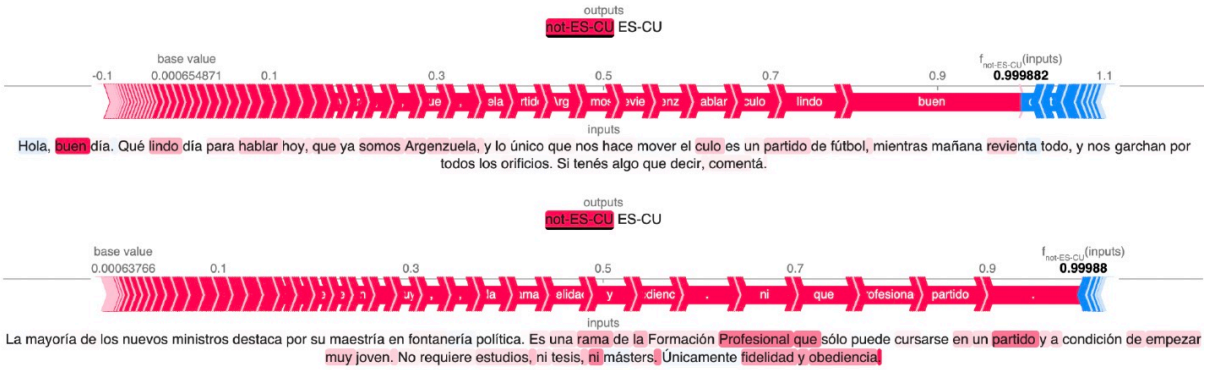


Figure 8: Agreement index for error instances in Top-N using DM_{mean} score metric.

Another key point is understanding why the model fails to retrieve certain common examples. We focus on the last two common examples in the ranking for each dataset, using SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to analyze the model's behavior. SHAP is based on Game Theory and assigns importance scores to features, showing how much each feature influ-



(a) DSL-TL dataset examples.



(b) CUBANSPVARIETY dataset examples.

Figure 9: For each dataset, we analyze the last two common examples in the ranking obtained using $DM_{mean-pred}$. The model is trained on binary classification for variety detection. The final output of the models for the predicted variety/class is highlighted. Red-colored terms influence the final decision towards *ES-ES* or *ES-CU* depending on the dataset, while blue-colored terms influence the model classification towards *ES-AR* or *not-ES-CU* classes.

ences the model's prediction. Figure 9a presents the SHAP scores for the last two common examples in the DSL-TL dataset ranking. For the first example, the words *Argentina*, *Rosario*, and *Marcelo* are the most influential for predicting the *ES-AR* label. The first two refer to the country and one of its major cities, while *Marcelo* is a common name in Argentina. For the second example, *auto* (commonly used in Argentina to mean "car," as opposed to *coche* in Spain) is the most significant feature, followed by *Puerto* and *Madero*, a well-known place in Argentina. While named entities influence the first example, the second example, with the word *auto*, suggests a potential annotation error, as it points to the Argentinian variety.

In Figure 9b, we provide the corresponding analysis for the CUBANSPVARIETY dataset. For the first example, the word *buen* (from the phrase *buen día*, which is used in Spanish varieties other than the Cuban one) is the most significant, along with *Argenzuela* (a blend of Argentina and Venezuela), *garchan*, and *tenés*, which are characteristic of the Argentinian variety. This example likely represents an annotation mistake. For the second example, the most influential words are *profesional*, *partido*, *fidelidad*, and *obediencia*, none of which are strong indicators of a specific variety. This suggests that common topics in Cuban tweets may affect the model's prediction, potentially introducing biases into the classification process.

Regarding the named entities, we followed the precedent set by previous works in variety Identification, such as the study introducing the DSL-TL dataset (Zampieri et al., 2024), retained named entities. Consequently, we included them in our initial approach **while emphasizing the importance of analyzing their influence**. We agree that a set of experiments where we could switch the named

entities with neutral entities (or even adversarial entities (eg switch SosCuba with SosMexico) would be interesting. In our case, while evidence suggests that named entities contribute to model errors, **our preliminary analysis demonstrates the model's robustness to their presence**. For example, the sentence "Mi mensaje para el pueblo de cuba emoji bandera cuba emoji :. ¡No están solos!. Cuenten con nosotros para seguir apoyando su lucha por la libertad y la democracia. soscuba url" was ranked second using the Datamaps mean approach. Although it contained clear markers such as "Cuba" and "soscuba," the model correctly identified it. This is not an isolated case, and further analysis of correctly classified examples can provide additional evidence of the system's robustness.

7 Conclusion

In this work, we examine the effectiveness of Datamaps methods in identifying common examples across closely related language varieties. Our results demonstrate the value of training dynamics in detecting difficult examples early in the model's learning process, as reflected by the effectiveness of $DM_{\text{mean-pred}}$ across both datasets. This confidence-based approach consistently outperformed the variability-based method, suggesting that tracking model confidence over predicted labels offers a reliable way to identify common examples automatically across different domains. Although the performance difference between variability-based and confidence-based approaches is less significant for theinformal dataset, the overall results indicate that confidence-based Datamaps can be a powerful tool for improving data quality in different contexts.

Although these methods may not fully solve the challenges of variety and dialect annotation, they offer a promising step forward, particularly when combined with automatic techniques and targeted human annotation.

We hope that this initial dataset, freely accessible under a CC-BY-SA license upon publication, the first centered on Cuban, a Caribbean variety of Spanish, will prove a valuable resource for future research on this topic.

8 Limitations

One limitation of our work is that the analysis focuses on binary classification scenarios, explicitly distinguishing between two main classes in each dataset without incorporating multi-class approaches or more complex variety distinctions. While this setup allows us to study common examples effectively, expanding the approach to multi-variety settings could provide a more comprehensive understanding of the challenges posed by language variety identification.

Another limitation is inherent in the way the annotations in the CUBANSPVARIETY dataset were built. Since all annotators were Cuban native speakers, the dataset focuses on Cuban versus non-Cuban distinctions. Incorporating annotators from other Spanish-speaking regions would allow for broader variety distinctions and more nuanced annotations, which could reduce potential biases introduced by a single-region perspective. **However, the framework for annotations was designed with enough flexibility to make it extensible for further annotations in variants different from Cuban** with the final aim of creating a dataset which cover most of the Spanish varieties. In this scenario, common examples between specific varieties will be determined by overlapping between annotation made by native speakers from each variant.

Finally, as discussed in Section 6, named entities, including hashtags, play a significant role in model behavior. Managing these entities, such as replacing them with special tokens, could be an effective way to reduce bias and improve generalization. This is especially important in tasks like language variety classification, where named entities might disproportionately influence predictions.

9 Ethical Considerations

This work involves using social media data, particularly from Twitter, which may contain sensitive or controversial content. Although we anonymize the data by replacing user mentions and URLs, the content could still involve personal opinions, political statements, or even hate speech, especially in datasets like the CUBANSPVARIETY dataset, which includes tweets related to politically sensitive events such as the July 11th protests in Cuba. Given the nature of the protests, some tweets may contain offensive content. We are aware of the potential privacy implications when working with such data, and we have adhered to Twitter's data usage policy to ensure compliance with ethical standards. Researchers accessing this dataset should consider the ethical implications when using politically charged content or messages that might harm

individuals or communities.

Furthermore, identifying language varieties, especially in socially and politically sensitive contexts, risks reinforcing stereotypes or biases associated with particular regions. In this work, we frame our approach as a technical solution for linguistic diversity and not as a tool for making any sociopolitical or cultural assumptions about the speakers of these varieties. However, we acknowledge that any automated system trained on real-world data is susceptible to unintended biases arising from imbalanced datasets or biased annotations. The annotations in the CUBANSPVARIETY dataset are from native Cuban speakers, and while this helps in identifying Cuban Spanish, it may introduce a regional bias.

Acknowledgments

This work received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101021607. The authors warmly thank the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Maha J. Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey. *ArXiv*, abs/2009.12622.

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2021. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *CoRR*, abs/2104.14728.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, choclo and mazorca: on the varieties of Spanish. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3689–3711, Mexico City, Mexico. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

R. A. Hudson. 1996. *Sociolinguistics*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, Brisbane, Australia.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Juan Manuel Pérez, Damián A. Furman, Laura Alonso Alemany, and Franco Luque. 2022. Robertuito: a pre-trained language model for social media text in spanish. *Preprint*, arXiv:2111.09453.

Spanish Wikipedia. 2021. Spanish wikipedia — Wikipedia, the free encyclopedia. [Online; accessed 8-March-2022].

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15. Citeseer.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. CLTeam1 at SemEval-2024 task 10: Large language model based ensemble for emotion detection in Hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 365–369, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Vaidya and Aditya Kane. 2023. Two-stage pipeline for multilingual dialect detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 222–229, Dubrovnik, Croatia. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Marcos Zampieri and Preslav Nakov. 2021. *Dialect and Similar Language Identification*, page 187–203. Studies in Natural Language Processing. Cambridge University Press.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Mahesh Bangera. 2024. Language variety identification with true labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10100–10109, Torino, Italia. ELRA and ICCL.

A Data Preprocessing:

Following previous works (Pérez et al., 2022; Castillo-lópez et al., 2023), we pre-processed the data by replacing user mentions with the token *@usuario* (or *@user* in English), allowing up to two consecutive mentions. URLs were substituted with the token *url*, and hashtags were segmented into words assuming Camel Case typing (e.g., #CubaIslaBella becomes *Cuba isla bella*). Emojis were replaced with their corresponding descriptions using the *emoji* python library [5], and any repeated emojis were removed. Laughs were normalized to *jaja*, following the standard in Spanish, and for letter repetitions, we kept up to two. We also removed repeated spaces and replaced line breaks with periods.

#sentences	1762
#tokens	41374
Avg length	23.48
Length variation (std)	13.49
Vocab size (unique words)	13336

Table 2: DSL-TL Overview.

B Annotation Guidelines for CubanSpVariety

The following guidelines were provided to the annotators to ensure consistent labeling of the dataset:

- **cuban_variety**: A boolean value indicating whether the tweet belongs to the target Spanish variety (Cuban). This value should be set to *true* only if the annotator can clearly identify evidence that the tweet belongs to the Cuban variety.
- **not_cuban_variety**: A boolean value indicating that the tweet does not belong to the target Cuban variety. This value should be set to *true* only if it is clear that the tweet does not belong to the Cuban variety, even if the specific variety cannot be identified.
- **specific_variety**: A string indicating the specific variety if the annotator can easily identify it. The value should remain empty if the specific variety cannot be identified. The possible varieties are based on the Spanish varieties map presented in the appendix of *Analyzing Zero-Shot Transfer Scenarios Across Spanish Variants for Hate Speech Detection*. These are:
  - Other Caribbean variety
  - Central American varieties (Costa Rica, El Salvador, Panamá)
  - Mexican
  - Spain
  - Rioplatense (Argentina, Uruguay)
  - Chilean
  - Habla de las tierras altas (Perú, Venezuela, Colombia, Bolivia, Ecuador)
- **unable_to_identify_variety**: A boolean value set to *true* if the annotator cannot identify any specific variety for the tweet.
- **irrelevant**: A boolean value set to *true* if the tweet's content is considered irrelevant. This can be due to the tweet's size or other characteristics that lead to a lack of meaningful content.

Annotator	Age	Gender
Annotator 1	26	Male
Annotator 2	26	Female
Annotator 3	23	Female

Table 3: Socio-demographic attributes of the annotators

These annotations guidelines are extensible for speakers form varieties different from Cuba by changing the variety target. It makes it possible to extend the varieties covered in the dataset in a direct way.

C Hyper-parameters

The model will be released under the Creative Commons CC-BY-SA license, allowing for open access and use with appropriate attribution.

All experiments were conducted using a single NVIDIA RTX 8000 GPU, with each experiment taking less than two hours to complete. We used the `AutoModelForSequenceClassification` from Hugging Face's Transformers library (Wolf et al., 2020) for sequence classification tasks.

D Variety Identification Results

D.1 Variety Identification Benchmarks on CubanSpVariety dataset

In this section, we present the benchmark results for the CUBANSPVARIETY dataset. We use the same

[5] Emoji python library website.

Hyper-parameter	Value
Max sequence length	512
Batch size	32
FP16	Enabled
Learning rate	1e-5
Epochs	10
Scheduler	linear
Warmup ratio	0.1
Weight decay	0.01
Save strategy	Epoch
Logging steps	10
Seed	{42,151,2021,15,98}

Table 4: Hyper-parameters used for the fine-tuning.

experimental setting for this task, as explained before. We present the dataset's benchmark for both approaches, single and multi-class. For the multi-class approach, we follow the procedure suggested by Keleg and Magdy (2023); Bernier-colborne et al. (2023) of using one binary classifier per label. For the metrics, we used the macro average across all possible varieties.

Table 5 shows the final results. We can notice a significant improvement in the model's performance in the multi-class scenario. This strengthens the point about single-class approach limitations for variety identification.

D.2 Variety Identification Benchmarks on DSL-TL dataset

In this section, we present the benchmark results for the DSL-TL dataset. Table 6 shows the final results. As for the CUBANSPVARIETY dataset, there is a significant improvement in the model's performance in the multi-class scenario.

Approach	Acc	Precision	Recall	f1-score
single-class	67.54 ± 1.42	65.86 ± 1.69	64.45 ± 1.01	64.62 ± 1.05
multi-class	**78.69 ± 0.86**	**82.64 ± 0.91**	**87.80 ± 1.28**	**85.06 ± 0.61**

Table 5: Benchmarks for Variety Identification task on CUBANSPVARIETY dataset. We present the results for both the single-class and the multi-class approaches.

Approach	Acc	Precision	Recall	f1-score
single-class	76.76 ± 0.74	76.18 ± 0.80	75.78 ± 0.75	76.76 ± 0.74
multi-class	**77.65 ± 0.27**	**82.00 ± 0.29**	**83.99 ± 0.30**	**82.97 ± 0.25**

Table 6: Benchmarks for Variety Identification task on DSL-TL dataset. We present the results for both the single-class and the multi-class approaches.