Language Identification for Low-Resource Formosan Languages

Henry Gagnier
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

Abstract

Formosan languages are a critically endangered group of Austronesian languages spoken in Taiwan, with severely limited representation in natural language processing (NLP) research and no support in existing language identification (LID) tools. We present the first systematic evaluation of machine learning models for the language identification of Kavalan, a Formosan language with fewer than 300 known speakers. We construct two benchmarks: a deployment-oriented benchmark with languages commonly confused with Kavalan by existing tools, and a linguistically motivated benchmark of typologically related Formosan languages. We evaluate random forest, support vector machine (SVM), and three pre-trained multilingual models using repeated stratified cross-validation. SVM models with character *n*-gram features achieve the strongest performance on both benchmarks, with a macro F1 of 0.993 on the deployment benchmark and a macro F1 of 0.906 on the Formosan benchmark, while remaining computationally inexpensive and effective with a low amount of data. Pre-trained multilingual models degrade significantly on the Formosan benchmark, with XLM-RoBERTa falling to a macro F1 of 0.505. These results demonstrate that traditional *n*-gram-based approaches are effective with low-resource Formosan LID and establish a foundation for downstream NLP tasks supporting the documentation and revitalization of low-resource Formosan languages.

Keywords: Formosan languages, low-resource languages, Kavalan, language identification

1. Introduction

The loss of indigenous languages threatens the preservation of unique indigenous identities, vital knowledge, cultural heritage, and the well-being of indigenous communities (Ajani et al., 2024). New natural language processing (NLP) and machine learning technologies have enabled the documentation, preservation, and revitalization of indigenous languages (Ngoc Le and Sadat, 2020). Language identification of indigenous languages is often difficult due to the low-resource nature of many indigenous languages, leading to insufficient training data (Haas and Derczynski, 2020).

Language identification (LID) is the task of automatically determining the language of a given text (Jauhiainen et al., 2019). Despite advances in LID for high-resource languages, LID for low-resource and endangered languages remains a significant challenge due to the scarcity of training data and the difficulty of constructing representative corpora (Haas and Derczynski, 2020). Additional challenges include language confusion arising from lexical or structural similarity between related languages, as well as code-switching. Recent work has begun to address LID for low-resource and indigenous languages specifically. Bestgen (2017) use a SVM approach to rank first in the Discriminating between Similar Languages (DSL) task. Yang et al. (2025) evaluated random forest on the identification of Navajo, an Athabaskan language. Sindane and Marivate (2024) evaluated the use of *n*-grams with naive Bayes, and the use of pre-trained multilingual models, including Afri-centric models, on the identification of low-resource South African languages.

Kavalan is a seriously endangered Formosan language spoken in southeastern Taiwan (Hsieh and Huang, 2007). Despite being claimed moribund, Blust (2010) reports that Kavalan had around 300 speakers in 1994. Although research on Kavalan in NLP is limited, recent work has started to focus on the processing of Formosan languages. Lin et al. (2025) evaluated LLMs on the translation, summarization, and automatic speech recognition of Atayal, Amis, and Paiwan. Zheng et al. (2022) developed a parallel corpus for Amis-Mandarin translation. Kargaran et al. (2023) included Formosan languages such as Rukai and Amis in a model identifying low-resource languages. Zheng et al. (2024) investigated the machine translation of Formosan languages using bilingual lexical resources, including the translation of Kavalan to Mandarin. While Kavalan and Formosan languages have been introduced to NLP, current work faces challenges, is limited, and has not focused on the language identification of Formosan languages (Lin et al., 2025).

The purpose of this study is to (1) evaluate traditional models and pre-trained multilingual models on the language identification of Kavalan and (2) identify potential challenges with the identification of Kavalan and other endangered Formosan languages. This work aims to improve the inclu-

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 242–252
11–12 May 2026.

sion of low-resource and endangered languages in NLP through the evaluation of machine learning models for the identification of Kavalan texts.

2. Materials and Methods

2.1. Data and Preprocessing

We will now discuss the data resources needed: a corpus of sentences in Kavalan and a benchmark of sentences in other languages for language identification benchmarking.

2.1.1. Kavalan Corpus

We used a Kavalan text corpus from the ePark corpus (Indigenous Languages Research and Development Foundation, 2025), which contains high-quality, officially produced, and verified language educational material, developed by the Indigenous Languages Research and Development Foundation as part of the FormosanBank project (Scheppat et al., 2025) . Daily conversations, and reading and writing sentences were selected, resulting in 326 Kavalan sentences of varying length and content. Table 1 provides illustrative examples of Kavalan sentences from the ePark corpus.

2.1.2. langdetect Confuser Languages Benchmark Construction

In order to identify the languages similar to Kavalan based on existing language identification systems, we classified all Kavalan (ckv) sentences using the langdetect 1.0.9 package [1], a direct port of Google's language detection library, to create a realistic and challenging classification scenario with languages that Kavalan is commonly confused with. langdetect supports 55 languages, none of which are Formosan languages. Corpora of sentences from the ten languages Kavalan was most commonly misidentified as, consisting of 326 sentences each or the same size as the Kavalan corpus, were gathered from the Tatoeba corpus (Tiedemann, 2012). The Tatoeba corpus was selected due to its multilingual coverage, including all ten confuser languages, its sentence-level format compatible with our experimental setup, and its established use in prior LID research. Sentences from the Tatoeba corpus containing symbols and numbers other than punctuation were not used to maintain uniformity with the ePark corpus. The ten languages identified were Indonesian (id, msa), Tagalog (tl, tgl), Swahili (sw, swa), Somali (so, som), Albanian (sq, sqi), Finnish (fi, fin), Estonian (et, est), Latvian (lt, lav), Italian (it, ita), and Welsh (cy, cym). We acknowledge that these confuser languages were selected based on the output of an existing tool rather than linguistic analysis, and that several (e.g., Latvian, Somali) are typologically unrelated to Kavalan. This benchmark, therefore, reflects a realistic deployment scenario rather than a typologically motivated challenge. Figure 1 displays the distribution of langdetect misclassifications, and Figure 2 displays the sentence length distribution across the eleven languages.

[1] https://pypi.org/project/langdetect/

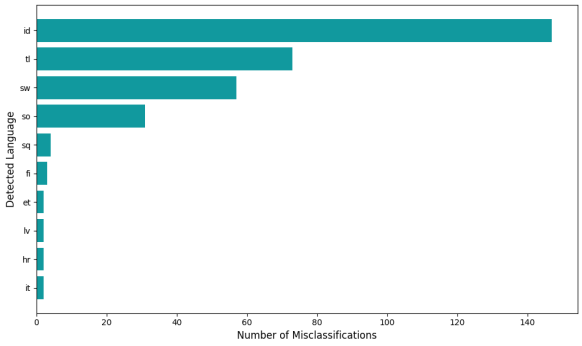

Figure 1: Distribution of languages that Kavalan sentences were most frequently misclassified as by langdetect, labeled with ISO 639-1 codes

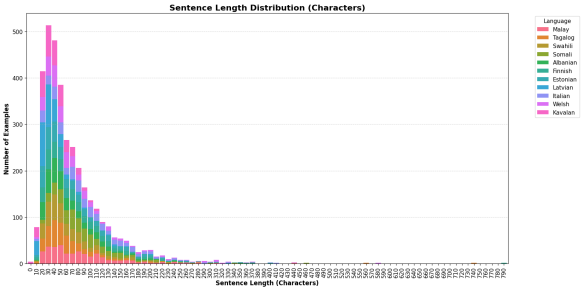

Figure 2: Distribution of the sentence length in characters of the eleven identified languages

2.1.3. Formosan Languages Benchmark Construction

To evaluate model performance in a linguistically motivated and more challenging setting, we constructed a second benchmark using other Formosan languages from the ePark corpus. Sentences in Kavalan were evaluated against the ten Formosan languages with the most available data in ePark, similarly drawn from the daily conversation, reading and writing sentence types. The ten Formosan languages selected were Amis (ami), Bunun (bnn), Rukai (dru), Paiwan (pwn), Puyuma (pyu), Thao (ssf), Sakizaya (szy), Tao (tao), Atayal (tay), and Truku (trv), representing six distinct branches of the Formosan family. Amis and Sakizaya are the closest relatives of Kavalan within the East Formosan branch, while Tao is typologically the most dis-

Kavalan	Mandarin	English
padadames pa ita padadames pa ita aita na kebalan	加油吧！加油吧！我們噶瑪蘭	Come on! Come on! We are Kavalan
aiku seRia suwani nizu maitis	我和弟弟都很害怕	My younger brother and I are both very afraid
usiq busaRai wasu	1 隻白色的狗	One white dog
sangangay ti ya quyu temita	狐狸看得口水都要流出來了	The fox's mouth watered as he looked
upitu bultedan qemiqedat	7 顆星星閃閃爍爍	Seven stars twinkling and shining

Table 1: Example sentences from the Kavalan ePark corpus

tant, belonging to the Malayo-Polynesian rather than Formosan subgroup, providing a range of linguistic distances within this benchmark (Blust, 2013). We sampled 326 sentences per language to match the number of Kavalan sentences sampled. As all languages in this benchmark are Formosan, this benchmark represents a more challenging setup than the deployment scenario. To facilitate future research and reproducibility, we have released both benchmarks, which are publicly available at `https://github.com/henrygagnier/kavalan-LID-benchmark`.

2.2. Machine Learning Models

We now go into the machine learning models we benchmarked for Kavalan LID: random forest, a support vector machine, and three pre-trained multilingual models. All models were evaluated using repeated stratified k-fold cross-validation (CV) with 5 splits and 3 repetitions (15 folds total), using a random seed of 8. Models were trained on a NVIDIA Tesla T4 GPU. This is important given the limited amount of data for Kavalan and other Formosan languages. Model performance was compared using pairwise two-sided Wilcoxon signed-rank tests (Wilcoxon, 1945) on the 15 per-fold macro F1 scores, with a significance threshold of $\alpha = 0.05$.

2.2.1. Support Vector Machine

A linear support vector machine (SVM) is a well-established, strong baseline for language identification (Joachims, 1998; Kruengkrai et al., 2005). We evaluated SVMs with TF-IDF character *n*-gram features of length 3 to 5 with vocabulary sizes of 5,000 *n*-grams and 50,000 *n*-grams. SVM was implemented using LinearSVC in the scikit-learn 1.6.1 library (Pedregosa et al., 2011) with $C = 1.0$.

2.2.2. Random Forest

We implemented random forest (RF) with 200 trees using the scikit-learn 1.6.1 library (Pedregosa et al., 2011) in Python 3.12. Features were extracted from Kavalan sentences using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. Character-level analysis and character *n*-grams of length 3 to 5 were extracted with the vocabulary limited to 5,000 *n*-grams. This approach often increases context understanding and feature representation (Setiawan et al., 2024), capturing sub-word patterns that are useful for identifying low-resource languages with limited vocabulary data.

2.2.3. Pre-trained Multilingual Models

Transformer models have emerged as state-of-the-art solutions for NLP tasks, excelling in understanding and generating natural language (G et al., 2023). Multilingual BERT (mBERT, `bert-base-multilingual-cased`) (Pires et al., 2019; Devlin et al., 2019), XLM-RoBERTa (XLM-R, `xlm-roberta-base`) (Conneau et al., 2020), and rebalanced mBERT (RemBERT, `google/rembert`) (Chung et al., 2021) are pre-trained multilingual models, each trained in at least 100 languages and performing well on multilingual benchmarks. mBERT, XLM-R, and RemBERT were implemented with the transformers 4.57.1 library in Python 3.12 with a batch size of 8, learning rate of 2e-5, 1 training epoch, weight decay of 0.01, and FP16 enabled. The models were also trained using repeated stratified 5-fold CV with 3 repeats, similarly to the random forest classifier.

3. Results

3.1. langdetect Confuser Languages Benchmark

We first look at the overall performance of the models on the langdetect confuser languages benchmark (Table 2). All models achieved fairly high accuracy, with SVM (50k vocab) performing the best overall with an accuracy of 0.993. SVM (5k vocab) had a similar accuracy of 0.989. Among the pre-trained multilingual models, RemBERT performed the best with an accuracy of 0.985, marginally above mBERT and XLM-R. Random forest was the weakest performer overall, with an accuracy of 0.949, while still being competitive with other

models despite its architectural simplicity. We report pairwise Wilcoxon signed-rank tests on per-fold macro F1 scores (Table 4), indicating that all model pairs differed significantly ($p < 0.05$) with the exception of mBERT and XLM-R and mBERT and RemBERT. SVM (50k vocab) significantly outperformed every other model.

Model	Accuracy	Macro F1
Random Forest	0.949 ± 0.007	0.949 ± 0.007
SVM (5k vocab)	0.989 ± 0.003	0.989 ± 0.003
SVM (50k vocab)	**0.993 ± 0.003**	**0.993 ± 0.003**
mBERT	0.982 ± 0.004	0.982 ± 0.004
XLM-R	0.980 ± 0.005	0.980 ± 0.005
RemBERT	0.985 ± 0.006	0.985 ± 0.006

Table 2: Performance of all models on the langdetect-based benchmark (mean ± SD across 15 folds)

We now look at the results of the classification of Kavalan in the same setup, which highlights more clearly differences in the classification of Formosan languages (Table 3). SVM (50k vocab) achieved a perfect precision and an F1 of 0.997. SVM (5k vocab) also had a strong performance with an F1 of 0.991. This demonstrates that SVMs are able to generalize to low-resource Formosan languages. XLM-R and RemBERT had F1 scores of 0.929 and 0.949, respectively. These scores are both lower than their overall F1 scores, likely due to Formosan languages not being included in their training data.

Model	Precision	Recall	F1
Random Forest	0.945	0.945	0.945
SVM (5k vocab)	0.994	0.988	0.991
SVM (50k vocab)	**1.000**	**0.994**	**0.997**
mBERT	0.958	0.993	0.975
XLM-R	0.887	0.974	0.929
RemBERT	0.950	0.949	0.949

Table 3: Performance of all models on the identification of Kavalan specifically evaluated against similar languages identified by langdetect

To better understand if some classes are more challenging than others, we analyze the confusion matrices for the four models (Figures 3-8). We find that errors are spread fairly evenly among classes, although some classes were challenging for particular models. In RF, errors in Latvian and Finnish are prevalent. The SVM models had fewer errors, with no language being particularly challenging. The pre-trained multilingual models performed very well with minimal error, although increased error in Kavalan can be observed in RemBERT and XLM-R.

Model A	Model B	*p*-value	Sig.
Random Forest	SVM (5k vocab)	<0.001	*
Random Forest	SVM (50k vocab)	<0.001	*
Random Forest	mBERT	<0.001	*
Random Forest	XLM-R	<0.001	*
Random Forest	RemBERT	<0.001	*
SVM (5k vocab)	SVM (50k vocab)	0.001	*
SVM (5k vocab)	mBERT	0.003	*
SVM (5k vocab)	XLM-R	<0.001	*
SVM (5k vocab)	RemBERT	0.007	*
SVM (50k vocab)	mBERT	<0.001	*
SVM (50k vocab)	XLM-R	<0.001	*
SVM (50k vocab)	RemBERT	<0.001	*
mBERT	XLM-R	0.303	
mBERT	RemBERT	0.277	
XLM-R	RemBERT	0.018	*

Table 4: Pairwise Wilcoxon signed-rank tests on macro F1 scores across 15 folds. * denotes $p < 0.05$.

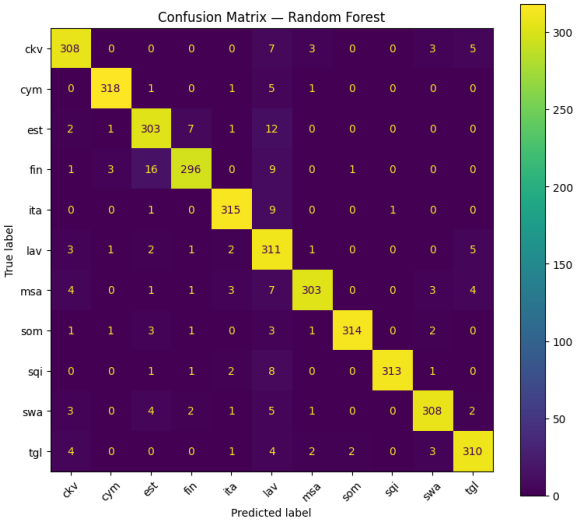

Figure 3: Confusion matrix of the classification results of random forest on the langdetect-based benchmark

3.2. Formosan Languages Benchmark

We now look at the overall performance of the models on the Formosan languages benchmark (Table 5). Overall performance dropped across all models relative to the langdetect-based benchmark, reflecting the greater difficulty of discriminating among typologically related languages. SVM (50k vocab) achieved the greatest accuracy of 0.906, followed by SVM (5k vocab) with an accuracy marginally lower at 0.898. RemBERT and random forest have comparable accuracies to the SVMs of 0.888 and 0.856, respectively. mBERT declined to an accuracy of 0.815. Most notably, the accuracy of XLM-R fell to 0.551. This is a reduction of 0.475 points, indicating that XLM-R generalizes poorly to the linguistically motivated Formosan setting. This performance is significantly

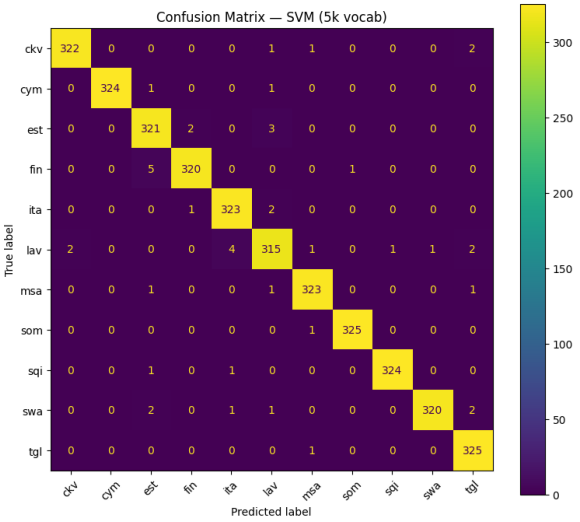

Figure 4: Confusion matrix of the classification results of SVM (5k vocab) on the langdetect-based benchmark

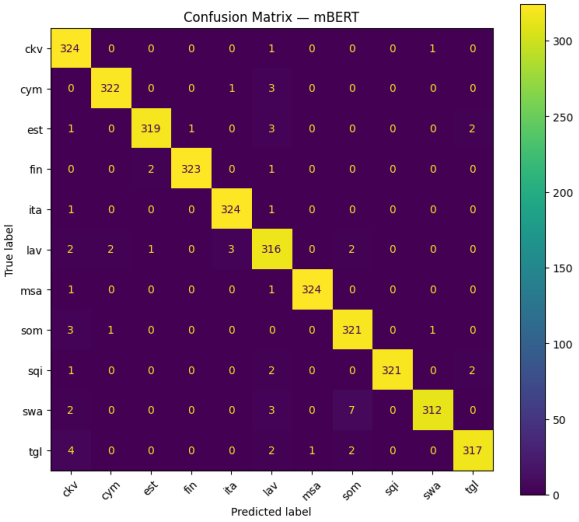

Figure 6: Confusion matrix of the classification results of mBERT on the langdetect-based benchmark

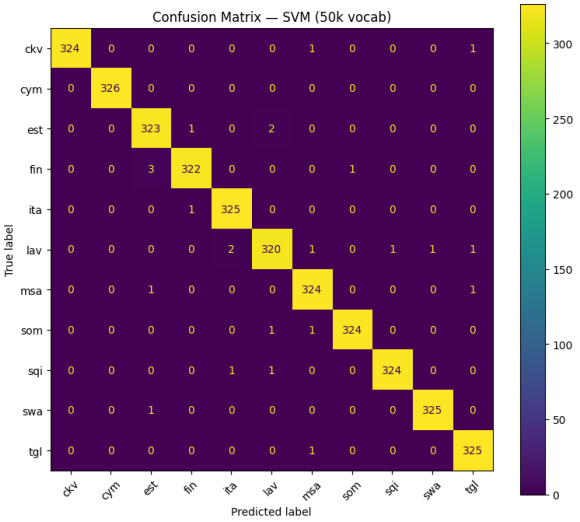

Figure 5: Confusion matrix of the classification results of SVM (50k vocab) on the langdetect-based benchmark

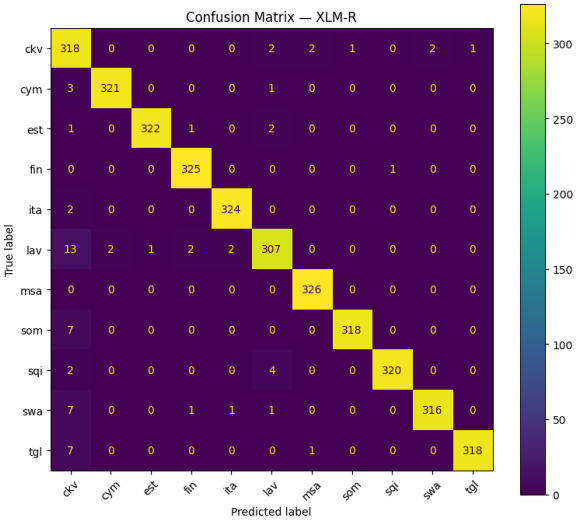

Figure 7: Confusion matrix of the classification results of XLM-R on the langdetect-based benchmark

worse than all models, including the much simpler random forest model. We find that all model pairs differed significantly besides SVM (5k vocab) and RemBERT through pairwise Wilcoxon signed-rank tests (Table 7). Zooming into the identification of Kavalan (Table 6), scores generally decrease from the langdetect-based benchmark but are higher than most Formosan languages. The SVM (50k vocab) led with an F1 of 0.994, followed by the SVM (5k vocab) at 0.985. RemBERT had the highest performance among the pre-trained multilingual models with an F1 of 0.945 on Kavalan, and mBERT had an F1 of 0.849. XLM-R had the lowest F1 of all models of 0.508, indicating that the model frequently failed to recognize Kavalan sentences as such, misclassifying them across multiple other Formosan languages, as visible in Figure 13.

The confusion matrices for the Formosan benchmark (Figures 9–14) reveal that Atayal (tay) and Amis (ami) were the most common sources of misclassification across models, with random forest in particular confusing a substantial number of Kavalan (ckv) sentences with Amis and Tay. SVM models had much less confusion overall, while mBERT and RemBERT showed scattered but moderate errors across several language pairs. XLM-R displayed severe confusion across the majority of Formosan language pairs, consistent with its low macro F1 performance.

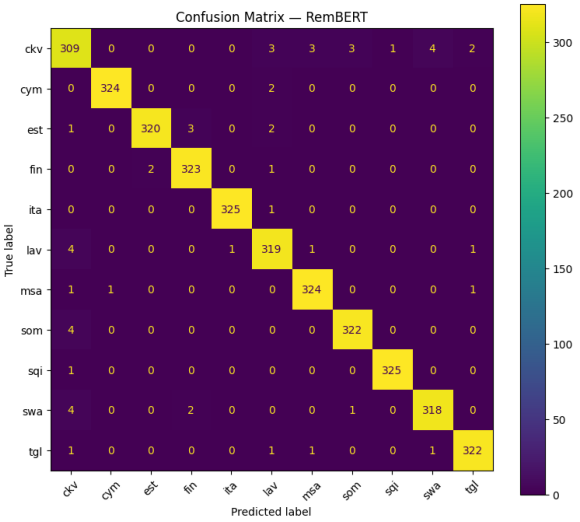

Figure 8: Confusion matrix of the classification results of remBERT on the langdetect-based benchmark

Model	Accuracy	Macro F1
Random Forest	0.856 ± 0.011	0.856 ± 0.011
SVM (5k vocab)	0.898 ± 0.007	0.899 ± 0.007
SVM (50k vocab)	0.906 ± 0.007	0.906 ± 0.007
mBERT	0.815 ± 0.017	0.813 ± 0.020
XLM-R	0.551 ± 0.053	0.505 ± 0.061
RemBERT	0.888 ± 0.014	0.892 ± 0.014

Table 5: Performance of all models on the Formosan language benchmark (mean ± SD across 15 folds)

Looking at the XLM-R confusion matrix (Figure 13), Kavalan sentences were frequently misclassified as Thao (ssf, n=94) and Sakizaya (szy, n=28), while only 4 sentences were misclassified as Amis (ami), its closest genealogical relative in the benchmark. This demonstrates that the misclassification pattern does not follow linguistic proximity, as Thao belongs to a distinct and distantly related branch. XLM-R appears to be transferring from superficially similar languages in its pre-training data, producing systematic misclassification concentrated in specific languages rather than random errors distributed across the benchmark.

4. Discussion

We construct two benchmarks based on a potential deployment setting and linguistic and typological motivations, and evaluate random forest, SVM models, and pretrained multilingual models on the benchmarks. We find that the language identification of Kavalan is possible with high accuracy, although performance varies depending on the model and the other languages in the benchmark. On the langdetect-based benchmark, all models performed strongly with SVM (50 vocab), achieving the best overall macro F1 of 0.993, and nearly perfectly identifying Kavalan individually. On the more challenging Formosan languages benchmark, SVM models remain the best-performing models, while the performance of pre-

Model	Precision	Recall	F1
Random Forest	0.930	0.942	0.936
SVM (5k vocab)	0.991	0.979	0.985
SVM (50k vocab)	0.997	0.991	0.994
mBERT	0.834	0.864	0.849
XLM-R	0.680	0.406	0.508
RemBERT	0.927	0.963	0.945

Table 6: Performance of all models on the identification of Kavalan specifically evaluated against typologically similar Formosan languages

Model A	Model B	*p*-value	Sig.
Random Forest	SVM (5k vocab)	<0.001	*
Random Forest	SVM (50k vocab)	<0.001	*
Random Forest	mBERT	<0.001	*
Random Forest	XLM-R	<0.001	*
Random Forest	RemBERT	<0.001	*
SVM (5k vocab)	SVM (50k vocab)	0.001	*
SVM (5k vocab)	mBERT	<0.001	*
SVM (5k vocab)	XLM-R	<0.001	*
SVM (5k vocab)	RemBERT	0.121	
SVM (50k vocab)	mBERT	<0.001	*
SVM (50k vocab)	XLM-R	<0.001	*
SVM (50k vocab)	RemBERT	0.007	*
mBERT	XLM-R	<0.001	*
mBERT	RemBERT	<0.001	*
XLM-R	RemBERT	<0.001	*

Table 7: Pairwise Wilcoxon signed-rank tests on macro F1 scores across 15 folds for the Formosan benchmark. * denotes $p < 0.05$.

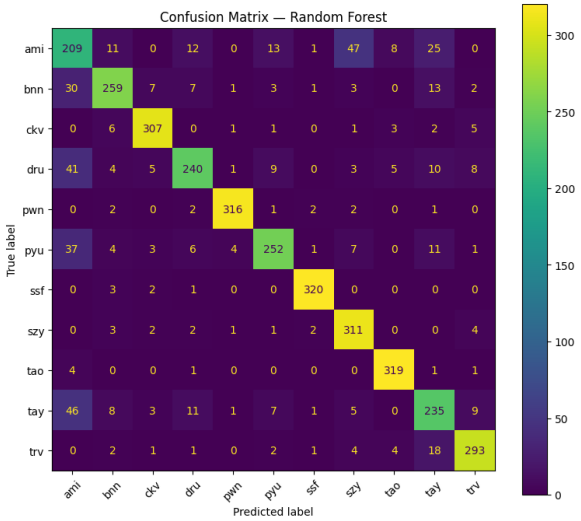

Figure 9: Confusion matrix of the classification results of random forest on the Formosan languages benchmark

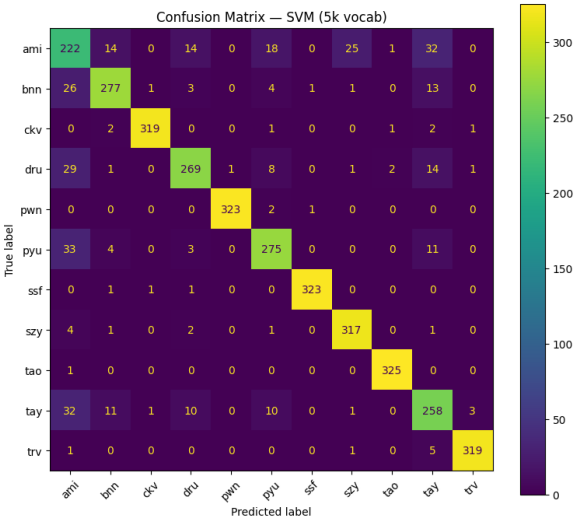

Figure 10: Confusion matrix of the classification results of SVM (5k vocab) on the Formosan languages benchmark

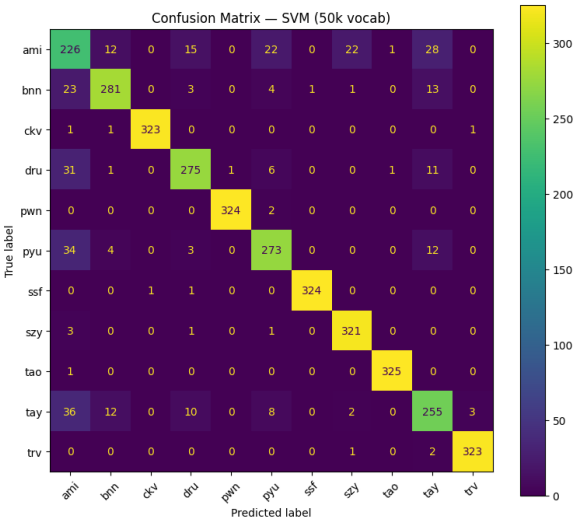

Figure 11: Confusion matrix of the classification results of SVM (50k vocab) on the Formosan languages benchmark

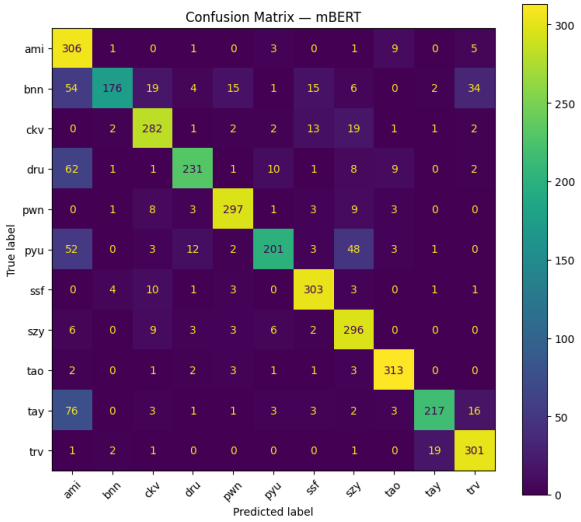

Figure 12: Confusion matrix of the classification results of mBERT on the Formosan languages benchmark

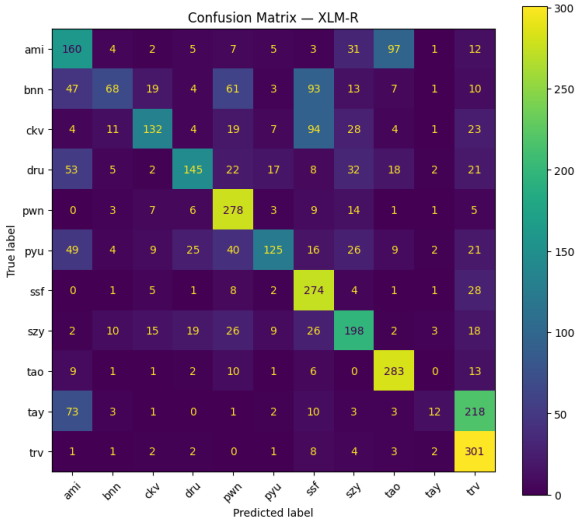

Figure 13: Confusion matrix of the classification results of XLM-R on the Formosan languages benchmark

trained multilingual models degraded significantly. XLM-R failed the Formosan languages benchmark with a macro F1 of 0.505. These results demonstrate that traditional *n*-gram-based SVM models are superior to large pre-trained multilingual models for the identification of extremely low-resource languages like Kavalan. This finding is significant as SVM-based approaches are computationally inexpensive and work with low amounts of data (e.g., 326 sentences), making this approach viable for low-resource language communities and research.

The strong performance of SVM models with character *n*-gram features is consistent with prior work on low-resource language identification. Bestgen (2017) and Sindane and Marivate (2024) both have found that the *n*-gram approaches performed well on low-resource languages. In our work, we find that SVM models perform significantly better than pre-trained multilingual models. The underperformance of mBERT and XLM-R is consistent with the known limitations of these models on languages absent from their pre-training data (Ebrahimi et al., 2022). As no Formosan languages are included in the training data of mBERT, XLM-R, or RemBERT, these models cannot draw on language-specific representations from typologically distant languages, which appears insufficient for language identification in the Formosan lan-

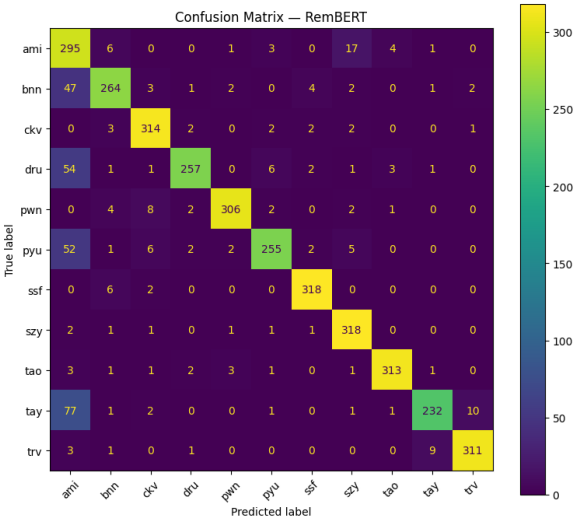

Figure 14: Confusion matrix of the classification results of RemBERT on the Formosan languages benchmark

guage family.

We find the XLM-R collapsed on the Formosan benchmark. While underperformance relative to SVM models was anticipated given the absence of Formosan languages from XLM-R's pre-training data, a macro F1 of 0.505 was unexpected, as XLM-R has performed well on unseen low-resource language tasks in the past (Winata et al., 2022). XLM-R's SentencePiece tokenizer, trained with no Formosan language representation, likely over-segments Formosan words in ways that obscure the morphological patterns, such as focus affixes and reduplication that character n-grams naturally capture. XLM-R may also resort to spurious transfer from superficially similar languages in its training data, such as Malay or Tagalog, producing systematic rather than random misclassification, consistent with the structured confusion visible in Figure 13. RemBERT's stronger performance may reflect its rebalanced training objective, which upweights lower-resource languages and may create more language-agnostic representations less susceptible to such transfer. RemBERT maintained a macro F1 of 0.892 on the same benchmark, suggesting that architectural and training differences between these models may significantly affect outputs in low-resource transfer scenarios. Random forest performed better than both mBERT and XLM-R in the linguistically motivated benchmark, suggesting that a simple ensemble of character n-gram features captures discriminative orthographic or morphological features that pre-trained multilingual models often cannot.

Future work should expand the benchmarks to cover more Formosan languages from more sources to provide a more comprehensive picture of language identification difficulty across the family. Given the complexity of Formosan languages, future works should also explore linguistically motivated features such as morpheme-level segmentation or language-specific tokenization, to improve identification performance beyond character *n*-grams. LID tools should also be deployed for language revitalization applications as LID is necessary of downstream tasks such as machine translation (MT) and automatic speech recognition (ASR) of Kavalan and related Formosan languages. The construction of larger and more diverse Kavalan corpora and corpora for other Formosan languages would be useful for improving model training and evaluating robustness of variation, such as code-switching to Mandarin, Taiwanese, and English. Evaluating these models on a comparably constructed benchmark for a well-resourced language such as Indonesian would further situate these results and clarify how much of the performance gap is attributable to low-resource conditions specifically.

This study provides the first systematic evaluation of machine learning models for the language identification of Kavalan, an extremely endangered Formosan language with under 300 known speakers. We construct two benchmarks reflecting a realistic deployment scenario and representing a linguistically motivated challenge and find that SVM models with character *n*-gram features consistently outperform pre-trained multilingual models, and achieve near-perfect identification of Kavalan in both settings. These findings demonstrate that effective LID tools for low-resource Formosan languages can be built with traditional and computationally accessible approaches, and build a foundation for the inclusion of Formosan languages in NLP research that could support their documentation and revitalization.

5. Conclusions

The study constructs the first benchmarks for Formosan language identification. We evaluate random forest, SVM models, mBERT, XLM-R, and RemBERT on our benchmark.

We find that SVM models perform the best in a realistic deployment setup with Kavalan, with a macro F1 of 0.993, and random forest and pre-trained multilingual models remain competitive with SVMs. In our linguistically motivated benchmark with solely Formosan languages, we find that SVM models continue to perform the best with a macro F1 of 0.906. Pre-trained multilingual models degrade in this setup, with the macro F1 of XLM-R falling to 0.505, indicating that some pre-trained multilingual models are unable to general-

ize to Formosan language tasks.

This work creates a foundation for robust language identification models for Formosan languages and displays the potential for SVMs in settings with extremely limited data. These findings contribute to advancing Kavalan and Formosan NLP research, create reliable models for LID, and advance the inclusion of low-resource and Formosan languages in NLP.

Limitations

There are several limitations that should be considered in this study. First, the Kavalan corpus used in this work is small, consisting of 326 sentences drawn from the ePark corpus. While this reflects the scarcity of Kavalan text data and the text data of other low-resource Formosan languages, it limits the diversity of linguistic contexts and variation represented in training and evaluation, and results may not fully generalize to naturalistic Kavalan language use. Second, the confusable languages in the langdetect-based benchmark were selected based on the output of an existing tool rather than linguistic analysis. While this reflects a potential deployment scenario, it means the benchmark is not the most difficult, as current tools lack the support of other typologically similar languages to Kavalan. Third, this work focuses on written text and does not address the identification of spoken Kavalan, which presents additional challenges due to the limited availability of resources for Kavalan.

6. Ethics

This research uses publicly available educational materials from the ePark corpus, developed by the Indigenous Languages Research and Development Foundation. We seek to actively contribute to the accessibility and visibility of Kavalan within computational linguistics. We acknowledge that Kavalan is critically endangered and that technological work on indigenous languages must support, not extract from, community-led revitalization efforts. Our publicly released benchmarks aim to enable NLP tools that could aid language documentation and preservation. Future applications should be developed in collaboration with Formosan language communities to ensure technological advances serve their self-determined preservation goals.

7. Bibliographical References

Yusuf Ayodeji Ajani, Bolaji David Oladokun, Shuaib Agboola Olarongbe, Margaret Nkechi Amaechi, Nafisa Rabiu, and Musediq Tunji Bashorun. 2024. Revitalizing indigenous knowledge systems via digital media technologies for sustainability of indigenous languages. *Preservation, Digital Technology & Culture*, 53(1):35–44.

Yves Bestgen. 2017. Improving the character N-gram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain. Association for Computational Linguistics.

Robert Blust. 2010. The Austronesian languages of Asia and Madagascar (review). *Oceanic Linguistics*, 49(1):302–312.

Robert Blust. 2013. *The Austronesian Languages*. Asia-Pacific Linguistics, Australian National University, Canberra. Revised edition.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Mohan G, R Prasanna Kumar, Elakkiya R, Venkatakrishnan R, Harrieni Shankar, Y Sree Harshitha, Harini K, and M Nikhil Reddy. 2023. Transformer-based models for language identification: A comparative study. In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–6. IEEE.

René Haas and Leon Derczynski. 2020. Discriminating between similar Nordic languages. *arXiv preprint arXiv:2012.06431*.

Fuhui Hsieh and Shuanfan Huang. 2007. Documenting and revitalizing Kavalan. In David Rau, editor, *Documenting and Revitalizing Austronesian Languages*, pages 94–104. National Foreign Language Resource Center and University of Hawa'i Press.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218. Association for Computational Linguistics.

C. Kruengkrai, P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara. 2005. Language identification based on string kernels. In *IEEE International Symposium on Communications and Information Technology (ISCIT 2005)*, volume 2, pages 926–929.

Kaiying Kevin Lin, Hsiyu Chen, and Haopeng Zhang. 2025. FormosanBench: Benchmarking low-resource Austronesian languages in the era of large language models. *arXiv preprint arXiv:2506.21563*.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through preprocessing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4869–4874, Florence, Italy. Association for Computational Linguistics.

Hunter Scheppat, Joshua Hartshorne, Dylan Leddy, Eric Le Ferrand, and Emily Prudhommeaux. 2025. Integrating diverse corpora for training an endangered language machine translation system. In *Proceedings of the Eighth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 162–169, Honolulu, Hawa'i, USA. Association for Computational Linguistics.

Yudi Setiawan, Nur Ulfa Maulidevi, and Kridanto Surendro. 2024. The optimization of N-gram feature extraction based on term occurrence for cyberbullying classification. *Data Science Journal*, 23.

Thapelo Sindane and Vukosi Marivate. 2024. From N-grams to pre-trained multilingual models for language identification. *arXiv preprint arXiv:2410.08728*.

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

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,

pages 777–791, Online. Association for Computational Linguistics.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025. Is it Navajo? accurate language detection in endangered Athabaskan languages. *arXiv preprint arXiv:2501.15773*.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for Amis-Mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84, Taipei, Taiwan. Association for Computational Linguistics.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for Formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259. Association for Computational Linguistics.

8. Language Resource References

Indigenous Languages Research and Development Foundation. 2025. 族語樂園. Indigenous Languages Research and Development Foundation. Digital Center of Taiwan Formosan Languages Production, University of Taipei. Accessed: 2025-08-30.