Morphological Feature Extraction for Fine-Grained Sorani Kurdish Dialect Identification: A Hybrid Transformer-Linguistic Approach

Soumedhik Bharati **Shibam Mandal** **Subham Majumdar** **Swarup Kr Ghosh** **Sayani Mondal**
Sister Nivedita University
{soumedhikbharati, shibammandal603, subhammajumdar.09123, swarupg1, sayani.mondal9}@gmail.com

Abstract

As reported, approximately 6 million people in Iraq and Iran speak in Sorani Kurdish, which exhibits substantial regional variation but lacks computational resources for dialect identification. We present the first fine-grained sub-dialect classification system for six Sorani varieties namely, Sulaymaniyah, Erbil, Iranian Sorani, Ardalani, Babani, and Mukriani. This investigation combines cross-lingual contextual embeddings (XLM-RoBERTa) with morphological features derived from explicit linguistic rules, including 24 patterns capturing verb prefixes, pronominal clitics, and definite markers. The suggested morphology-augmented XLM-R model has been trained on a unified dataset of $16,409$ sentences without manual annotation, and achieves 91.91% accuracy, outperforming pure transformers (91.79%) and traditional machine learning baselines (SVM 86.41%). Key ablation studies reveal that morphological features serve as effective regularizers for geographically proximate dialects.

1 Introduction

Sorani Kurdish (Central Kurdish) is a low-resource language written in the Persian-Arabic script (Abdalla et al., 2025). It features complex morphology and significant regional variation. While linguistic studies have documented these variations, computational approaches to Kurdish dialect identification remain sparse. Prior work has largely been limited to binary classification between wide Iranian and Iraqi varieties, achieving high accuracy (96%) but failing to distinguish finer sub-dialects (Malmasi, 2016). To the best of our knowledge, no computational work exists on identifying specific Sorani sub-dialects. This codebase and the dataset are publicly available.[1].

1.1 Contribution

We propose a fine-grained, six-way classification task covering the major Sorani varieties, including *Sulaymaniyah, Erbil, Iranian Sorani (Sanandaji), Ardalani, Babani, and Mukriani*. We have investigated two primary research questions: (1) Can transformer-based models effectively distinguish closely related sub-dialects? (2) Does explicit morphological feature extraction provide a complementary signal to contextual embeddings in low-resource settings? Our primary contributions are threefold. First, we introduce the first 6-way sub-dialect identification dataset for Sorani, comprising 16,409 sentences derived from metadata-rich corpora. Second, we propose a hybrid architecture that combines XLM-RoBERTa embeddings with a vector of 24 linguistically motivated morphological features such as verb prefixes, clitics, and markers. Finally, we exhibit that the morphology-augmented model achieves 91.91% accuracy, giving the best performance among the models we evaluated for fine-grained Sorani dialect identification.

2 Related Work

Recent efforts have improved resources for Kurdish. This includes morphological analysers and named entity recognition (Ahmadi, 2020a; Naserzade et al., 2023). However, most existing tools treat Sorani as a monolith, ignoring the dialectal nuances important for downstream tasks like automatic speech recognition (ASR) or culturally aware machine translation (Ahmadi, 2020b). Dialect identification is a well-established task for languages like Arabic, where distinguishing between Egyptian, Levantine, and Gulf dialects is common (Althobaiti, 2020; Alyami and Alzaidy, 2020). For Kurdish, character n-grams were used to distinguish Kurmanji from Sorani, as well as Iranian Sorani from Iraqi Sorani Malmasi (2016). The six granular sub-dialects within Sorani were missing from previous work.

[1] https://github.com/Soumedhik/sorani-kurdish-dialect-identification

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 172–176
March 28, 2026. ©2026 Association for Computational Linguistics

Table 1: Dataset composition across six Sorani sub-dialects. The split is approximately 70/15/15.

Dialect	Train	Val	Test	Total
Sulaymaniyah	1,968	416	393	2,777
Erbil	1,908	407	400	2,715
Iranian Sorani	1,931	411	400	2,742
Ardalani	1,936	412	404	2,752
Babani	1,924	413	400	2,737
Mukriani	1,935	410	402	2,747
Total	**11,602**	**2,408**	**2,399**	**16,409**

Table 2: Subset of the 24 morphological patterns used for feature extraction.

Category	Feature	Example Pattern
Verb Prefix	`de_present` `na_negative`	`de-` (present habitual) `na-` (negation)
Clitics	`clitic_3sg` `clitic_1pl`	`-y` (Sulaymaniyah) vs `-ê` (Iran) `-ayn`
Definiteness	`def_eke` `def_aka`	`-eke` (Iranian Sorani) `-aka` (Iraqi Sorani)
Plural	`pl_an`	`-an`
Lexical	`zor_very`	`zor` vs `gele`

Combining neural representations with hand-crafted features has proven effective for low-resource languages (Abdulmumin et al., 2021). While deep learning models like XLM-R excel at capturing semantic context, they often require large datasets to learn morphological rules that can be explicitly encoded via regex patterns (Baziotis et al., 2022). We adopt this hybrid strategy to utilise linguistic knowledge where data volume is limited.

3 Datasets and Linguistic Features

The datasets used in this research and the linguistic features extraction methods are discussed.

3.1 Dataset Construction

We have constructed a dataset of $16,409$ sentences collected from Kurdish news websites and digital media sources representing six Sorani sub-dialects, using an automated collection and normalization pipeline, followed by data augmentation to balance dialect classes. We treat Sulaymaniyah and Babani as distinct labels following established sociolinguistic usage in regional media, while acknowledging their close linguistic relationship. The training pipeline operates without manual annotation and produces stratified train/validation/test splits depicted in Table 1.[2]

Preprocessing included Unicode normalization (unifying Farsi/Arabic characters like *yeh* and *kaf*) and diacritic handling.

3.2 Morphological Feature Extraction

We have developed a custom extractor targeting 24 documented morphological discriminators (Naserzade et al., 2023; Ahmadi, 2021). These features capture systematic variation across Sorani sub-dialects at multiple linguistic levels. Specifically, we model variation in verb prefixes that mark tense and aspect (e.g., the alternation between *de-* and *a-* for present habitual constructions), differences in pronominal clitics used for person marking (e.g., 3rd singular *-y* versus *-ê*), and phonological variation in definite suffixes (e.g., *-aka* in Erbil versus *-eke* in Iranian Sorani). Together, these patterns encode dialect-specific morpho-syntactic cues that are difficult for purely distributional models to learn reliably in low-resource settings.

To illustrate the dialectal variation captured by the dataset, Table 2 is complemented by naturally occurring corpus examples. For example, the present habitual appears as *de-nûsim* (“I write”) in Sulaymaniyah, while Iranian Sorani frequently uses *a-nûsim*. Similarly, definite nouns occur as *kitêb-aka* in Iraqi Sorani and *kitêb-eke* in Iranian Sorani. Such region-consistent patterns motivate the inclusion of explicit morphological indicators.

4 Methodology

The details of model architectures and background are presented in this section. The overall processing pipeline of the proposed morphology-augmented model is summarized in Algorithm 1.

4.1 XLM-RoBERTa

We have fine-tuned **XLM-RoBERTa-base** (Conneau et al., 2020), a multilingual transformer pre-trained on 100 languages with limited Kurdish coverage. Although Central Kurdish (Sorani, ckb) is not explicitly represented in the pre-training corpus, prior work has shown that XLM-R transfers effectively to unseen languages via shared script and subword representations. The model uses a SentencePiece tokenizer with a max length of 128 tokens. We fine-tuned the model for 25 epochs with

[2] Sources include region-specific Kurdish news portals from Sulaymaniyah, Erbil, Sanandaj, and Mahabad, selected based on self-declared regional coverage and author metadata. A full list of URLs is released with the dataset.

Algorithm 1 Morphology-Augmented Classification

```
Input: Sentence $S$, MorphPatterns $P$
Output: Dialect Label $\hat{y}$
$T \leftarrow \text{Tokenize}(S)$
$H \leftarrow \text{XLM-R}(T)$
$h_{\text{cls}} \leftarrow H_{[\text{CLS}]}$
$v_{\text{morph}} \leftarrow \text{ExtractFeatures}(S, P)$    $\triangleright \in \mathbb{R}^{24}$
$h_{\text{morph}} \leftarrow \text{MLP}(v_{\text{morph}})$
$h_{\text{morph}} \leftarrow \text{Proj}(h_{\text{morph}})$
$h_{\text{attn}} \leftarrow \text{MultiHeadAttn}(Q = h_{\text{morph}}, K = H, V = H)$
$z \leftarrow \text{Concat}(h_{\text{cls}}, h_{\text{attn}} + h_{\text{morph}})$
$\hat{y} \leftarrow \text{Softmax}(\text{Classifier}(z))$
```

a batch size of 16, using a learning rate of 2×10^{-5} for the encoder and 1×10^{-3} for the classification head. To prevent overfitting on this small dataset, we evaluated both frozen and unfrozen encoder strategies.

4.2 Morphology-Augmented XLM-R

The proposed architecture employs a **deep fusion with cross-attention** strategy. Let x denote the input sentence and $m \in \mathbb{R}^{24}$ represent the extracted morphological feature vector. Let $H = \text{XLM-R}(x)$ denote the sequence of contextual embeddings produced by the encoder, and $h_{\text{cls}} = H_{[\text{CLS}]} \in \mathbb{R}^{768}$ the corresponding CLS representation.

The morphological vector is passed through a deep projection network with residual connections, yielding $h_{\text{morph}} = \mathcal{F}_{\text{MLP}}(m) \in \mathbb{R}^{256}$ after a learned projection for dimensional alignment, where $\mathcal{F}_{\text{MLP}} : \mathbb{R}^{24} \to \mathbb{R}^{256}$ denotes a multi-layer perceptron with layer normalization and GELU activation.

To integrate the two modalities, we apply a multi-head cross-attention mechanism with $N = 4$ heads, selected empirically to balance representational capacity and overfitting on the 16k-sentence dataset. The morphological representation acts as the query, while the sequence of contextual embeddings provides the key and value, producing $h_{\text{attn}} = \text{MHA}(h_{\text{morph}}, H, H)$ following the standard multi-head attention formulation (Vaswani et al., 2023).

The attention output is combined with the original morphological representation via a residual connection, and concatenated with the CLS embedding to form the final representation $z = [\, h_{\text{cls}}\, ;\, (h_{\text{attn}} + h_{\text{morph}})\,] \in \mathbb{R}^{1024}$. This vector is passed to a classification head $\mathcal{C} : \mathbb{R}^{1024} \to \mathbb{R}^{6}$ with hidden dimension $d_h = 384$, followed by layer normalization, GELU activation, and softmax to obtain the dialect prediction $\hat{y} = \text{softmax}(\mathcal{C}(z))$.

Table 3: Overall performance comparison. The morphology-augmented model achieves the best results.

Model	**Accuracy**	**Macro F1**
Random Baseline	16.70%	0.1670
Logistic Regression	75.11%	0.7511
Linear SVM	86.41%	0.8639
XLM-R (Frozen)	91.79%	0.9179
Morph-Augmented XLM-R	**91.91%**	**0.9190**

5 Results and Analysis

The results produced by the suggested model, along with a comparison with the state-of-the-art, have been discussed in this section.

5.1 Baseline methods

We have trained Logistic Regression (LR) (Cox, 1958) and Linear SVM (Cortes and Vapnik, 1995) models using a combination of character n-grams (2-5 chars, 4,000 features), word unigrams (500 features), and 24 morphological counts. We have used TF-IDF weighting for text features and standard scaling for morphological counts.

5.2 Model Comparison

As shown in Table 3, transformer-based approaches significantly outperform traditional baselines. The linear SVM achieved 86.41% accuracy, confirming that character n-grams capture substantial dialectal signal. However, XLM-R improved this by over 5 percentage points, reaching 91.79%.

The proposed **morphology-augmented XLM-R** achieves the best overall performance, with 91.91% accuracy and 91.90% macro F1. The gain over vanilla XLM-R is modest but consistent (+0.12% absolute). McNemar's test shows that XLM-R significantly outperforms logistic regression ($p < 0.001$), and the morphology-augmented model significantly outperforms the SVM baseline ($p < 0.001$). The difference between the two XLM-R variants is not statistically significant ($p = 0.6625$), indicating that morphological features provide complementary but subtle benefits.

5.3 Per-Dialect Analysis

Table 4 details the performance by dialect. The model performs consistently well across all classes

Figure 1: Model comparison across six Sorani dialects

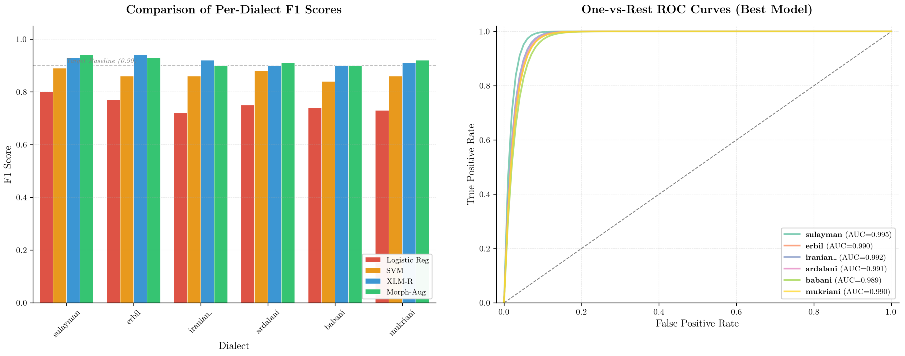

Figure 2: XLM-R performance analysis.

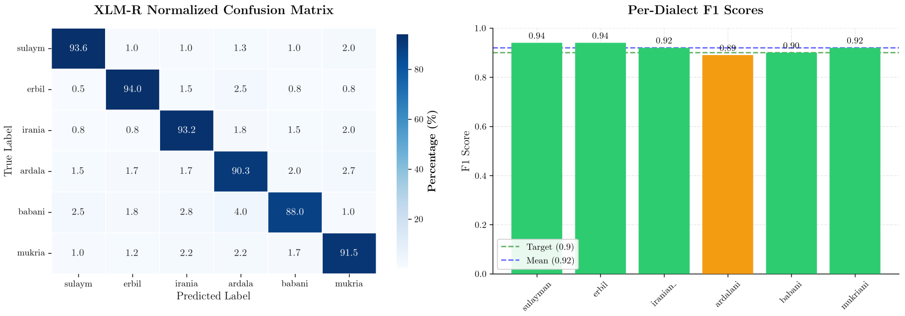

Table 4: Per-class performance of the best model.

Dialect	Precision	Recall	F1-Score
Sulaymaniyah	0.9437	0.9389	0.9413
Erbil	0.9282	0.9375	0.9328
Iranian Sorani	0.9098	0.9325	0.9210
Ardalani	0.8831	0.9158	0.8991
Babani	0.9305	0.8700	0.8992
Mukriani	0.9227	0.9204	0.9215

Table 5: Top confusion pairs in error analysis.

True	Predicted	Error %
Babani	Ardalani	8.8%
Babani	Iranian Sorani	7.2%
Erbil	Ardalani	5.7%
Mukriani	Ardalani	4.6%

(F1 > 0.89). Sulaymaniyah, the standard educational dialect, achieves the highest F1 (0.94). Babani shows a slightly lower F1-score (0.899), likely due to overlap with other Southern Sorani varieties.

5.4 Error Analysis

We analyze the confusion matrix in Table 5 to examine linguistic similarity across dialects. The most frequent errors occur between *Babani* and *Ardalani* (8.8%), which is linguistically expected given their geographic proximity in Western Iran and shared clitic patterns. Confusion between *Erbil* and *Ardalani* further reflects overlapping phonological features across the Iraq–Iran border.

5.5 Discussion

Multilingual transformers struggle with fine-grained dialectal distinctions in low-resource settings. Incorporating a lightweight morphological feature vector (24 dimensions) improves stability for minority dialects such as Babani, yielding a 0.3% F1 gain in ablation studies. This reflects the complementary role of explicit morphological constraints over purely statistical representations (Abdulmumin et al., 2021).

6 Conclusion

We present a computational study of Sorani Kurdish sub-dialects, achieving the highest accuracy with a morphology-augmented XLM-RoBERTa model. By integrating contextual embeddings with a small set of linguistically motivated morphological features, the approach effectively distinguishes closely related Sorani varieties that challenge purely data-driven models. This highlights the value of hybrid neural–linguistic methods for fine-grained dialect identification in under-resourced languages.

7 Limitations

Our study relies primarily on text collected from regional news sources. While this ensures grammatical standardization, it may not fully capture the colloquial morphological variations found in spoken dialects or social media text. Additionally, while XLM-RoBERTa is pre-trained on 100 languages, Central Kurdish (Sorani) is not a primary training language compared to Northern Kurdish (Kurmanji). We rely on the model's ability to transfer representations via the shared Perso-Arabic script and Persian lexical overlap. Finally, we acknowledge that the distinction between varieties such as Sulaymaniyah and Babani is often fluid; our classification relies on geographic source metadata which may contain inherent overlaps.

8 Ethical Considerations

This work studies dialect identification for Sorani Kurdish using publicly available text from news and digital media sources. No new data were collected, and no human subjects were involved. The dataset does not contain personally identifiable or sensitive information to the best of our knowledge.

Dialect identification carries potential risks if misused, including stereotyping or unintended profiling of speakers. Our goal is purely analytical: to improve linguistic modeling for an under-resourced language and support downstream NLP research. We do not advocate deploying such systems in high-stakes or user-facing settings without careful validation, transparency, and community oversight.

Finally, we emphasize that dialect boundaries in Sorani Kurdish are fluid and socially constructed. Model predictions should therefore be interpreted as probabilistic signals rather than definitive labels, and used with appropriate linguistic and cultural awareness.

References

Bakhtawar Abdalla, Rebwar Mala Nabi, Hassan Eshkiki, and Fabio Caraffini. 2025. Named entity recognition for the kurdish sorani language: Dataset creation and comparative analysis. *Preprint*, arXiv:2511.22315.

Idris Abdulmumin, Bashir Shehu Galadanci, Abubakar Isa, Habeebah Adamu Kakudi, and Ismaila Idris Sinan. 2021. A hybrid approach for improved low resource neural machine translation using monolingual data. *Preprint*, arXiv:2011.07403.

Sina Ahmadi. 2020a. KLPT – Kurdish language processing toolkit. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 72–84, Online. Association for Computational Linguistics.

Sina Ahmadi. 2020b. KurdishBLARK: Basic language resources and tools for Kurdish. https://kurdishblark.github.io/. Accessed: 2025-12-31.

Sina Ahmadi. 2021. Hunspell for sorani kurdish spell checking and morphological analysis. *Preprint*, arXiv:2109.06374.

Maha Jarallah Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey. *Preprint*, arXiv:2009.12622.

Reem Alyami and Rabeah Alzaidy. 2020. Arabic dialect identification in social media. In *ICCAIS 2020 - 3rd International Conference on Computer Applications and Information Security*, ICCAIS 2020 - 3rd International Conference on Computer Applications and Information Security, United States. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2020 IEEE.

Christos Baziotis, Mikel Artetxe, James Cross, and Shruti Bhosale. 2022. Multilingual machine translation with hyper-adapters. *Preprint*, arXiv:2205.10835.

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.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

David R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–242.

Shervin Malmasi. 2016. Subdialectal differences in Sorani Kurdish. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 89–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Morteza Naserzade, Aso Mahmudi, Hadi Veisi, Hawre Hosseini, and Mohammad MohammadAmini. 2023. Ckmorph: A comprehensive morphological analyzer for central kurdish. *International Journal of Digital Humanities*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.