NLP final project - cross Arabic dialectal pos tagging comparison

Nir Shoham (ID. 322657073), Omer Kuriel (ID. 206998635)
Submitted as final project report for the NLP course, Reichman University, 2025

1 Introduction:

Arabic is a linguistically rich language with significant dialectal variation across regions. While Modern Standard Arabic (MSA) is widely used in formal contexts, most native speakers communicate in dialects that differ lexically, morphologically, and syntactically from MSA. These dialects pose a challenge for natural language processing (NLP), particularly due to the scarcity of annotated data and the limited generalizability of models trained primarily on MSA.
Our motivation to explore this topic stems from both linguistic and personal interests. As a former Arabic translator, I (Omer) have developed a deep appreciation for the language's complexity and internal diversity. This project provided an opportunity to further investigate the connections and divergences between Arabic dialects. We chose to focus on part-of-speech (POS) tagging, a core task in NLP that forms the foundation to many downstream applications. While a substantial amount of work has been done on dialect identification, we found far fewer studies examining how existing POS tagging models perform across dialects - especially in a comparative framework. This gap in the literature motivated our decision to evaluate model performance in a controlled, dialect-diverse setting.
Through this work, we aim to contribute both linguistic insights and empirical benchmarks for evaluating the capacity of modern NLP tools to handle dialectal Arabic.

1.1 Related Work:

Recent years have seen a surge in the development of Arabic pre-trained language models, with efforts such as AraBERT, MARBERT, ARBERT, and the CAMeLBERT family achieving state-of-the-art results on various Arabic NLP tasks. Among these, CAMeLBERT (Inoue et al., 2021) offers an especially valuable contribution by systematically examining how language variant (Modern Standard Arabic, Dialectal Arabic, and Classical Arabic), pre-training data size, and task type affect model performance across five tasks, including part-of-speech (POS) tagging.
A key insight from their work is that the proximity between the language variant used during pre-training and the variant used during fine-tuning is more important than the size of the pre-training corpus. This conclusion directly motivated the design of our project, which seeks to explore this variant-proximity effect in greater depth, particularly within and across Arabic dialects.

In contrast to prior work that typically trains and evaluates models on dialects in isolation, we ask: how does fine-tuning on one dialect affect performance on another? What kinds of linguistic features (e.g. verbs with prefixes, rare vocabulary) pose consistent challenges across dialects?
Our project aspires to offer a complementary micro-level analysis of dialectal POS tagging under data-scarce conditions. We provide new annotated resources, comparative evaluations, and a linguistically motivated analysis of how dialectal proximity shapes model behavior.

2 Methodology:

Our project is divided into two main phases: (1) zero-shot evaluation using a pre-trained Modern Standard Arabic (MSA) POS tagging model, and (2) fine-tuning the model on our manually annotated dialectal POS tagging dataset to assess cross-dialect generalization.

We began by constructing a gold-standard dataset of 700 sentences, manually POS-tagged using a unified tagset. These sentences were sampled from seven subsets of the MADAR corpus: five dialects from MADAR Corpus-6 (Doha, Beirut, Cairo, Tunis, Rabat), MSA, and the Jerusalem dialect from MADAR Corpus-26. Each subset contains 100 sentences, annotated by Omer - a former Arabic translator, with close attention to dialect-specific syntax, and morphological variation.

For Phase 1, we used the CAMeL-Lab/bert-base-arabic-camelbert-ca-pos-msa model from Hugging Face - a CAMeLBERT-CA model fine-tuned on the Penn Arabic Treebank (PATB) for POS tagging in MSA. While this model is not trained on dialectal Arabic, it served as a strong baseline for zero-shot evaluation. We applied the model to all dialect subsets without additional training and assessed its ability to generalize to unseen dialectal data. Predictions were aligned with our tagset to ensure consistency across subsets.

In Phase 2, we fine-tuned the CAMeLBERT-CA POS-MSA model on each individual dialect dataset and evaluated its performance across the remaining dialects. This setup allowed us to study cross-dialect relations and to examine whether fine-tuning on a specific variety improves performance on others - especially closely related ones. We also evaluated in-dialect fine-tuning performance. In both phases, we generated accuracy scores and confusion matrices for each dialect pair.

All modeling tasks were implemented using Hugging Face's transformers library. We trained and evaluated models in Google Colab, using GPU acceleration. Each fine-tuning session (on one dialect) took approximately 20 seconds, with 3 epochs, a learning rate of 5e-5, and a batch size of 16.

We used 3 epochs and a learning rate of 5e-5 because BERT-style token classification usually learns quickly, and three passes balance learning with overfitting risk. A very small batch size (~10%) gives more frequent, noisier updates that can help generalization on low-resource, dialect-specific data. Using Hugging Face Transformers keeps the process simple and reproducible.

Our codebase is written in Python and structured into modular components for data loading, preprocessing, model training, and evaluation. Throughout the project, we encountered and addressed several technical challenges:

- **Dialectal orthographic variation**: Unlike MSA, which follows standardized spelling conventions, dialectal Arabic is primarily a spoken language and lacks orthographic standardization. This results in significant variability in how words are written - for instance, the word for “I want” in Palestinian Arabic may appear as بدي, بدّي, or بديي. In addition, sounds may be represented inconsistently (e.g., ق rendered as ء or omitted entirely), and word endings may alternate between ة and ه. Such variation introduces orthographic noise that degrades model performance, especially for models trained on standardized MSA.
- **Unknown tokens**: When the model encountered a token it did not recognize, it applied subword tokenization to break it down into smaller, known components. This often resulted in a single word being split into multiple sub-tokens, each receiving its own POS tag. For example, the word مائدة ("table") was tokenized as: "ما" (prefix), "##ئ" (middle part), "##دة" (suffix). Although this is a single word in the gold dataset, the model treated it as three separate units. This posed challenges for evaluation, as our methodology is based on a one-to-one alignment between

gold-standard tokens and predicted tags. To resolve this, we adopted a consistent strategy: we retained only the POS tag assigned to the **first sub-token** and used it to represent the entire word for evaluation purposes.

Additionally, in the raw tokenized output before fine-tuning, sub-tokens beyond the first for each word often received the -100 label in the training data. This label is a special ignore index used by Hugging Face's transformers library to ensure that these positions do not contribute to the loss calculation during training. Thus, only the first sub-token's prediction is considered in both loss computation and evaluation.

- **Prepositional constructions:** One significant limitation of the model is its inability to assign composite part-of-speech (POS) tags such as prep+noun or verb+pron, which were initially commonly used in our gold-standard annotations to capture the fused morphological and syntactic structure of dialectal Arabic. As a result, the model frequently struggled with accurately tagging fused prepositional forms such as بجانب ("beside"), which consist of a preposition combined with another element (e.g., noun, pronoun, or adverbial component). Rather than identifying these as compound units, the model typically segmented them into separate parts - for example, "بـ" (prepositional prefix) and "جانب" (side) - but assigned a single POS tag to the full token, often defaulting to the tag of the main component (e.g., NOUN, PRON, etc.).

  To ensure alignment between the model's output and our gold-standard annotations, we adopted a pragmatic evaluation strategy that followed the model's segmentation logic. Specifically, in cases where the fused prepositional form was followed by a pronoun (e.g., "عندك" 'you have'), we assigned the prepositional POS tag (prep+pron) to the entire token to reflect its syntactic role and morphological cohesion. In all other cases (e.g., "بجانب"), we used the POS tag of the post-prepositional element (e.g., noun) as the label for the entire token. This approach allowed us to maintain consistent evaluation across structurally mismatched outputs while preserving linguistic fidelity.

Despite these challenges, our methodology allowed for a controlled, comparative investigation of dialectal POS tagging, combining linguistic sensitivity with practical evaluation of model robustness in low-resource scenarios.

3 Experimental results

Metrics:

Cross dialect - Conservative F1 (Wilson Lower Bound–Adjusted Precision & Recall): We first wanted to rank the "best-performing" POS tag across all dialects. Our initial evaluation of POS tagging performance relied solely on recall (TP / (TP + FN)). After evaluating the results obtained, this approach proved problematic, since rare tags with very low frequency (e.g., occurring only once in the gold data) achieved perfect recall from a single correct prediction and appeared as top performers despite poor overall behavior. In addition, the recall metric does not take into consideration FP's, which leads to inflated impressions of performance. To address this, we implemented a more statistically conservative method. For each tag, we first calculated support for recall and for precision which represent the number of gold and predicted instances respectively. These values serve as the sample sizes for a Wilson lower bound calculation, applied separately to the recall $\hat{R} = \frac{TP}{TP+FN}$ and precision $\hat{P} = \frac{TP}{TP+FP}$ The Wilson lower bound for a proportion $\hat{p}$ with sample size n (95% confidence, z=1.96) is given by:

$$LB(\hat{P}, n) = \frac{\hat{P} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{P}(1-\hat{P})}{n} + \frac{z^2}{4n^2}}}{1+\frac{z^2}{n}}$$

Using support as n in this formula produces conservative estimates that shrink scores for small-sample tags and penalize tags with many false positives. Conservative F1 is then computed from these adjusted values. The formula is given by:

$$F1_{LB} = \frac{2 \cdot PRECISION_{LB} \cdot RECALL_{LB}}{PRECISION_{LB} + RECALL_{LB}}$$

This metric integrates the Wilson-adjusted precision and recall into a single score, ensuring that both small sample sizes and high false positive rates are penalized.
Additionally, we excluded the punctuation (punc) tag from performance summaries, as punctuation tagging is trivially solved and would skew aggregate scores.

Per dialect - Macro F1 and Micro F1 scores: Macro-F1 computes the F1 score independently for each class and averages them, giving equal weight to all tags regardless of frequency. Micro-F1 aggregates all true positives, false positives, and false negatives across classes before computing the F1 score, giving greater weight to high-frequency classes.

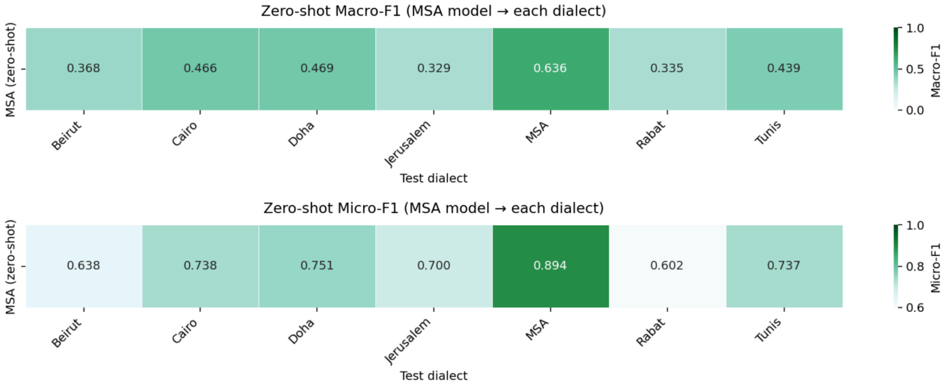
Cross dialect - Most Common Mismatch Pairs: For each dialect and in the zero-shot setting, we identify the most frequent confusion pairs (gold vs. predicted tags). This reveals systematic model biases, e.g., over-predicting nouns for functional tokens, misclassifying interrogative parts, collapsing verb forms into more common categories, and highlights where targeted fine-tuning could yield the largest gains.

Zero shot - Metrics results:

Cross dialect - Conservative F1: The conjunction (conj) tag proved to be the most reliable performer across dialects, achieving consistently high F1 scores with minimal variation. This can be attributed to its relatively unambiguous orthographic representation and limited morphological variability across Arabic varieties, making it easier for an MSA-trained model to recognize and classify.
By contrast, noun (noun) appears frequently and achieves strong recall, but its precision suffers due to systematic over-prediction. The model often defaults to tagging unknown or morphologically complex tokens as nouns, inflating false positives. While this behavior boosts recall, it leads to inflated noun counts at the expense of more specific categories (e.g., verbs, proper nouns, pronouns). This “noun inflation” effect is one of the major sources of nominal over-tagging identified in our error analysis.

Macro F1 and Micro F1 scores: As shown in Figure 1, MSA outperformed all other dialects with (Macro-F1 = 0.636, Micro-F1 = 0.894). Across dialects, macro-F1 scores remained substantially lower (≈0.33–0.47) than micro-F1 scores (≈0.60–0.75), indicating a long-tail effect: the model performs strongly on frequent classes (e.g., noun, punctuation) but underperforms on low-frequency or morphologically diverse categories (e.g., interrogatives, particles). This macro–micro gap confirms that the model’s relatively high aggregate performance is driven by over-reliance on high-frequency, morphologically stable tags, while rare or highly dialectal forms suffer from poor generalization.

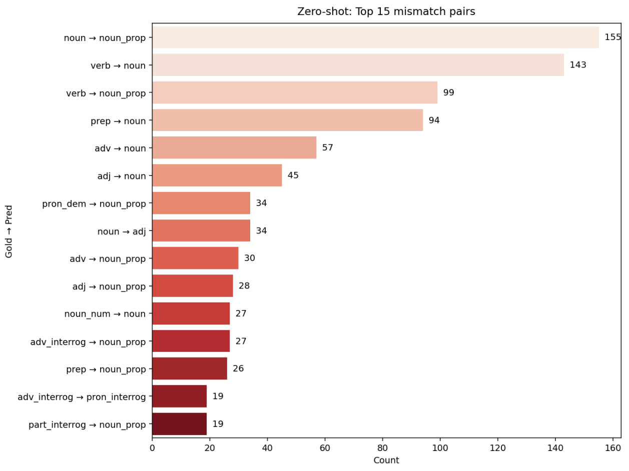
Figure 1: *zero-shot Macro-F1 (top) and Micro-F1 (bottom) for the MSA-trained model tested on each dialect.* MSA achieves the highest score for both metrics. The consistent gap between macro- and micro-F1 reflects the model's stronger performance on high-frequency tags and weaker performance on rare or morphologically diverse tags.

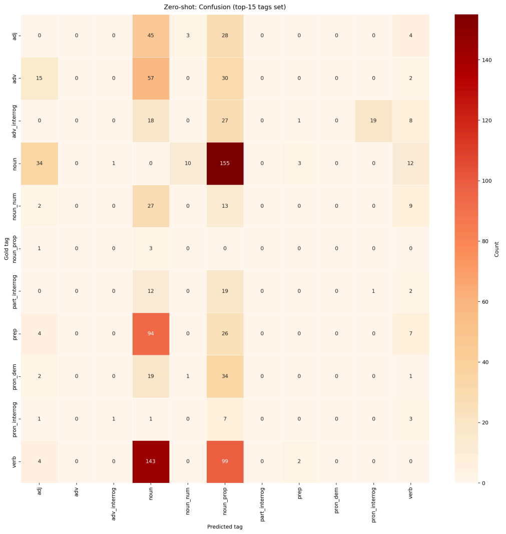


Cross dialect - Most Common Mismatch Pairs: As shown in Figure 2 and Figure 3, The model tends to overuse noun and noun_prop for unknown tokens (proper-noun and noun inflation): The model frequently backs off to noun_prop or noun when it can't lexically anchor a token, especially with dialect items (e.g., "بدي"), clitic fusion ("نضمنلك", "زيدني"), or informal spellings ("هدا" vs "هذا"). The most frequent zero-shot error is noun → noun_prop (155) which indicates that whenever the model doesn't recognize a certain noun it backs off to tag it as a specific noun (proper noun). The second frequent zero shot error is verb → noun. Further linguistic investigation revealed that whenever there is a verb that is "wrapped" with a prefix (modal prefix) and or with a pronoun suffix, the model has difficulty identifying the core POS tag in this token, and backoffs to a noun tag. This bias "up-labels" uncertain tokens as proper names or nouns, yielding syntactically implausible sequences and heavy precision penalties. This phenomena will be further discussed in the next page under "Failure to Handle Modal Verb Forms with Prefixes" part.

Figure 2. *Zero-shot: Top 15 most frequent gold–predicted mismatch pairs across dialects.* The highest mismatch counts occur in noun-to-proper-noun and verb-to-noun confusions, reflecting systematic over-prediction of nominal categories.

Figure 3. *Zero-shot: Confusion matrix for the top 15 POS tags (gold vs. predicted).* The darkest cells highlight the same dominant errors seen in Figure 2, confirming consistent model bias toward nominal over-tagging, particularly with dialect-specific or morphologically complex tokens.





Zero shot - Linguistic Error analysis:

Applying the **MSA-trained CAMeLBERT-CA-POS-MSA** model to dialectal Arabic in a zero-shot setting revealed several recurring, linguistically grounded error patterns. These patterns highlight the model's limited capacity to generalize beyond the conventions of standardized Arabic and reflect the inherent challenges posed by morphosyntactic variation across dialects. The following sections outline the most prominent phenomena, beyond the previously discussed tendency to overpredict nouns and proper nouns, accompanied by examples:

1. Failure to Handle Modal Verb Forms with Prefixes: The model exhibits consistent difficulty in modal verbs that begin with the prefix "بـ", which marks indicative/future tense in many Levantine, Egyptian, and Gulf dialects (e.g., "بتحكيلو" 'you tell him'). While native speakers readily recognize these as verbal forms, the model often misclassifies them as proper nouns or incorrectly segments them, reflecting a gap in its internalized representation of dialect-specific verbal morphology.

- **Example:**
 Input: **"بتكلف خمسة و تمانين سنت"** (*bitkalif khamse w-tamānīn sent*, 'It costs eighty-five cents.')
 Gold: [verb, noun_num, conj, noun_num, noun]
 Model: [noun, noun_num, conj, noun, verb]

Such errors suggest that the model treats the **"بـ"** prefix as orthographically independent, failing to map it onto its morphosyntactic role as a tense/aspect marker, and thereby misaligning the token's grammatical category.

2. Inability to Generalize Across Pronunciation Variants: A recurrent source of error stems from the model's inability to generalize across dialectal pronunciation and orthographic variants of Modern Standard Arabic (MSA) words. Many dialectal forms differ from MSA by only a few phonemes or letters. differences that are semantically transparent to human speakers but absent from the model's training corpus. For instance, the Levantine form **"هدا"** (*hāda*, 'this') replaces the MSA **"هذا"** (*hādhā*), while **"ما أحلاه"** (*mā aḥlāh*, 'how nice it is!') may be pronounced and written as **"محلاه"** (*maḥlāh*) in certain dialects.

- **Example 1:**
 Input: **"محلاه مريولك"** (*maḥlāh maryūlak*, 'How nice your shirt is!')
 Gold: [interj, noun]
 Model: [noun, noun_prop]
- **Example 2:**
 The MSA **"ثمانين"** (*thamānīn*, 'eighty') is often pronounced and written as **"تمانين"** (*tamānīn*) in Levantine and Egyptian dialects.
 Input: **"بتكلف خمسة و تمانين سنت"** (*bitkalif khamse w-tamānīn sent*, 'It costs eighty-five cents.')
 Gold: [verb, noun_num, conj, noun_num, noun]
 Model: [noun, noun_num, conj, noun, verb]

Such orthographic and phonological shifts pose a challenge for the model, which tends to treat the variant form as an unknown token and default to high-frequency nominal categories, disrupting both syntactic and semantic accuracy.

Fine-tuned models - Metrics results and Error analysis:

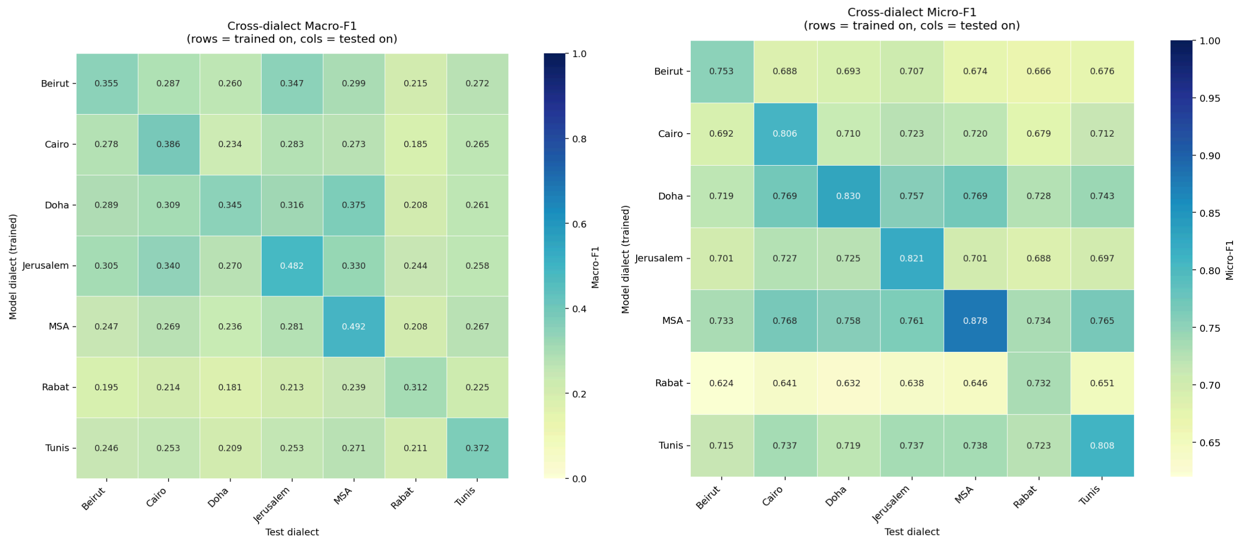
Cross dialect - Conservative F1: Across most fine-tuned models, nouns and verbs achieve relatively high recall but low precision because they absorb a disproportionate share of false positives. The confusion analysis after fine-tuning shows large off-diagonal mass into noun, e.g., prep→noun (728), verb→noun (718), adj→noun (578), adv→noun (541) and a secondary noun↔verb confusion (see Fig. 4, 5). This reflects fallback behavior under dialectal ambiguity (clitic fusion, modal prefixes): uncertain tokens are "collapsed" into broad, high-frequency categories. As a result, Precision_LB for noun/verb is markedly lower than their recall, and the Wilson-adjusted F1 (F1_LB) down-ranks these tags despite strong coverage, unlike more stable categories such as conj.

Cross dialect - Macro F1 and Micro F1 scores: As shown in Figure 4, a clear pattern of Levantine affinity emerges in the cross-dialect results. The Beirut–Jerusalem pair yields some of the strongest off-diagonal scores, with Beirut→Jerusalem achieving a Macro-F1 of 0.347 (Micro-F1 ≈ 0.707) and Jerusalem→Beirut scoring 0.305 (Micro-F1 ≈ 0.701). Notably, the Jerusalem model attains the highest performance on Beirut data among all non-Beirut models. This pattern lends empirical support to the hypothesis that geographical proximity within the Levant correlates with dialectal and linguistic similarity.

In contrast, Rabat is the weakest source overall in off-diagonal performance, with Macro-F1 mostly in the 0.18–0.24 range and Micro-F1 between 0.62–0.65. Transfer *to* Rabat is similarly low in Macro-F1 (e.g., Beirut→Rabat 0.215, Cairo→Rabat 0.185). Tunis exhibits a comparable pattern, with cross-dialect Macro-F1 scores around 0.20–0.27. These results align with the substantial geographic and structural gap between Maghrebi and Eastern dialects.

MSA functions as a strong Micro-F1 "hub": the model fine-tuned on MSA sentences performs relatively well across all dialects except MSA itself. Overall, cross-dialect transfer closely mirrors patterns of geographic and linguistic distance: Levant↔Levant transfer yields the strongest results, while Maghrebi dialects remain the most challenging, particularly for rare categories in Macro-F1.

Figure 4: *Fine-tuned Macro-F1 (left) and Micro-F1 (right) with rows representing the training dialect the model was fine tuned on and columns representing the test dialect.*

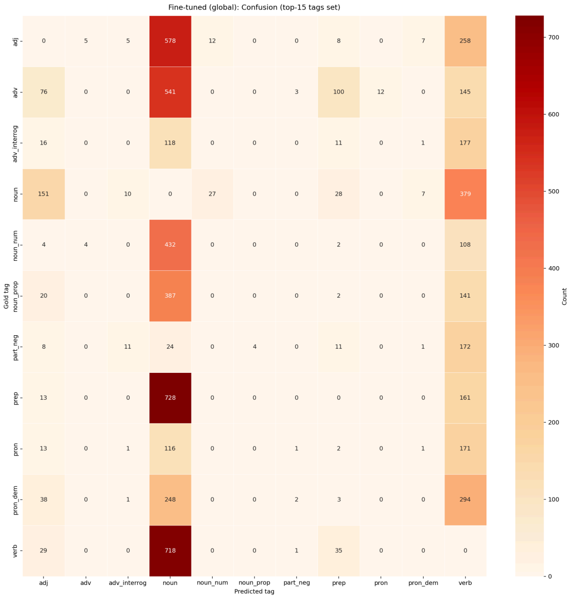


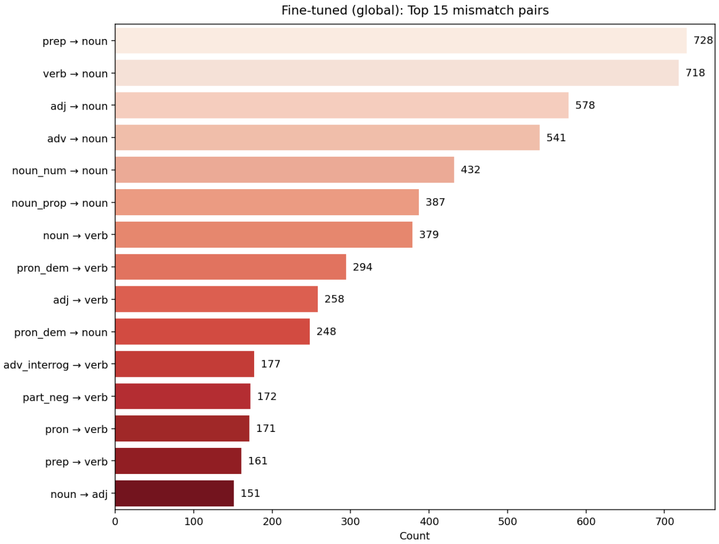
	Beirut	Cairo	Doha	Jerusalem	MSA	Rabat	Tunis
Beirut	0.355	0.287	0.260	0.347	0.299	0.215	0.272
Cairo	0.278	0.386	0.234	0.283	0.273	0.185	0.265
Doha	0.289	0.309	0.345	0.316	0.375	0.208	0.261
Jerusalem	0.305	0.340	0.270	0.482	0.330	0.244	0.258
MSA	0.247	0.269	0.236	0.281	0.492	0.208	0.267
Rabat	0.195	0.214	0.181	0.213	0.239	0.312	0.225
Tunis	0.246	0.253	0.209	0.253	0.271	0.211	0.372

	Beirut	Cairo	Doha	Jerusalem	MSA	Rabat	Tunis
Beirut	0.753	0.688	0.693	0.707	0.674	0.666	0.676
Cairo	0.692	0.806	0.710	0.723	0.720	0.679	0.712
Doha	0.719	0.769	0.830	0.757	0.769	0.728	0.743
Jerusalem	0.701	0.727	0.725	0.821	0.701	0.688	0.697
MSA	0.733	0.768	0.758	0.761	0.878	0.734	0.765
Rabat	0.624	0.641	0.632	0.638	0.646	0.732	0.651
Tunis	0.715	0.737	0.719	0.737	0.738	0.723	0.808

Cross dialect - Most Common Mismatch Pairs: In zero-shot, mBERT's bias tends to favor over-tagging as "noun" or "noun_prop" when it's uncertain. Proper nouns in particular often get over-predicted when capitalization or orthographic cues aren't reliable in dialectal Arabic, so noun_prop can become the "safe" fallback label. As shown in Figure 6, After fine-tuning we see a shift in the fallback tags, from "noun" and "noun_prop" to "noun" and "verb. This may occur due to the model's exposure to real dialectal POS distributions and more varied syntactic contexts, which helps it moderate noun_prop predictions downward. However, fine-tuning also strengthens its representation of common high-frequency verbal forms in dialects. This might be the source of the different over-prediction bias: verbs are now seen as equally "safe" fallback labels alongside nouns, particularly in contexts where morphology could plausibly fit both.

Figure 6. *Fine tuned*: *Confusion matrix for the top 15 POS tags (gold vs. predicted).* After fine-tuning, proper-noun inflation disappears; most errors now collapse diverse categories into nouns, with notable verb confusion as well.
Figure 7. *Fine tuned: Top 15 most frequent gold–predicted mismatch pairs across dialects.* The highest mismatch counts occur in prep-to-noun and verb-to-noun confusions, reflecting systematic over-prediction of nominal categories.





4 Discussion

Overall, our findings suggest that cross-dialect transfer performance may be influenced by geographic and linguistic distance, with stronger transfer often occurring within regional clusters and weaker transfer across major dialect groups. Challenges remain in reducing the performance gap between divergent dialect families. These observations could inform future work on domain adaptation, dialect clustering, and the development of resources for underrepresented dialects.

These findings highlight the need for dialect-aware pre-training or adaptation strategies. They also demonstrate that dialectal orthography, clitic fusion, and modal prefixes are persistent sources of error in zero-shot POS tagging, and should be focal points for future model development and annotation efforts.