Multi-Task Learning for Dialect Identification and POS Tagging in Arabic Digital Texts with Code-Mixing and Lexical Borrowing

Abstract

Arabic digital text often combines Modern Standard Arabic, regional dialects, colloquial morphology, non-standard spelling, and borrowing-like forms, complicating both sentence-level dialect identification and token-level POS tagging. This study compares single-task learning (STL) and multi-task learning (MTL) under a shared MARBERT-based setup on a manually reviewed Arabic digital-text corpus. After review and retention, the corpus comprised 831 sentences and 10,244 tokens, while the main six-class comparison used 810 sentences and 10,001 tokens. Evaluations across three random seeds (13, 42, and 77) showed that STL outperformed MTL on both tasks. Mean macro-F1 declined from 0.575 to 0.454 for dialect identification and from 0.512 to 0.158 for POS tagging under MTL. The results indicate that, in noisy Arabic digital text, task relatedness alone does not guarantee beneficial parameter sharing, especially when sentence-level regional abstraction and token-level morphosyntactic discrimination must be learned simultaneously.

Keywords: Arabic NLP; dialect identification; POS tagging; multi-task learning

Resumen

El texto digital árabe combina con frecuencia árabe estándar moderno, dialectos regionales, morfología coloquial, grafías no estandarizadas y formas cercanas al préstamo léxico, lo que complica tanto la identificación dialectal a nivel oracional como el etiquetado gramatical a nivel de token. Este estudio compara aprendizaje monotárea (STL) y aprendizaje multitarea (MTL) bajo una configuración compartida basada en MARBERT sobre un corpus árabe digital revisado manualmente. Tras la revisión, el corpus quedó en 831 oraciones y 10.244 tokens, mientras que la comparación principal de seis clases utilizó 810 oraciones y 10.001 tokens. La evaluación sobre tres semillas aleatorias (13, 42 y 77) mostró que STL superó a MTL en ambas tareas. El macro-F1 medio descendió de 0.575 a 0.454 en identificación dialectal y de 0.512 a 0.158 en etiquetado POS bajo MTL. Los resultados indican que la afinidad lingüística entre tareas no garantiza por sí sola una compartición beneficiosa de parámetros en texto digital árabe ruidoso.

Palabras clave: PLN árabe; identificación dialectal; etiquetado POS; aprendizaje multitarea

1. Introduction

Arabic digital communication rarely unfolds within a single, stable linguistic code. Across social media, user comments, and informal online writing, speakers shift between Modern Standard Arabic (MSA), regional dialects, and foreign languages, often within the same discourse and sometimes even within the same lexical item. This pattern reflects the broader Arabic linguistic ecology, in which standardized written Arabic coexists with multiple spoken varieties that are central to everyday interaction. Earlier resource-building work showed that written Arabic data with substantial dialectal content were both necessary and scarce, while recent survey work confirms that Arabic code-switching remains a major yet unevenly developed area in NLP (Zaidan & Callison-Burch, 2011, pp. 37–38; Hamed et al., 2025, pp. 4561–4562).

From a computational perspective, the challenge of Arabic mixed text lies not only in the coexistence of varieties, but also in their interaction at morphological, syntactic, and orthographic levels. Arabic digital writing is marked by spelling instability, colloquial morphology, lexical borrowing, borrowing-like forms, and social-media conventions that exceed the assumptions of monolingual or MSA-centered pipelines. Survey evidence further indicates that Arabic code-switched NLP remains uneven in task coverage: word-level language identification is relatively well represented, whereas POS tagging and other structurally richer tasks remain less resourced and less studied (Darwish et al., 2020, p. 1; Hamed et al., 2025, pp. 4564–4565; Obeid et al., 2020, pp. 7022–7024).

Within this landscape, POS information is especially important. Attia et al. (2019, pp. 18–20) showed that POS features improve word-level code-switching identification in Arabic and that function words and content words pattern differently across varieties in intra-sentential switching. This finding is particularly relevant when viewed alongside work on dialectal POS tagging, which has shown that robust tagging in Arabic social-media text must handle dialect diversity, non-standard spelling, and morphologically rich colloquial usage rather than only formal MSA (Darwish et al., 2020, p. 1; Abo Mokh et al., 2022, pp. 238–239). Arabic digital text also requires careful treatment of lexical origin: visually unfamiliar forms are not always evidence of external borrowing, just as orthographic irregularity does not in itself prove non-Arabic lexical status. In the present study, lexical borrowing is therefore not a peripheral issue, but rather, part of the interpretive problem linking dialect signals, POS ambiguity, and annotation reliability in mixed Arabic text.

These observations suggest that dialect identity and grammatical function should not be modeled as unrelated signals when the data themselves are structurally mixed. From a broader machine-learning perspective, multi-task learning improves generalization by training related tasks through a shared representation (Caruana, 1997, pp. 41–42). Yet the literature still lacks a focused study that jointly models sentence-level dialect identification and token-level POS tagging on reviewed Arabic digital text while interpreting results in light of lexical borrowing and mixed forms. This paper addresses that gap through a controlled comparison between single-task learning (STL) and multi-task learning (MTL) under a unified Arabic pretrained backbone. The experiments are conducted on a single manually reviewed corpus whose sentence-level dialect labels, token-level POS labels, and borrowing-related lexical judgments were adjudicated before modeling. The final comparison is reported across three random seeds under a common MARBERT-centered setup.

More specifically, the study makes four contributions. First, it develops a manually reviewed Arabic digital-text resource through a three-layer adjudication process covering sentence-level dialect labels, token-level POS labels, and borrowing-related lexical decisions. Second, it offers a controlled comparison between STL and MTL on the same reviewed corpus while keeping the main six-class experimental setting analytically distinct from the broader corpus documentation. Third, it evaluates the two training regimes through aggregate metrics and linguistically motivated analyses of dialect-sensitive cases, corrected POS patterns, and borrowing-aware interpretation. Fourth, it shows that shared learning was not beneficial under the present configuration: STL outperformed MTL on both tasks, and the degradation under MTL was especially severe for token-level POS tagging.

2. Related Work

2.1. Arabic dialect resources and dialect identification

Arabic dialect identification has developed in close connection with corpus-building efforts and shared evaluation campaigns. A foundational step was the Arabic Online Commentary Dataset, which drew attention to the scarcity of written Arabic resources with substantial dialectal content despite the overwhelming availability of MSA-oriented material (Zaidan & Callison-Burch, 2011, pp. 37–38). Later work such as AIDA approached MSA–dialect alternation as a code-switching problem and combined language modeling with morphological analysis to identify dialectal material at token level (Elfardy et al., 2014, pp. 94–95). More recently, the NADI shared tasks have confirmed that Arabic dialect identification remains sensitive to dataset design, label granularity, and evaluation setup, to the extent that NADI 2024 adopted a multi-label setting for one of its core tasks in order to better reflect overlap across Arabic varieties (Abdul-Mageed et al., 2024, pp. 709–710).

2.2. Arabic code-switching and mixed-language resources

The broader literature on Arabic code-switching has grown substantially, but it remains uneven in both task coverage and resource depth. Hamed et al. (2025, pp. 4564–4565) show that Arabic code-switched NLP has concentrated more heavily on word-level language identification than on richer grammatical tasks.

Resources such as the Cairo Student Code-Switch Corpus expanded the empirical basis for Arabic–English mixed-language research (Balabel et al., 2020, pp. 3973–3974), while Arabizi-oriented work further demonstrated that mixed Arabic digital writing often spans scripts, orthographies, and language boundaries at the same time (Shehadi & Wintner, 2022, pp. 194–195). This resource landscape confirms that Arabic digital text cannot be reduced to a simple dialect-versus-MSA distinction; rather, it frequently involves layered variation that is lexical, grammatical, orthographic, and script-based all at once.

2.3. POS tagging, grammatical signal, and dialectal variation

POS tagging occupies an important place in Arabic mixed-text analysis because grammatical category is not independent of dialectal or code-switching behavior. Attia et al. (2019, pp. 18–20) showed that POS information improves code-switching identification in Arabic and that function words and content words pattern differently across varieties in intra-sentential switching. This point resonates with broader work on dialectal POS tagging. Darwish et al. (2020, p.1) demonstrated that robust POS tagging over Arabic social-media text must account for dialect diversity and non-standard writing, while Abo Mokh et al. (2022, pp. 238–239) showed that POS tagging for Arabic dialects remains fragile under out-of-domain conditions. Taken together, these studies suggest that POS in Arabic digital text is not merely a preprocessing convenience, but a linguistically meaningful layer whose quality directly affects downstream interpretation.

2.4. Multi-task learning in Arabic NLP

From a machine-learning perspective, multi-task learning is attractive because related tasks can sometimes improve one another through shared representations and auxiliary supervision (Caruana, 1997, pp. 41–42). In Arabic NLP, this logic has already been explored in adjacent settings. Khered et al. (2025, p. 21) proposed an MTL framework for dialectal Arabic identification and translation to MSA, reporting gains over their baseline configuration. However, that work is translation-centered and encoder–decoder based, not a direct comparison between sentence classification and token classification over the same reviewed corpus. More generally, recent MTL research also warns that task relatedness alone does not guarantee positive transfer: tasks may cooperate, but they may also compete when they demand different kinds of representations or different levels of abstraction (Standley et al., 2020, pp. 9120–9132; Guo et al., 2020, pp. 3854–3863; Zhang & Yang, 2022, pp. 5586–5609).

2.5. Research gap and the position of the present study

The gap is not simply that Arabic NLP needs more multi-task learning. The more specific gap is that Arabic code-switched and mixed-text NLP still has stronger support for surface classification than for morphosyntactic interpretation, stronger coverage for word-level identification than for POS tagging, and fewer jointly reviewed datasets than the complexity of the data requires. Existing work does not yet provide a controlled comparison between STL and MTL for sentence-level dialect identification and token-level POS tagging on a manually reviewed Arabic digital-text corpus that also treats lexical borrowing as an interpretive layer rather than collapsing all non-standard forms into undifferentiated noise. The present study occupies precisely this space: it tests whether two linguistically related tasks benefit from shared supervision when both are modeled on the same reviewed Arabic digital-text resource, and when borrowing like forms are incorporated into interpretation rather than ignored or misread as superficial irregularity.

3. Methodology and Experimental Design

3.1. Corpus construction, review layers, and reporting conventions

The study draws on three aligned review layers that operate over different analytical units: sentence-level dialect labels, token-level POS tags, and a type-level borrowing lexicon. The raw sentence table comprised 839 entries, while the raw POS table contained 10,354 token instances. The auxiliary borrowing layer comprised 350 lexical types, 60 of which were manually re-examined during adjudication. After

review and retention, the final working corpus contained 831 reviewed sentences and 10,244 reviewed tokens. The study therefore remains anchored in a single corpus whose annotations were refined through review and adjudication before final experimentation.

To keep corpus description distinct from the main controlled comparison, the manuscript separates the broader reviewed corpus from the primary six-class evaluation setting. The reviewed resource remains a seven-class macro-dialect corpus (EGY, GLF, IRQ, LEV, MAGH, MSA, SUD). In the updated retained distribution, the reviewed corpus contains 156 EGY, 220 GLF, 46 IRQ, 101 LEV, 99 MAGH, 188 MSA, and 21 SUD, with corresponding token counts shown in Table 2. Because the Sudanese class is comparatively sparse, the main controlled STL–MTL comparison is framed as a six-class experiment over 810 sentences and 10,001 token instances, with the final sentence split set at 648 training sentences, 81 validation sentences, and 81 test sentences. The Sudanese material is retained only for supplementary descriptive reporting. This distinction allows the paper to preserve full corpus documentation while maintaining a controlled comparison across the six main macro-dialect labels.

Table 1. Manual adjudication across the three review layers

Layer	Raw units	Flagged review items	Uncertain	Dropped
POS	10,354	corrected token decisions 117	0	0
Borrow lexicon	350	reviewed lexical types 60	0	0
Dialect sentences	839	flagged sentences 46	9	5

Table 2. Retained reviewed corpus distribution by macro-dialect

Macro label	Sentences	Token instances	Role in reporting
EGY	156	1,922	Core comparison
GLF	220	2,577	Core comparison
IRQ	46	625	Core comparison
LEV	101	1,036	Core comparison
MAGH	99	1,243	Core comparison
MSA	188	2,598	Core comparison
SUD	21	243	Supplementary only

3.2. Three-layer manual adjudication protocol

To reduce annotation noise and make later error analysis linguistically interpretable, the study adopted a three-layer manual adjudication protocol. The first layer targeted token-level POS tags whenever the original label conflicted with the local syntactic function of the token. This was especially important for colloquial inflected verbs, borrowed technical items, and ambiguous forms that had been over-assigned to broad categories such as proper nouns. The second layer targeted lexical origin, where candidate borrowing items were re-evaluated to separate genuine borrowing from native Arabic forms that had been mistakenly absorbed into a borrowing inventory. The third layer targeted sentence-level dialect labels, where disputed cases were reassessed for macro-dialect assignment, standard-register drift, weak dialect signal, metadata noise, and template-like duplication. This protocol was motivated by the fact that Arabic digital text often combines orthographic instability, colloquial morphology, and lexical borrowing in ways that directly affect both POS interpretation and code-switch-sensitive classification.

The adjudication outcomes were operationalized conservatively. In the POS layer, 117 token-level decisions were corrected. In the borrowing layer, 60 reviewed lexical types were returned to native Arabic rather than kept as borrowing evidence. In the dialect layer, 32 flagged sentences were revised, 9 were

marked uncertain, and 5 were dropped. This conservative policy was adopted to protect the controlled comparison from avoidable label noise rather than to maximize corpus size. It also prevents later analyses from over-attributing native Arabic lexical material to borrowing simply because forms are digitally infrequent or orthographically atypical.

3.3. Task formulation

The study models two supervised tasks derived from the same retained sentence pool but operating at different granularities. The first is sentence-level dialect identification, where each sentence is assigned one macro-dialect label from the reviewed inventory. The second is token-level POS tagging, where each token receives a POS label from the harmonized tag inventory after manual correction of flagged cases. The borrowing layer is not treated as a third predictive task. Instead, it functions as an adjudication and interpretive layer used to analyze model behavior on native, borrowed, and potentially hybrid lexical material, this design follows a linguistic rather than merely computational intuition. Prior work on Arabic code-switching has shown that POS information carries a strong signal for identifying switching behavior, and that function words and content words do not distribute randomly across varieties in intra-sentential Arabic code-switching. For that reason, POS is treated here as a theoretically motivated auxiliary task rather than as a decorative annotation layer.

3.4. Core experimental comparison: STL versus MTL

The primary comparison in this study is a controlled comparison between two training regimes rather than between two pretrained model families. Under the STL setting, the two tasks are trained independently as STL-Dialect, a sentence-level dialect classifier, and STL-POS, a token-level POS tagger. Each model receives the same reviewed textual material but is optimized for only one task at a time, Under the MTL setting, a single shared encoder is trained jointly on both tasks. Sentence representations feed the dialect-classification head, while token-level contextual representations feed the POS-tagging head. The purpose is to determine whether joint supervision encourages the encoder to recover morphosyntactic regularities that improve dialect discrimination in noisy or structurally mixed Arabic digital text. The comparison is deliberately controlled: both regimes use the same reviewed corpus, the same backbone, the same shared split, and the same three-seed evaluation protocol.

3.5. Backbone selection and the role of alternative models

The final controlled comparison adopts MARBERT as the primary pretrained backbone. This choice is methodologically justified because MARBERT was introduced specifically for diverse Arabic varieties and multi-dialectal language understanding, making it a strong fit for dialect-sensitive and social-media-oriented Arabic text. By contrast, AraT5 was introduced in a text-to-text Arabic generation framework. Although it is a powerful Arabic pretrained model family, it is not the cleanest starting point for the present controlled comparison, which combines sentence classification and token classification under a shared encoder. If a second backbone is later added, it can be reported as a supplementary robustness check; it does not define the central claim of the paper.

3.6. Preprocessing and representation support

Preprocessing was intentionally conservative. The goal was to reduce noise without erasing linguistically meaningful variation. Duplicate or clearly unreliable sentence-level cases were removed during adjudication. Token-level preprocessing retained colloquial spellings and mixed forms wherever these contributed to dialect or POS interpretation. Arabic-specific preprocessing support and morphological utilities were informed by CAMeL Tools, which provides Arabic-oriented functionality for preprocessing, morphological modeling, and dialect-related NLP tasks. At the POS level, reviewed corrections were preserved as the authoritative labels for the supervised token-classification component. At the sentence level, only retained instances were carried into the dialect-classification pipeline.

3.7. Evaluation strategy

Because the dialect labels are imbalanced, dialect identification is reported primarily with macro-F1 accompanied by weighted F1, accuracy, per-class F1, and the confusion matrix. For POS tagging, token level macro-F1 is treated as the main comparative indicator, with weighted F1, accuracy, and per-tag reports used to clarify where each regime succeeds or fails. All final comparisons were conducted under the same reviewed split and across three seeds (13, 42, and 77) so that the paper's claims reflect stable behavior rather than a single favorable run. In final confirmed run metadata, the experiment used UBC-NLP/MARBERT with a maximum length of 128, a batch size of 8, a learning rate of 2e-05, weight decay of 0.01, a warmup ratio of 0.1, and three epochs for both STL and MTL. The split seed was 42, and all runs were executed on CUDA.

3.8. Linguistically guided error analysis

The final stage of the methodology is a linguistically guided error analysis built directly on the adjudicated tables. Rather than treating all errors as equivalent, the analysis distinguishes at least three broad families: POS-sensitive misclassification, especially where colloquial inflected forms had been misread in the raw annotation; lexical-origin interference, where borrowed or technically Arabicized forms blur the boundary between native and non-native lexical material; and weak-signal dialect cases, where the sentence itself provides insufficient dialectal evidence even after review. This analytic layer is essential because the paper's claim is interpretive as well as statistical. The aim is not only to test whether MTL improves performance, but also to explain whether that improvement corresponds to recoverable morphosyntactic structure in Arabic code-mixed text.

4. Results and Analytical Interpretation

4.1. Review outcomes and the retained analytical pool

The review process yielded a cleaner and more analytically reliable corpus for the final comparison. After review and retention, the corpus comprised 831 sentences and 10,244 token instances, while the main six-class experimental subset comprised 810 sentences and 10,001 token instances, split into 648 training sentences, 81 validation sentences, and 81 test sentences. Disputed sentence-level cases were not left as raw label noise, but were checked, revised where possible, and excluded where the evidence remained weak. At token level, POS decisions were corrected whenever local grammatical function clearly conflicted with the original annotation, while the borrowing layer filtered out visually foreign or suspicious forms that did not, in context, constitute genuine external borrowing. The result is therefore not simply a smaller resource, but a more interpretable one: the final STL–MTL comparison is conducted on a corpus whose main points of annotation instability had already been examined before the modeling stage. This is methodologically important because the experiment is not being asked to compensate blindly for raw noise; it is being asked to compare two training regimes on a reviewed Arabic digital-text resource.

4.2. POS correction profile

The POS review layer was especially revealing. Of the 117 corrected token-level decisions, 49 were reassigned to nouns and 44 to verbs, meaning that 93 of 117 corrections (79.5%) fell into these two categories alone. This concentration is linguistically significant: many raw errors arose not from rare edge cases, but from systematic over-reading of colloquial or technically unfamiliar forms as proper nouns or otherwise exceptional items.

Table 3. POS corrections within the flagged subset by revised label

Revised POS label	Count
Noun	49

Verb	44
Preposition	8
Adjective	5
Adverb	4
Conjunction	2
Subordinating conjunction	2
Pseudo-verb	1
Demonstrative pronoun	1
Interjection	1

The dominant review families were loan common noun (42), colloquial verb (31), and function word (10), which together account for 70.9% of the corrected subset. Tokens such as ريستارت ,ايرور, and للويندوز behave as common lexical items in context rather than named entities, whereas forms such as تهنج, حطيته, and يقولي clearly function as verbs despite non-standard spelling. The corrected subset therefore points to a recurring weakness of raw annotation in Arabic digital text: surface unfamiliarity was repeatedly mistaken for grammatical exceptionalism.

4.3. Borrowing review as an interpretive filter

The borrowing layer performed a different but equally important function: it acted as a restraint against false explanatory inflation. All 60 flagged items in the borrowing list were ultimately reassigned to native Arabic rather than retained as borrowing evidence. Items such as بياخذ ,متواطئا ,بيطلع, and رح had initially appeared suspect because of orthographic form or contextual distribution, yet manual review showed that they belonged within Arabic lexical usage rather than an external borrowing inventory, this matters because, in Arabic digital text, visual foreignness and genuine lexical borrowing are not coextensive. The result supports a more cautious interpretive line in which borrowing is treated as an analytical lens, not a catch-all explanation for every non-standard form. This caution is consistent with broader Arabic linguistic scholarship, which distinguishes borrowing, Arabization, and integrated usage rather than collapsing them into one undifferentiated category (Al-Akash, 2018, pp. 1185–1219; Al-Dulaimi & Tahsin, 2017, pp. 1203–1221).

4.4. Dialect revision profile

The dialect review layer likewise reveals that disagreement was structured rather than random. Among the 46 flagged sentence-level cases, the most common family was wrong region reassigned (17), followed by standard register (12) and weak dialect signal (9). These three families alone account for 38 of the 46 flagged cases (82.6%). The pattern suggests that the main instability in sentence-level labeling does not lie in gross annotation failure, but in three recurring interpretive tensions: confusion between closely related regional signals, drift toward MSA when a sentence is lexically generic or instructional, and genuinely weak evidence in short or formulaic digital sentences.

Table 4. Dominant dialect-review families among the flagged sentences

Review family	Count	Typical effect on corpus
Wrong region reassigned	17	Revised and retained under a different macro label
Standard register	12	Revised toward MSA and retained
Weak dialect signal	9	Marked uncertain and excluded from the core comparison

Review family	Count	Typical effect on corpus
Duplicate or template	5	Dropped from the retained resource
City-to-macro normalization	3	Normalized to a macro label and retained

These patterns are visible in cases where Gulf-oriented academic colloquial forms had initially been associated with Egyptian or Iraqi location labels, and in technically worded instructional sentences that were better read as MSA despite regional metadata. The resulting dataset is therefore better viewed as a reviewed macro-dialect resource than as a raw projection of source labels.

4.5. Main STL–MTL comparison

STL consistently outperformed MTL on both tasks. For sentence-level dialect identification, STL achieved a mean macro-F1 of 0.575 (SD = 0.031), compared with 0.454 (SD = 0.017) for MTL. The same direction held for weighted F1 (0.645 vs. 0.553) and accuracy (0.667 vs. 0.588). The contrast was sharper for token-level POS tagging. STL achieved a mean macro-F1 of 0.512 (SD = 0.014), whereas MTL dropped to 0.158 (SD = 0.025). Weighted F1 and accuracy followed the same pattern, with STL reaching 0.917 weighted F1 and 0.923 accuracy, compared with 0.726 and 0.773 for MTL. In short, shared learning did not improve either task under the present setup, and the token-level gap remained especially large.

Table 5. Mean STL–MTL comparison across the three final seeds

Task	Regime	Macro-F1	Weighted-F1	Accuracy
Dialect identification	STL	0.575	0.645	0.667
Dialect identification	MTL	0.454	0.553	0.588
POS tagging	STL	0.512	0.917	0.923
POS tagging	MTL	0.158	0.726	0.773

This result is methodologically important because the comparison controls backbone choice. Both settings were implemented under the same MARBERT-centered experimental framework; the primary difference is therefore the training regime rather than a change in model family. Under the present configuration, the result cannot be explained away as a backbone substitution effect.

4.6. Seed-level stability

The observed degradation under MTL was not an artifact of a single random initialization. The same direction appeared in all three seeds. For dialect identification, the MTL macro-F1 deficit relative to STL was -0.078 in seed 13, -0.180 in seed 42, and -0.104 in seed 77. The contrast was sharper for POS tagging, where the MTL macro-F1 deficit reached -0.328, -0.344, and -0.389, respectively. This consistency matters because the paper's central claim concerns the effect of training regime rather than the accidental behavior of one run. The result is therefore not a seed-specific irregularity but a stable empirical pattern.

Table 6. Seed-level STL–MTL macro-F1 deltas

Seed	Δ Dialect Macro-F1 (MTL − STL)	Δ POS Macro-F1 (MTL − STL)
13	0.078-	0.328-
42	0.180-	0.344-
77	0.104-	0.389-

4.7. Dialect-class findings

To make the aggregate result interpretable at class level, Table 7 reports representative class-level contrasts from the detailed saved output, rather than a three-seed aggregate. The class-level dialect breakdown shows that the underperformance of MTL was not uniformly distributed across labels. The decline was especially pronounced in Iraqi (IRQ), which collapsed to zero F1 under MTL, despite remaining at 0.333 F1 under STL. Noticeable degradation also appeared in Maghrebi (MAGH) and Levantine (LEV), whereas Egyptian (EGY), Gulf (GLF), and MSA remained comparatively stronger but still below their STL counterparts. This pattern suggests that the shared model was least robust in the dialect classes that either had lower support or required finer regional discrimination.

Table 7. Representative dialect-class contrasts between STL and MTL

Dialect	STL F1	MTL F1	Support
EGY	0.857	0.824	16
GLF	0.667	0.582	24
IRQ	0.333	0.000	5
LEV	0.636	0.476	11
MAGH	0.471	0.286	9
MSA	0.710	0.667	16

The confusion patterns reinforce this interpretation. Under MTL, confusion rose among closely related or distributionally dominant labels, while the weaker classes lost robustness most quickly. The reviewed corpus therefore reduced noise, but it did not eliminate class-sensitive asymmetries.

4.8. POS-tag findings

Table 8 reports representative POS-tag contrasts from the detailed saved output in order to show where the aggregate degradation under MTL becomes linguistically visible. The POS results reveal an even stronger form of degradation. Under STL, performance remained comparatively robust across several high- and mid-frequency tags, including noun, proper noun, verb, adjective, adverb, conjunction, foreign, and preposition. Under MTL, however, performance contracted sharply toward a narrower set of dominant categories, while a number of finer-grained tags dropped to zero or near-zero F1. This indicates that the jointly trained encoder did not preserve the token-level distinctions required for reliable POS tagging in noisy Arabic digital text. Rather than helping the model exploit shared grammatical signal, the joint setting appears to have blurred the local morphosyntactic contrasts on which POS tagging depends.

Table 8. Representative POS-tag contrasts between STL and MTL

POS tag	STL F1	MTL F1	Support
Noun	0.932	0.773	259
Proper noun	0.902	0.742	171
Verb	0.894	0.717	161
Adjective	0.889	0.000	38
Adverb	0.857	0.000	11
Conjunction	0.750	0.000	10
Foreign	1.000	0.000	18
Preposition	0.978	0.779	68

4.9. Training dynamics

The training-history trace indicates that the joint model did learn over time. Validation loss decreased across epochs, while the dialect macro-F1 rose from 0.281 to 0.496 and the POS macro-F1 rose from 0.111 to 0.148 on the validation split. This means that the MTL model was not simply failing to optimize. Yet the final jointly trained solution remained below STL on both tasks. The negative result is therefore better interpreted not as non-convergence in the narrow sense, but as evidence that the learned shared representation was less suitable than the task-specific STL alternatives.

4.10. Why did MTL underperform?

A linguistically plausible explanation is that the two tasks place pressure on different representational levels. Sentence-level dialect identification rewards broader regional abstraction, while POS tagging depends on highly local morphosyntactic sensitivity. In reviewed Arabic digital text, where noisy orthography, colloquial inflection, weak dialect cues, and borrowing-like forms often co-occur, these two forms of supervision do not necessarily align. The reviewed subsets presented earlier in this section make this tension visible. POS correction was concentrated in grammatically sensitive categories, borrowing review warned against superficial readings of non-standard forms, and dialect review showed that some sentence-level labels remained structurally weak even after adjudication. Taken together, these patterns support the interpretation that the present MTL configuration introduced interference across representational levels rather than beneficial regularization. Recent MTL research helps make sense of this asymmetry: task relatedness does not by itself guarantee productive sharing, and some task pairs compete rather than cooperate inside a common representation. The present results are also consistent with the broader concern that excessive sharing can induce negative transfer rather than balanced gains when the underlying tasks demand different granularities of representation (Standley et al., 2020, pp. 9120–9132; Guo et al., 2020, pp. 3854–3863; Zhang & Yang, 2022, pp. 5586–5609).

4.11. Implications, limitations, and next experimental direction

These findings should not be read as a blanket rejection of multi-task learning for Arabic NLP. Rather, they show that the usefulness of shared supervision is configuration specific. The present result applies to a reviewed Arabic digital-text resource, a six-class dialect setting, a sentence-plus-token pairing, and the current encoder-sharing and optimization design. Within this configuration, STL was more robust. This is itself a scientifically useful result, because it demonstrates that linguistic relatedness between tasks does not by itself guarantee MTL gains in noisy Arabic digital text. The next experimental step, therefore, is not to abandon MTL as an idea, but to frame the present evidence as identifying a boundary condition. Future work may test whether alternative loss balancing, larger reviewed corpora, or different coupling schemes between sentence-level and token-level supervision can reduce the observed interference. In the present paper, however, the central empirical conclusion is clear: the MTL setup did not improve either task, and the stronger result was obtained under separate STL optimization.

5. Conclusion

This study examined whether sentence-level dialect identification and token-level POS tagging benefit equally from shared supervision in reviewed Arabic digital text. Under the present configuration, the answer is no. Using a single corpus whose sentence-level dialect labels, token-level POS tags, and borrowing-related lexical judgments were manually reviewed before final experimentation, the study compared single-task learning (STL) and multi-task learning (MTL) within a common MARBERT-centered framework across three random seeds. The results showed that STL remained more robust than MTL on both tasks, with the gap proving especially large for token-level POS tagging.

This asymmetry is the central contribution of the paper. It shows that the two tasks, although linguistically related, do not place the same demands on the shared encoder. Sentence-level dialect

identification can remain recoverable through broader regional abstraction once weak-signal and misassigned cases have been adjudicated. POS tagging, by contrast, depends on fine-grained local morphosyntactic distinctions that remain especially vulnerable in Arabic digital text, particularly when noisy spelling, colloquial inflection, weak dialect cues, and borrowing-like forms coexist within the same sentence. In this sense, the findings do not support a blanket rejection of multi-task learning, but they do identify a clear boundary condition: linguistic relatedness alone is not sufficient to guarantee useful sharing. The borrowing layer is especially important in this interpretation. Manual review showed that visual foreignness and genuine lexical borrowing are not coextensive, and that a number of flagged items that appeared superficially non-native were ultimately reassigned to ordinary Arabic usage. Borrowing therefore functions not as a third predictive task, but as an interpretive layer that clarifies how mixed Arabic digital text should be read and modeled.

Taken together, the results contribute both empirically and methodologically. Empirically, the paper provides a controlled comparison between STL and MTL on a manually reviewed Arabic digital-text resource. Methodologically, it shows that review and adjudication can materially change what a model comparison means in noisy, mixed, and borrowing-sensitive text. Future work may test alternative loss balancing, more selective sharing strategies, adapter-based task separation, or larger reviewed corpora. Under the present design, however, the conclusion is clear: shared supervision did not improve either task; rather, task-specific optimization provided the stronger overall solution, particularly when reliable token-level grammatical discrimination was required.

References:

- Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7088–7105). Association for Computational Linguistics.
- Abdul-Mageed, M., Keleg, A., Elmadany, A., Zhang, C., Hamed, I., Magdy, W., Bouamor, H., & Habash, N. (2024). NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference* (pp. 709–728). Association for Computational Linguistics.
- Abo Mokh, N., Dakota, D., & Kübler, S. (2022). Improving POS tagging for Arabic dialects on out-of-domain texts. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 238–248). Association for Computational Linguistics.
- Al-Akash, O. M. (2018). *Masālik al-lughawiyyīn al-ʿArab fī muʿālajat ẓāhirat al-iqtirāḍ al-lughawī* [مسالك اللغويين العرب في معالجة ظاهرة الاقتراض اللغوي]. *Ḥawliyyat Kulliyyat al-Lughah al-ʿArabiyyah bi-Jirjā, 22*(2), 1185–1219.
- Al-Dulaimi, H. H., & Tahsin, A. R. (2017). *Al-iqtirāḍ al-lughawī wa-atharuhu al-ījābī wa-al-salbī fī al-lughah* [الاقتراض اللغوي وأثره الإيجابي والسلبي في اللغة]. *Al-Majallah al-ʿIlmiyyah bi-Kulliyyat al-Ādāb, 30*, 1203–1221.
- Attia, M., Samih, Y., Elkahky, A., Mubarak, H., Abdelali, A., & Darwish, K. (2019). POS tagging for improving code-switching identification in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (pp. 18–29). Association for Computational Linguistics.
- Balabel, M., Hamed, I., Abdennadher, S., Vu, N. T., & Çetinoğlu, Ö. (2020). Cairo Student Code-Switch (CSCS) Corpus: An annotated Egyptian Arabic-English corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3973–3977). European Language Resources Association.
- Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41–75.
- Darwish, K., Attia, M., Mubarak, H., Samih, Y., Abdelali, A., Màrquez, L., Eldesouki, M., & Kallmeyer, L. (2020). Effective multi-dialectal Arabic POS tagging. *Natural Language Engineering, 26*(6), 677–690.
- Elfardy, H., Al-Badrashiny, M., & Diab, M. (2014). AIDA: Identifying code switching in informal Arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 94–101). Association for Computational Linguistics.

- Guo, P., Lee, C.-Y., & Ulbricht, D. (2020). Learning to branch for multi-task learning. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 3854–3863). PMLR.
- Hamed, I., Sabty, C., Abdennadher, S., Vu, N. T., Solorio, T., & Habash, N. (2025). A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 4561–4585). Association for Computational Linguistics.
- Khered, A., Benkhedda, Y., & Batista-Navarro, R. (2025). A multi-task learning approach to dialectal Arabic identification and translation to Modern Standard Arabic. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages* (pp. 21–31). Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.
- Nagoudi, E. M. B., Elmadany, A., & Abdul-Mageed, M. (2022). AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 628–647). Association for Computational Linguistics.
- Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., & Habash, N. (2020). CAMeL Tools: An open-source Python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 7022–7032). European Language Resources Association.
- Shehadi, S., & Wintner, S. (2022). Identifying code-switching in Arabizi. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 194–204). Association for Computational Linguistics.
- Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020). Which tasks should be learned together in multi-task learning? In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 9120–9132). PMLR.
- Zaidan, O. F., & Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: An annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 37–41). Association for Computational Linguistics.
- Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(12), 5586–5609.