Boosting Fake News Detection in Arabic Dialects with Consistency-Aware LLM Merging Techniques

Abdelouahab Hocini and Kamel Smaïli

LORIA, University of Lorraine, F-54600, France

Abstract. This work explores the use of Large Language Models (LLMs) for fake news detection in multilingual and multi-script contexts, focusing on Arabic dialects. We address the challenge of insufficient digital data for many Arabic dialects by using pretrained LLMs on a diverse corpus including Modern Standard Arabic (MSA), followed by fine-tuning on dialect-specific data. We examine AraBERT, DarijaBERT, and mBERT for performance on North African Arabic dialects, incorporating code-switching and writing styles such as Arabizi. We evaluate these models on the BOUTEF dataset, which includes fake news, fake comments, and denial categories. Our approach fine-tunes both Arabic and Latin script text, with a focus on cross-script generalization. We improve accuracy using an ensemble strategy that merges predictions from AraBERT and DarijaBERT. Additionally, we introduce a new custom loss function, named CALLM to enforce consistency between models, boosting classification performance. The use of CALLM achieves significant improvement in F1-score (12.88 ↑) and accuracy (2.47 ↑) compared to the best model (MarBERT).

1 Introduction

The rapid progress of Large Language Models (LLMs) has greatly advanced Natural Language Processing (NLP), especially for high-resource languages. However, low-resource languages, such as the North African Arabic dialects, continue to present challenges due to limited data, lack of standardization, and diverse writing styles [18]. These dialects, spoken in Algeria, Tunisia, and Morocco, combine Modern Standard Arabic (MSA), French, Spanish, and English, with frequent code-switching between Arabic and Latin scripts. Initially, these dialects were primarily spoken and not written, but this has changed due to the influence of social networks. These complexities make traditional NLP models ill-suited for handling these colloquial languages.

To address this, transformer-based models such as AraBERT[4], DarijaBERT[8], and DziriBERT[2] have been developed to capture MSA and dialectal nuances. While AraBERT focuses on Arabic-script MSA, DarijaBERT supports both Arabic and Latin scripts, and DziriBERT specializes in Algerian dialects. However, these models struggle to generalize across scripts and code-switched text. Meanwhile, mBERT[6] offers cross-lingual capabilities but lacks specialization in North African dialects.

In this work, we develop an LLM-based aggregation strategy to enhance fake news detection by merging AraBERT and DarijaBERT using a custom loss function. We fine-tune these models on BOUTEF[16], a multimodal fake news corpus composed of 8 languages and dialects, leveraging a three-step training approach.

Additionally, we evaluate multiple merging strategies to improve classification accuracy: Mean Merge, Dense Merge, Custom Loss function.

2 Related Works

Recent advancements in Natural Language Processing (NLP) have led to the development of pretrained language models tailored for Arabic and its dialects[1]. AraBERT[4] is one

David C. Wyld et al. (Eds): NLPCL, BIGML, AISC – 2025
pp. 25-33, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151803

of the most widely used models, trained on Modern Standard Arabic (MSA) for tasks like sentiment analysis, named entity recognition (NER), and question answering (QA), with limited dialectal data.

To better capture North African dialects, DarijaBERT[8] specializes in Moroccan Darija written in Latin script (Arabizi). Despite its Moroccan focus, it remains effective for Algerian and Tunisian dialects, which share over 30% of their vocabulary[15]. Similarly, DziriBERT[2] is the first Transformer-based model trained for Algerian dialects, supporting both Arabic and Latin scripts. Despite being trained on only 1M tweets, it effectively captures dialectal nuances.

Other Arabic-focused models include CAMeLBERT[12] and MarBERTv2[3], the latter extending AraBERT by incorporating news-based corpora, improving Arabic text processing. Meanwhile, multilingual models like mBERT[6] and XLM-R[5] provide cross-lingual capabilities, outperforming mBERT on Arabic tasks. However, neither model is explicitly trained on dialectal Arabic, limiting their effectiveness for code-switching and Arabizi text.

Beyond language models, ensemble learning has been widely used in NLP to improve classification accuracy by combining predictions from multiple models, reducing bias, and enhancing robustness. KL divergence-based regularization[14] ensures prediction consistency in ensembles, while Knowledge Distillation (KD)[9] transfers knowledge from large to small models. Recent advances like Multiple Teacher Knowledge Distillation (MT-KD)[13, 19] enable knowledge transfer from multiple models into a unified model.

Inspired by MT-KD and ensemble learning, we propose an approach that merges outputs from specialized Arabic LLMs using a consistency loss function to improve fake news classification across Arabic dialects. Unlike distillation, we directly aggregate model predictions, leveraging their strengths while ensuring robustness for dialectal and code-switched text.

	Arabic Script		Latin Script		BOUTEF	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
AraBERT	**86.6**	**94.02**	-	-	78.14	90.7
DaridjaBERT	-	-	**78.5**	88.9	74.41	88.1
MarBERT	80	91.63	-	-	84.23	**93.30**
mBert	76.35	90.43	71.36	83.58	**84.62**	92.82
XLM-R	79.86	90.86	77.84	**90.9**	78.62	89.39

Table 1. Test Performance Comparison of Script-Specialized (AraBERT, DarijaBERT, MarBERT) and Multilingual (mBERT, XLM-R) Models on Arabic Script, Latin Script, and BOUTEF Corpus.

3 LLMs for Multilingual and Multiscript Fake News Detection

Many Arabic dialects suffer from a lack of digital text data, making it difficult to train Large Language Models (LLMs) from scratch. Since LLMs require vast corpora, dialectal Arabic alone is insufficient. A practical solution is to pretrain an LLM on a large, diverse corpus containing Modern Standard Arabic (MSA) and related languages, followed by fine-tuning on dialectal data. This enables the model to develop a general linguistic foundation while adapting to dialectal variations. In the following experiments, we use AraBERT, DarijaBERT, and mBERT, which are designed for Arabic language processing. To assess their effectiveness in handling North African Arabic dialects, we fine-tune these models on dialect-specific data.

3.1 Evaluation set

To evaluate this approach, we fine-tune the previous LLM on BOUTEF[16], a multimodal fake news dataset that captures North African dialects, code-switching (MSA, Algerian, Tunisian, French, English, and Arabizi), and misinformation patterns. This corpus categorizes news into Fake, Fake Comment, and Denial. A key challenge is distinguishing Fake News from Fake Comments, as Fake Comments represent user interactions with fake posts, often containing sarcasm, misinformation, or subjective opinions [10], making them difficult to separate from fabricated news. Table 2 details its three-class structure:

- Fake: Fabricated news with misleading information.
- Fake Comment: User interactions with Fake posts, including sarcasm or partial misinformation.
- Denial: Rebuttals or corrections aimed at refuting Fake news.

	Fake	FC	Denial	Total
MSA	372	990	251	1613
CS	10	384	8	402
FRA	109	603	116	828
ANG	18	42	15	75
ALGDIA	26	909	2	937
ARABIZI	0	125	0	125
TUNDIA	1	466	0	467
ARA-TUN	0	86	0	86
Total	536	3605	392	4533

Table 2. Distribution of news samples in the BOUTEF corpus across three classes: Fake, FC (Fake Comment), and Denial. The dataset includes diverse text sources, covering different languages, dialects, and writing styles.*Note: CS represents code-switched texts, while ARABIZI and ARA-TUN correspond to Algerian Arabizi and Tunisian Arabizi.*

3.2 Performance of Fine-Tuning on Dialect and Script Variability

Fine-tuning our models on BOUTEF allows us to assess their ability to generalize across dialects and writing scripts. To achieve this, we split BOUTEF into two subsets and adopted a three-step fine-tuning approach: first on Arabic script text, then on Latin script (Arabizi), and finally on the full dataset to evaluate cross-script generalization. Table 1 shows the performance of each model on the test sets after the fine tuning. AraBERT performs best on Arabic script, while DarijaBERT excels in handling Latin-script Arabic (Arabizi). mBERT, with its multilingual pretraining, shows overall better performance across both scripts but may not be as specialized as the other two models.

3.3 Improving Fake News Detection Using Aggregated Insights from LLMs

To improve performance, we leverage multiple LLMs to generate independent predictions, which are then combined to determine the final output. This ensemble approach mitigates individual model biases and enhances accuracy by integrating the strengths of AraBERT (for Arabic script) and DarijaBERT (for Arabizi) in news classification. Algorithm 1 details how probability distributions from both models are merged using function M. This strategy improves classification across Arabic dialects and writing styles. Next, we describe the merging policies used to combine AraBERT and DarijaBERT predictions.

Algorithm 1 Merging Predictions from AraBERT and DarijaBERT

```
Input: News article text $X$
Output: Predicted label $Y$
Step 1: Preprocess input text $X$
Step 2: Obtain probability distributions:
  $P_{\text{AB}} \leftarrow f_{\text{AB}}(X)$    // Arabic script model
  $P_{\text{DB}} \leftarrow f_{\text{DB}}(X)$    // Latin script model
Step 3: Merge probability distributions using function $M$
  $P_m(C(i)) \leftarrow M(P_{\text{AB}}, P_{\text{DB}})$
Step 4: Determine final label:
  $Y \leftarrow \arg\max_i (P_m(C(i))$    // Choose class with highest probability
return $Y$
```

Dense Merge As a black-box approach, we opted for a Multi-Layer Perceptron (MLP) as the merging policy. The input layer has a size of $n \times N$, where n represents the number of classes and N is the total number of models used (AraBERT and DarijaBERT in this case). After freezing the weights of the BERT-based models, we trained the MLP for 5 epochs to fine-tune it for the task, allowing it to learn an optimal combination of the models' outputs.

Mean Merge The final class probabilities are computed as the arithmetic mean of each model's output, ensuring a simple yet effective fusion strategy by averaging the probabilities from AraBERT and DarijaBERT. Mathematically, for each class c, the merged probability is given by:

$$P_{\text{final}}(C(j)) = \frac{1}{N} \sum_{i=1}^{N} P_i(C(j)) \quad j \in [1 \ldots n] \tag{1}$$

where N is the number of models, $P_i(C(j))$ represents the probability assigned to class $C(j)$ by model i and n the number of classes.

Mean averaging alone is not sufficient because AraBERT, which specializes in Arabic script sequences, and DarijaBERT, which focuses on Latin script sequences, use different tokenizers. This results in distinct representations for the same input, leading to potential divergence in their predictions. To address this, we introduce a consistency term in our loss function, which penalizes discrepancies between the outputs of AraBERT and DarijaBERT. By enforcing better alignment between the models while maintaining classification accuracy, this approach ensures a more stable and effective merging strategy.

Consistent Aggregation of LLMs (CALLM) Inspired by previous works [13, 11, 7], we introduce a custom loss function to merge AraBERT and DarijaBERT effectively. While KL divergence is commonly used to align model outputs in ensemble learning, its asymmetry tends to make one model dominate the other rather than allowing a balanced fusion. To avoid this issue, we opt for the Euclidean distance, which enforces consistency between the models without overpowering one over the other. By penalizing discrepancies between AraBERT and DarijaBERT's outputs, our loss function ensures a stable fusion while preserving classification accuracy. The final loss is formulated as follows:

$$\mathcal{L} = (\alpha - 1) \cdot \sum_{i=1}^{2} y_i \cdot \log(\hat{y}_i) + \alpha \cdot \|\hat{y}_1 - \hat{y}_2\|_2^2 \tag{2}$$

where:

- y_i represents the ground truth label.
- $\hat{y}_i$ denotes the final predicted probabilities.
- $\hat{y}_1$ and $\hat{y}_2$ are, accordingly, the outputs of AraBERT and DarijaBERT.
- The term $\|\hat{y}_1 - \hat{y}_2\|_2^2$ encourages consistency between the two models, promoting similar predictions.
- α is a hyper-parameter that controls the trade-off between classification loss and consistency.

Table 3 shows that MarBERT (84.62% F1-score, 93.30% accuracy) outperforms AraBERT and DarijaBERT, likely due to its exposure to both MSA and dialectal Arabic. mBERT achieves similar results (84.23 F1-score), highlighting the strength of multilingual pretraining in handling diverse Arabic text. However, **CALLM** significantly improves performance, achieving 95.52% F1-score and 95.6% accuracy, a 10.9% relative gain over MarBERT. This demonstrates the effectiveness of enforcing consistency between AraBERT and DarijaBERT, ensuring better alignment despite their different tokenization schemes. The results confirm that script-aware model merging is key to improving fake news classification in Arabic dialects.

	F1-Score	Accuracy
AraBERT	78.14	90.7
DaridjaBERT	74.41	88.1
MarBERT	**84.62**	93.30
DziriBERT	84.26	**93.47**
mBERT	84.23	92.82
XLM-R	78.62	89.39
MeanMerge	90.37	93.96
DenseMerge	85.18	94.13
CustomLoss	**95.52**	**95.6**

Table 3. Test F1-Score and Accuracy Comparison of Individual Models and Merging Strategies for Fake News Classification in Arabic Dialects.

4 Ablation Study

To evaluate the impact of our design choices, we conducted an ablation study analyzing: (1) the effect of the consistency term weight α, (2) the choice of similarity metric (KL divergence vs. Euclidean distance), and (3) the effect of applying Euclidean distance at the logit level instead of on probability distributions.

4.1 The Trade-off Between Alignment and Accuracy

CALLM introduces a consistency term controlled by the hyperparameter α, which balances classification accuracy and alignment between AraBERT and DarijaBERT. To determine the optimal value of α, we conducted a sweep over the interval $[0, 1]$ using a logarithmic scale and analyzed its impact on model performance.

Table 4 summarizes the results. Setting $\alpha = 0$ (i.e., using only cross-entropy loss) provides a baseline F1-score of 84.75% and accuracy of 93.97%. Introducing a small consistency weight ($\alpha = 0.004$) leads to a peak performance of 95.52% F1-score and 95.6% accuracy, confirming that enforcing agreement between models enhances classification. However, increasing α beyond 0.04 leads to a gradual decline in performance, with a sig-

α	F1-Score	Accuracy
0	84.75	93.97
0.004	**95.52**	**95.6**
0.032	95.04	95.11
0.126	94.29	94.45
0.251	80.25	85.48
0.501	72.83	81.24
1.0	72.83	81.24

Table 4. F1-Score and Accuracy Trends Across Different α Values

nificant drop occurring at $\alpha > 0.1$. As α increases further, accuracy and F1-score degrade sharply, and for $\alpha \geq 0.5$, performance collapses entirely. This suggests that excessive consistency enforcement distorts model outputs, overriding meaningful individual predictions and leading to poor classification accuracy. These results suggest that small values of α (0.001 - 0.032) improve performance, while excessively high values degrade classification accuracy by enforcing too strong an alignment between AraBERT and DarijaBERT. The optimal trade-off is found at $\alpha = 0.004$, demonstrating that a moderate level of consistency regularization enhances robustness without restricting model expressiveness.

4.2 Comparison: KL Divergence vs. CALLM

In our proposed merging strategy, we enforce consistency between AraBERT and DarijaBERT using a regularization term. Two commonly used approaches for aligning model outputs are KL divergence and Euclidean distance. To assess their effectiveness, we compared their impact on classification performance by varying the weight α over a logarithmic scale[17].

Table 5 provides a numerical comparison. At lower values of α, CALLM consistently outperforms KL divergence, achieving a peak accuracy of 95.60% and F1-score of 95.52% at $\alpha = 0.004$. In contrast, KL divergence reaches its highest performance at $\alpha = 0.063$ but fails to match the stability of CALLM.

α	CALLM	KL Divergence
0.004	**95.52**	94.57
0.032	95.04	95.03
0.063	94.53	**95.27**
0.126	94.29	94.49
0.251	80.25	91.74
0.501	72.83	90.96
1.0	72.83	72.83

Table 5. F1-Score Performance of CALLM vs. KL Divergence Across Different α Values

At low α values, CALLM enforces alignment without introducing instability, leading to a higher peak performance than KL divergence. However, as α increases, KL divergence maintains better performance than CALLM, particularly for $\alpha > 0.2$. This suggests that KL divergence is more resistant to over-regularization but does not perform as well when a small consistency weight is applied. For high α values ($\alpha \geq 0.5$), both methods lead to performance degradation (noticeably more severe for CALLM), confirming that excessive consistency enforcement distorts classification accuracy. The results indicate that CALLM is a better choice when using a moderate consistency regularization, whereas KL divergence may be more suitable for higher consistency enforcement settings.

4.3 Consistency Enforcement: Logits vs. Probabilities

An alternative approach to enforcing consistency is to apply CALLM at the logit level rather than on the final probability distributions. The intuition behind this is that aligning logits—the raw model outputs before applying softmax—allows greater flexibility in the resulting probability distributions. This approach modifies our loss function as follows:

$$\mathcal{L}_{\text{logit}} = (1 - \alpha) \sum_i y_i \log \hat{y}_i + \alpha \cdot |z_1 - z_2|_2^2 \tag{3}$$

where z_1 and z_2 are the raw logits from AraBERT and DarijaBERT.

To assess the impact of this choice, we compared the performance of CALLM applied to probabilities vs. logits, sweeping across different values of α. The results are shown in Table 6.

α	CALLM-P	CALLM-L
0.004	**95.52**	95.09
0.016	94.74	92.99
0.032	95.04	91.03
0.063	94.53	86.30
0.126	94.29	81.24
0.251	80.25	72.83
0.501	72.83	72.83
1.0	72.83	72.83

Table 6. F1-Score Performance Comparison: CALLM on Probabilities vs. Logits.

The results indicate that applying CALLM to probabilities consistently outperforms logit-based consistency enforcement across all values of α. While logit-level consistency performs comparably for very small α values, its performance degrades rapidly as α increases, with accuracy dropping from 92.99% at $\alpha = 0.016$ to just 81.24% at $\alpha = 0.126$. At $\alpha > 0.2$, both methods collapse to poor performance, confirming that excessive consistency regularization is detrimental.

Thus, CALLM on probability distributions remains the preferred approach, achieving higher accuracy and stability for fake news classification in Arabic dialects.

4.4 Key Takeaways

Our ablation study reveals key insights into the impact of design choices in our merging strategy:

- Moderate consistency improves performance: Small α values $[0.001, 0.032]$ enhance accuracy and F1-score by aligning AraBERT and DarijaBERT, while excessive enforcement ($\alpha > 0.1$) degrades performance, emphasizing the need for balance.
- CALLM outperforms KL divergence at low α values ($\alpha = 0.004$), achieving higher peak performance. However, KL divergence is more stable when $\alpha > 0.2$.
- Applying CALLM at the probability level yields better accuracy and F1-score than logit-level enforcement.
- Across all experiments, excessive consistency enforcement ($\alpha \geq 0.5$) leads to a collapse in classification accuracy and F1-score, highlighting the importance of finding an optimal trade-off between alignment and individual model flexibility.

These findings highlight the importance of controlled consistency enforcement, with CALLM on probabilities proving the most stable and effective for Arabic dialect fake news classification.

5 conclusion

In this work, we introduced a consistency-aware LLM aggregation strategy for fake news classification in north African Arabic dialects, addressing the challenges posed by code-switching, diverse scripts, and dialectal variations. By merging AraBERT and DarijaBERT with a custom loss function (CALLM), we achieved state-of-the-art performance on the BOUTEF dataset, significantly outperforming individual models. Our results demonstrate that controlled model merging can enhance classification in low-resource NLP tasks, particularly for dialectal Arabic and script-diverse datasets. The findings suggest that script-aware LLM fusion is a promising approach for improving robustness, generalization, and cross-script alignment in fake news detection. For future work, we aim to incorporate adaptive weighting mechanisms to optimize model fusion. Additionally, we will evaluate generalization on other Arabic dialect datasets to assess cross-dialect robustness. Finally, we plan to extend our approach to other NLP applications beyond fake news classification.

References

1. Abbas Yousef, M., ElKorany, A., Bayomi, H.: Fake-news detection: a survey of evaluation arabic datasets. Social Network Analysis and Mining **14**(1), 1–18 (2024)
2. Abdaoui, A., Berrimi, M., Oussalah, M., Moussaoui, A.: Dziribert: a pre-trained language model for the algerian dialect. arXiv preprint arXiv:2109.12346 (2021)
3. Abdul-Mageed, M., Elmadany, A., Nagoudi, E.M.B.: 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, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.551, https://aclanthology.org/2021.acl-long.551
4. Antoun, W., Baly, F., Hajj, H.: Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104 (2020)
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. CoRR **abs/1911.02116** (2019), http://arxiv.org/abs/1911.02116
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
7. E. Almandouh, M., Alrahmawy, M.F., Eisa, M., Elhoseny, M., Tolba, A.: Ensemble based high performance deep learning models for fake news detection. Scientific Reports **14**(1), 26591 (2024)
8. Gaanoun, K., Naira, A.M., Allak, A., Benelallam, I.: Darijabert: a step forward in nlp for the written moroccan dialect. International Journal of Data Science and Analytics pp. 1–13 (2024)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531
10. Hocini, A., Smaili, K.: Detecting fake news: Exploring key features in multilingual arabic dialect corpus. In: Hdioud, B., Aouragh, S.L. (eds.) Arabic Language Processing: From Theory to Practice. pp. 236–248. Springer Nature Switzerland, Cham (2025)
11. Huang, Y., Feng, X., Li, B., Xiang, Y., Wang, H., Qin, B., Liu, T.: Ensemble learning for heterogeneous large language models with deep parallel collaboration (2024), https://arxiv.org/abs/2404.12715
12. Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., Habash, N.: The interplay of variant, size, and task type in Arabic pre-trained language models. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Online) (Apr 2021)
13. Jiang, Y., Feng, C., Zhang, F., Bull, D.: MTKD: Multi-Teacher Knowledge Distillation for Image Super-Resolution, p. 364–382. Springer Nature Switzerland (Oct 2024)
14. Malinin, A., Gales, M.: Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness (2019), https://arxiv.org/abs/1905.13472

15. Meftouh, K., Harrat, S., Smaïli, K.: PADIC: extension and new experiments. In: 7th International Conference on Advanced Technologies. Antalya, Turkey (Apr 2018), https://hal.science/hal-01718858
16. Smaïli, K., Hamza, A., Langlois, D., Amazouz, D.: Boutef: Bolstering our understanding through an elaborated fake news corpus. In: Hdioud, B., Aouragh, S.L. (eds.) Arabic Language Processing: From Theory to Practice. pp. 107–123. Springer Nature Switzerland, Cham (2025)
17. Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A., Romero-Gonzalez, J.A.: Loss functions and metrics in deep learning (2024), https://arxiv.org/abs/2307.02694
18. Touahri, I., Mazroui, A.: Survey of machine learning techniques for arabic fake news detection. Artificial Intelligence Review **57** (05 2024). https://doi.org/10.1007/s10462-024-10778-3
19. Wu, C., Wu, F., Huang, Y.: One teacher is enough? pre-trained language model distillation from multiple teachers. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4408–4413. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.387, https://aclanthology.org/2021.findings-acl.387/