Evaluating Calibration of Arabic Pre-trained Language Models on Dialectal Text

Ali Al-Laith
Copenhagen University, Denmark
alal@di.ku.dk

Rachida Kebdani
University of Verona, Italy
rashida-kebdani@hotmail.com

Abstract

While pre-trained language models have made significant progress in different classification tasks, little attention has been given to the reliability of their confidence scores. Calibration, how well model confidence aligns with actual accuracy, is essential for real-world applications where decisions rely on probabilistic outputs. This study addresses this gap in Arabic dialect identification by assessing the calibration of eight pre-trained language models, ensuring their predictions are not only accurate but also reliable for practical applications. We analyze two datasets: one with over 1 million text samples and the Nuanced Arabic Dialect Identification dataset(NADI-2023). Using Expected Calibration Error (ECE) as a metric, we reveal substantial variation in model calibration across dialects in both datasets, showing that prediction confidence can vary significantly depending on regional data. This research has implications for improving the reliability of Arabic dialect models in applications like sentiment analysis and social media monitoring.

1 Introduction

Arabic pre-trained language models (PLMs) have advanced significantly in dialect identification and classification, with most research focusing on improving accuracy and dataset development. However, these efforts often overlook calibration—how well a model's confidence scores align with the true probability of correct predictions(Nixon et al., 2019). Calibration is crucial for Arabic dialect applications, where nuanced regional variations in language can lead to significant social and cultural implications if predictions are unreliable. In real-world applications like sentiment analysis, social media monitoring, and policy-making, accurate yet calibrated predictions are essential to support informed decision-making.

This study addresses this gap by evaluating the calibration of existing Arabic pre-trained models on dialectal text. Using 1 million text samples automatically annotated and NADI-2023 datasets, we conduct calibration analysis exclusively on cases where all eight models unanimously agree on dialect labels, focusing on high-confidence predictions. We employ metrics such as Expected Calibration Error (ECE) to measure the alignment between model confidence and accuracy, assessing the trustworthiness of these models in dialect classification.

By focusing on calibration, this work goes beyond accuracy metrics to highlight the reliability of model predictions. Calibration evaluation not only aids in model selection for high-stakes applications but also informs areas for improvement, ensuring that Arabic dialect models are both accurate and dependable in practice.

2 Related Work

2.1 Arabic Dialect Datasets

Dialectal Arabic (DA) encompasses the diverse spoken forms of Arabic used across the Arab world, differing significantly from Modern Standard Arabic (MSA) in phonology, morphology, orthography, and syntax(Bouamor et al., 2014). DA is typically divided into regional groups, including Egyptian, North African, Levantine, Gulf, and Yemeni, with each containing sub-varieties like Tunisian, Lebanese, and Saudi dialects(Zaghouani and Charfi, 2018). Given DA's prevalence in daily communication, incorporating DA resources into LLM training is crucial for creating models that understand and generate Arabic as it is spoken in real-world contexts. The MADAR Twitter corpus, used in the MADAR shared task on fine-grained Arabic dialect identification, comprises 2,980 Twitter user profiles from 21 countries, facilitating dialect identification in Twitter user profiles (Bouamor et al., 2019). The Gumar corpus, a large-scale collection of Gulf Arabic, includes 1,236 forum novels totaling around 112 million words, with manual

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 68–76
January 20th, 2025.

document-level annotations for sub-dialect information across the Gulf Cooperation Council countries: Bahrain, UAE, Kuwait, Saudi Arabia, Oman, and Qatar (Khalifa et al., 2016). Nuanced Arabic Dialect Identification (NADI) introduced different datasets for Arabic dialect identification in different level such as country or city levels(Abdul-Mageed et al., 2022, 2023a, 2024). Baimukan et al. (2022) introduced the first unified three-level hierarchical schema (region-country-city) for dialectal Arabic classification. By mapping 29 datasets to this schema, they enabled their aggregation and demonstrated its effectiveness by building language models for dialect identification.

2.2 Arabic Dialect Pre-trained Language Models

The development of dialect-specific BERT-based models for Arabic has emerged to address the linguistic diversity across the Arab world, resulting in several models specialized for individual dialects. SudaBERT (Elgezouli et al., 2021), for instance, focused on Sudanese Arabic, outperforming Arabic-BERT (Talafha et al., 2020) in sentiment analysis (SA) for the Sudanese dialect, though Arabic-BERT showed stronger performance in Modern Standard Arabic (MSA) across both SA and named entity recognition (NER). Similarly, AraRoBERTa was designed for seven dialects (Saudi, Egyptian, Kuwaiti, Omani, Lebanese, Jordanian, and Algerian), employing RoBERTa architecture with various supervision approaches (AlYami and Al-Zaidy, 2022). AraRoBERTa performed particularly well in Saudi and Egyptian dialects due to larger dataset availability, while semi-supervised training improved results for certain dialects like Egyptian and Algerian.

For the Algerian dialect, DziriBERT was trained on over a million tweets, excelling in SA, emotion classification, and topic classification tasks, with MARBERT following closely (Abdaoui et al., 2021). Haddad et al. (2023) introduced Tun-BERT, targeting Tunisian Arabic, performed best in SA and dialect identification but was outperformed in reading comprehension by AraBERT (Antoun et al., 2020) and GigaBERT (Safaya et al., 2020). Moroccan Arabic, or Darija, has also been addressed with models like MorrBERT (Moussaoui and El Younnoussi, 2023), DarijaBERT, and Atlas-Chat (Shang et al., 2024). MorrBERT and its RoBERTa-based counterpart MorRoBERTa achieved high accuracy in SA and dialect identification, with DarijaBERT variants showing strong performance in dialect identification, SA, sarcasm detection, and topic classification. Atlas-Chat, the latest Moroccan Arabic model, achieved notable results in sentiment analysis and translation.

In addition, AlcLAM, a model focusing on Arabic dialects in general, excelled in dialect identification and offensive language detection compared to other models (Ahmed et al., 2024). SaudiBERT (Qarah, 2024b) and EgyBERT (Qarah, 2024a), specifically trained on Saudi and Egyptian dialects respectively, showed strong performances across various tasks such as sarcasm detection, gender identification, and event detection, often surpassing established models like AraBERT, CAMeLBERT (Inoue et al., 2021), and MARBERT. This growing body of dialect-specific models demonstrates the significance of tailoring architectures and training data to regional linguistic features, leading to enhanced performance in dialect-relevant NLP tasks across the Arab world.

2.3 Calibration of Pre-trained Language Models

Calibrating probabilistic predictive models is essential for reliable prediction and decision-making in AI. Naeini et al. (2015) introduced Bayesian Binning into Quantiles (BBQ), a non-parametric, computationally efficient calibration method that post-processes binary classification outputs, making it compatible with various classifiers and demonstrating high accuracy in experiments on real and simulated datasets. Desai and Durrett (2020a) examined calibration in BERT and RoBERTa models for tasks like natural language inference, paraphrase detection, and commonsense reasoning, evaluating both in-domain and out-of-domain settings to account for model uncertainty. Baan et al. (2022) introduced an instance-level calibration based on human uncertainty, validated through a ChaosNLI dataset case study, which examines temperature scaling under human judgment. Neural network classification models often rely on maximum predicted probabilities as confidence scores, which typically require post-processing calibration to improve reliability. By transforming multi-class calibration into a binary surrogate task, this approach enhances calibration efficiency and significantly improves results across various neural networks for image and text classification (LeCoz et al., 2024).

Jiang et al. (2021) explored language model calibration by assessing how well models like T5,

BART, and GPT-2 match predicted probabilities to correctness likelihoods, finding them poorly calibrated on QA tasks. Calibration methods such as fine-tuning and post-hoc adjustments showed improvement in confidence accuracy across diverse datasets. Zhang et al. (2021) extended calibration in QA by combining confidence scores with input context and data augmentation, achieving 5-10% accuracy gains on reading comprehension benchmarks and opening calibration study in open retrieval settings, showing robust gains across tasks. Yang et al. (2023) benchmarked multilingual Large Language Model (LLM) calibration on QA tasks across languages, covering encoder-only, encoder-decoder, and decoder-only models (110M to 7B parameters) across high- and low-resource languages. They found that decoder-only models, like LlaMa2, benefit from in-context learning, and incorporating cheaply translated samples improves calibration, particularly for non-English languages.

For stance detection, Li and Caragea (2023) used knowledge distillation with soft labels and iterative teacher-student learning to enhance model performance, implementing dynamic temperature scaling to calibrate predictions, which improved stance detection results on three datasets.

3 Methodology

3.1 Dataset

We use two types of annotated datasets: automatically annotated data using eight Pre-trained Language Models(PLMs), limited to samples with unanimous dialect labels, and manually annotated data by human annotators.

For the automatic annotations, we compile over 1 million text samples from multiple datasets. The first source is the Arabic Dialect Identification dataset[1], with more than 360,000 labeled Arabic sentences, built by integrating arabic_pos_dialect [2], IADD (Zahir, 2022)[3], QADI (Abdelali et al., 2020)[4], and the MADAR corpus (Bouamor et al., 2018)[5]. Additionally, we select over 500,000 tweets from AraSenCorpus, a collection of 4.5 million tweets in Modern Standard Arabic and dialects (Al-Laith et al., 2021), and over 200,000 samples from a 5.5 million tweet corpus for emotion and symptom classification (Al-Laith and Alenezi, 2021). As the collected tweets were crawled from social media, the data are expected to be noisy and should be cleaned up before performing any of the NLP tasks to get better results. We apply text preprocessing steps, including the removal of URLs, hashtags, mentions, and duplicate tweets.

For manual annotations, we use the NADI 2023 dialect identification dataset (Abdul-Mageed et al., 2023b), with PLMs predicting dialects across training and development sets, totaling 15,400 samples across 14 dialects (1,100 samples per dialect).

3.2 Pre-trained Language Models

We use the following Pre-trained Language Models (PLMs) to conduct the Arabic dialect prediction experiments:

1. **Arabic Dialect Identification Model**[6] **(Model 1)**: The model is trained to accurately identify spoken dialects in Arabic text. It was trained using a combination of publicly available datasets and fine-tuned on their own dataset. With high accuracy in identifying Arabic dialects, the model can be utilized in a variety of applications.

2. **CAMeLBERT-MSA DID MADAR Twitter-5 Model** [7] **(Model 2)**: The model is a dialect identification (DID) model specifically designed for Arabic (Inoue et al., 2021). It was fine-tuned from the CAMeLBERT-MSA model using the MADAR Twitter-5 dataset, which includes 21 labels. This model is particularly useful for identifying different Arabic dialects in social media texts.

3. **CAMeLBERT-Mix DID NADI Model**[8]**(Model 3)**: The model is a dialect identification (DID) model that was built by fine-tuning the CAMeLBERT-Mix model. For the fine-tuning, we used the NADI Coountry-level dataset[9], which includes 21 labels.

[1] https://github.com/Lafifi-24/arabic-dialect-identification

[2] https://huggingface.co/datasets/arabic_pos_dialect

[3] https://github.com/JihadZa/IADD

[4] https://github.com/qcri/QADI

[5] https://sites.google.com/nyu.edu/madar/?pli=1

[6] https://huggingface.co/lafifi-24/arbert_arabic_dialect_identification

[7] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa-did-madar-twitter5

[8] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix-did-nadi

[9] https://sites.google.com/view/nadi-shared-task

4. **ADI-NADI-2023**[10] **(Model 4)**: A BERT-based model fine-tuned to perform single-label Arabic Dialect Identification (Keleg and Magdy, 2023).

5. **Arabic-MARBERT-dialect-Identification-City Model**[11] **(Model 5)**: The model is a dialect identification model that was built by fine-tuning the MARBERT model. For the fine-tuning, I used MADAR Corpus 26 dataset, which includes 26 labels(cities).

6. **Bert base arabic camelbert MSA finetunedArabic Dialect Identification**[12] **(Model 6)**: The model was trained on QADI dataset from (Abdelali et al., 2020).

7. **CAMeLBERT-MSA DID NADI Model**[13] **(Model 7)**: It is a dialect identification (DID) model that was built by fine-tuning the CAMeLBERT Modern Standard Arabic (MSA) model[14]. For the fine-tuning, we used the NADI Coountry-level dataset[15], which includes 21 labels.

8. **NADI-2024-baseline**[16] **(Model 8)**: A BERT-based model fine-tuned to perform single-label Arabic Dialect Identification (ADI).

3.3 Dialect Selection

Table 1 displays the range of dialects encompassed by each of the pre-trained language models (PLMs) discussed. Some models offer predictions of Arabic dialects at the city level, we have aligned these cities with their respective countries for a more comprehensive understanding. Since the number of labels varies across models and some dialects such as the Qatari dialect has no sample annotated by all models, we have focused our analysis on the common labels, selecting 14 out of 22 labels shared among all models' label sets.

[10] https://huggingface.co/AMR-KELEG/ADI-NADI-2023

[11] https://huggingface.co/Ammar-alhaj-ali/arabic-MARBERT-dialect-identification-city

[12] https://huggingface.co/Abdelrahman-Rezk/bert-base-arabic-camelbert-msa-finetuned-Arabic_Dialect_Identification_model_1

[13] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa-did-nadi

[14] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa/

[15] https://sites.google.com/view/nadi-shared-task

[16] https://huggingface.co/AMR-KELEG/NADI2024-baseline

4 Experiments and Results

4.1 Dialect Prediction Experiment

The process of dialect prediction with Hugging Face models involves loading a pre-trained model and tokenizer to numerically encode the input text, enabling model processing. The model produces logits, which are then converted into probabilities, with the highest probability determining the sample's predicted label. This approach efficiently supports tasks such as text classification and named entity recognition, offering a standardized method for leveraging pre-trained models in NLP.

After predicting dialects for each sample with all 8 selected models, we computed the majority label separately for both the automatically and manually annotated datasets. Figure 1 displays the count of models agreeing on the same label, alongside sample counts and frequencies for each dataset. In the automatically annotated dataset, 127,646 samples had full agreement across all 8 models (around 8.5%), while only 9,852 samples (approximately 0.66%) received 8 different labels, indicating minimal consensus. In the manually annotated dataset, 2,581 samples had unanimous agreement, representing around 4%, while only 10 samples (less than 0.01%) received 8 different labels, further underscoring the rarity of full disagreement.

Figure 2 provides a detailed view of the percentage of samples identified per dialect and the number of models that concurred on each label for both datasets. In the automatically annotated data, models most frequently agreed on labels with 2 to 5 models in agreement, while in the manually annotated dataset, model agreement levels were generally higher, with 4 to 8 models showing more consistent label matches. This discrepancy highlights the influence of annotation style on model consensus, with the manually annotated dataset exhibiting slightly higher overall agreement among models.

For the calibration analysis, we focus on samples that received the same label from all models (127,646 samples) from the automatically annotated dataset, while we include all samples from the manually annotated NADI-2013 dataset for model calibration analysis.

4.2 Expected Calibration Error (ECE)

In this experiment, we use Expected Calibration Error (ECE), a metric that measures how well the

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Is Included?
Algeria	✓	✓	✓	✓	✓	✓	✓	✓	✓
Bahrain	✓	✓	✓	✓	✗	✓	✓	✓	✗
Djibouti	✗	✓	✓	✗	✗	✗	✓	✗	✗
Egypt	✓	✓	✓	✓	✓	✓	✓	✓	✓
Iraq	✓	✓	✓	✓	✓	✓	✓	✓	✓
Jordan	✓	✓	✓	✓	✓	✓	✓	✓	✓
KSA	✓	✓	✓	✓	✓	✓	✓	✓	✓
Kuwait	✓	✓	✓	✓	✗	✓	✓	✓	✗
Lebanon	✓	✓	✓	✓	✓	✓	✓	✓	✓
Libya	✓	✓	✓	✓	✓	✓	✓	✓	✓
MSA	✓	✗	✗	✗	✓	✗	✗	✗	✗
Mauritania	✗	✓	✓	✗	✗	✗	✓	✗	✗
Morocco	✓	✓	✓	✓	✓	✓	✓	✓	✓
Oman	✓	✓	✓	✓	✓	✓	✓	✓	✓
Palestine	✓	✓	✓	✓	✓	✓	✓	✓	✓
Qatar	✓	✓	✓	✓	✓	✓	✓	✓	✓
Somalia	✗	✓	✓	✗	✗	✗	✓	✗	✗
Sudan	✓	✓	✓	✓	✓	✓	✓	✓	✓
Syria	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tunisia	✓	✓	✓	✓	✓	✓	✓	✓	✓
UAE	✓	✓	✓	✓	✗	✓	✓	✓	✗
Yemen	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Arabic Dialects included in our analysis.

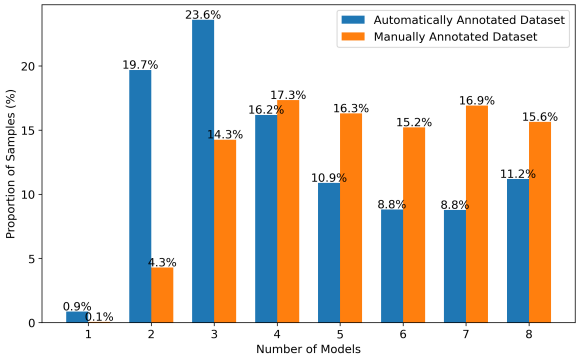

Figure 1: Sample proportion by number of models agreeing to assign the same dialect.

model's predicted probabilities reflect the true accuracy (Desai and Durrett, 2020b):

$$ECE = \sum_{k=1}^{K} \frac{|B_k|}{n} |acc(B_k) - conf(B_k)|$$

where $K = 10$ is the number of bins (confidence intervals), $|B_k|$ is the number of samples in bin k, $acc(B_k)$ is the accuracy in bin k, and $conf(B_k)$ is the average confidence in bin k. The ECE value reflects how well-calibrated a model's confidence estimates are, with lower ECE indicating better calibration.

4.2.1 Automatically Annotated Data

ECE is used to assess the calibration quality of eight Arabic pre-trained language models on dialectal text by comparing model confidence with actual accuracy on a subset where all models agreed on the same label. ECE is calculated by binning predicted confidence scores, then measuring the discrepancy between the average confidence and accuracy within each bin. This error quantifies how closely model confidence aligns with observed accuracy, indicating whether models tend to over- or under-predict. By focusing on samples with unanimous agreement, the experiment aims to reveal calibration disparities among models that exhibit high predictive consensus, offering insights into their reliability when applied to Arabic dialect classification. It is shown that both Model 1 & 6 achieved a relatively low ECE of 0.07, as shown in Figure 3, indicating that both models are reasonably well-calibrated. In contrast, Model 4 achieves high ECE of 0.44, indicating that the model is not well-calibrated.

4.2.2 Manually Annotated Data

We use the same ECE formula described in the previous section. The results of the experiment reveal significant variation in Expected Calibration Error (ECE) across the models, indicating differing levels of calibration quality. Model 4 exhibits the high-

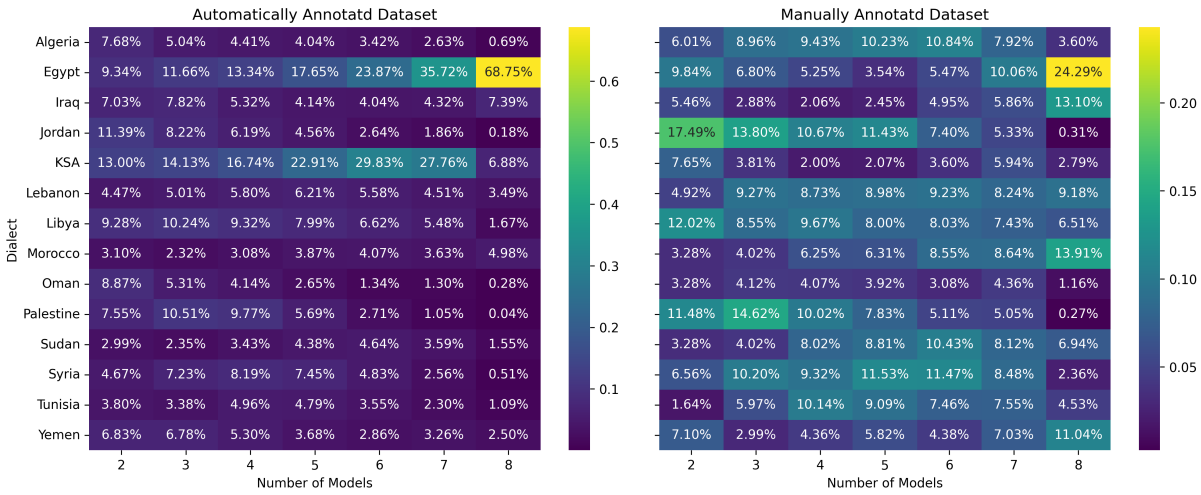

Figure 2: Sample proportion and number of models agreeing to assign the same dialect.

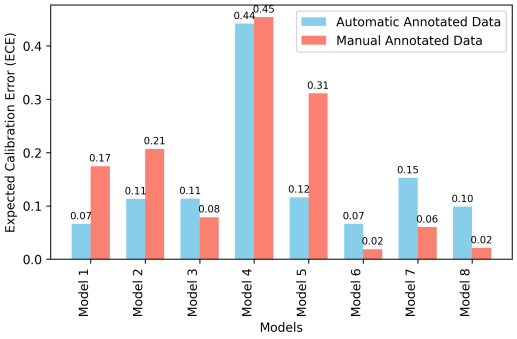

Figure 3: Expected Calibration Error (ECE) values for each model on both datasets.

est ECE at 0.45, suggesting poor calibration and a substantial gap between predicted probabilities and actual outcomes. Similarly, Models 2 and 5 show relatively high ECE values of 0.21 and 0.31, respectively, also pointing to weaker calibration. In contrast, Models 6, 7, and 8 achieve notably low ECE scores (0.02, 0.06, and 0.02), demonstrating better alignment between predictions and actual labels, indicating that these models are more reliably calibrated. Model 3 also shows moderate calibration with an ECE of 0.08. Overall, the results highlight the variance in calibration performance, with some models showing potential for practical application due to better-calibrated predictions, while others require further adjustment to improve reliability. Figure 4 shows the ECE values of each model on the NADI dataset.

5 Result Analysis and Discussion

The calibration analysis across models and dialects reveals distinct trends in model reliability on both automatically and manually annotated datasets. Models 1 and 8 demonstrate more consistent calibration across dialects and datasets, suggesting they are better suited for varied dialectal data and annotation styles. In contrast, Models 4 and 5 show higher calibration errors, especially on manually annotated data, indicating a greater sensitivity to the complexities introduced by human annotations. This difference underscores the potential need for fine-tuning or recalibration when applying these models to manually annotated datasets to enhance their predictive confidence.

Additionally, the calibration differences across dialects reveal that certain dialects, such as Palestinian and Sudanese, are more challenging for the models to interpret consistently, displaying higher calibration errors. This pattern suggests that these dialects might require additional data or targeted adjustments to improve model alignment. Overall, these findings emphasize the importance of considering both annotation type and dialect specificity when evaluating model calibration, as these factors can significantly impact model reliability in multilingual and multi-dialectal applications.

6 Limitation

This work has some notable limitations that could impact the generalizability and comprehensiveness of the findings. First, while the analysis provides insights into the calibration of eight pre-trained language models, it is constrained by the choice and availability of these models. Each model has been pre-trained on varying datasets, which may lack consistent or comprehensive coverage of spe-

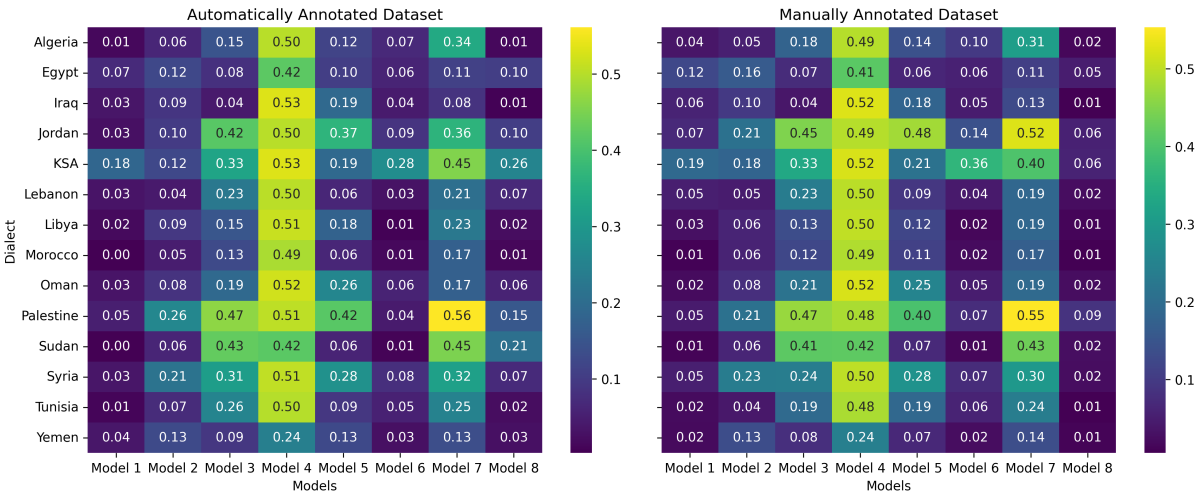

Figure 4: Expected Calibration Error (ECE) Values for Each Dialect and Model.

cific Arabic dialects, thereby limiting our ability to capture the full linguistic diversity within Arabic dialects. Consequently, the calibration results might reflect biases inherent in the pre-training datasets rather than purely dialectal features.

Second, the study relies solely on Expected Calibration Error (ECE) as the calibration metric, which, while informative, provides only a single perspective on model calibration quality. ECE does not capture all aspects of prediction reliability, such as miscalibration at different confidence levels or the potential impacts of class imbalance in dialect distribution. Integrating additional calibration metrics, like Brier Score or Maximum Calibration Error (MCE), might provide a more comprehensive evaluation of model performance across dialects.

Additionally, the study does not consider the contextual or pragmatic nuances present in real-world dialectal Arabic, as these models may not account for complex language variations or code-switching phenomena commonly seen in Arabic dialects. This limitation may impact the reliability of model predictions when applied to more dynamic or informal Arabic text data, such as social media posts, which often contain non-standard dialectal expressions.

Finally, the study focuses on calibration without incorporating linguistic or sociolinguistic factors that could influence model performance across dialects. Factors such as geographical proximity, historical language influences, and sociolinguistic prestige of certain dialects could affect model calibration in ways that ECE alone cannot capture. Future research could benefit from a more interdisciplinary approach that considers these factors, potentially enhancing model calibration for specific dialectal groups.

7 Conclusion and Future Work

The ECE analysis demonstrates considerable variability in model calibration performance across both automatically and manually annotated datasets for Arabic dialect prediction. Models 1, 6, and 8 exhibit relatively lower ECE scores, suggesting they maintain more reliable calibration across different dialects and annotation types. Conversely, Models 4 and 5 display notably higher calibration errors, particularly with manually annotated data, which highlights the impact of annotation style on calibration outcomes. This variability suggests that certain models are better suited to dialectal Arabic tasks, though a one-size-fits-all approach may not be feasible given the complexity of the data.

Since the data in the automatically annotated dataset was randomly sampled without balancing dialect distribution, future work can explicitly address this by exploring techniques like resampling or re-weighting to assess their impact on the reliability of the findings. We also plan to improve model calibration in Arabic dialect prediction with focus on dialect-specific calibration techniques, with a particular emphasis on dialects that exhibit higher calibration errors, such as Palestinian and Sudanese Arabic. Approaches such as fine-tuning models with dialect-specific data or applying post-hoc calibration methods may enhance model reliability for these challenging dialects. Additionally, investigating why certain models like Models 1, 6, and 8 perform better could yield insights into ar-

chitectural or pre-training factors that contribute to calibration efficacy. Incorporating domain-specific knowledge on linguistic features unique to each dialect may further enhance calibration, especially for dialects with distinct phonological or lexical characteristics.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023a. Nadi 2023: The fourth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2310.16117*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023b. Nadi 2023: The fourth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2310.16117*.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2407.04910*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

Murtadha Ahmed, Saghir Alfasly, Bo Wen, Jamaal Qasem, Mohammed Ahmed, and Yunfeng Liu. 2024. Alclam: Arabic dialectal language model. *arXiv preprint arXiv:2407.13097*.

Ali Al-Laith and Mamdouh Alenezi. 2021. Monitoring people's emotions and symptoms from arabic tweets during the covid-19 pandemic. *Information*, 12(2):86.

Ali Al-Laith, Muhammad Shahbaz, Hind F Alaskar, and Asim Rehmat. 2021. Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus. *Applied Sciences*, 11(5):2434.

Reem AlYami and Rabah Al-Zaidy. 2022. Weakly and semi-supervised learning for Arabic text classification using monodialectal language models. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 260–272, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. *arXiv preprint arXiv:2210.16133*.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The madar shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Shrey Desai and Greg Durrett. 2020a. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Shrey Desai and Greg Durrett. 2020b. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Mukhtar Elgezouli, Khalid N Elmadani, and Muhammed Saeed. 2021. Sudabert: A pre-trained encoder representation for sudanese arabic dialect. In *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–4. IEEE.

Hatem Haddad, Ahmed Cheikh Rouhou, Abir Messaoudi, Abir Korched, Chayma Fourati, Amel Sellami, Moez Ben HajHmida, and Faten Ghriss. 2023. Tunbert: pretraining bert for tunisian dialect understanding. *SN Computer Science*, 4(2):194.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. *arXiv preprint arXiv:2310.13661*.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of gulf arabic. *arXiv preprint arXiv:1609.02960*.

Adrien LeCoz, Stéphane Herbin, and Faouzi Adjed. 2024. Confidence calibration of classifiers with many classes. *Preprint*, arXiv:2411.02988.

Yingjie Li and Cornelia Caragea. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6316–6329.

Otman Moussaoui and Yacine El Younnoussi. 2023. Pre-training two bert-like models for moroccan dialect: Morroberta and morrbert. In *MENDEL*, volume 29, pages 55–61.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

Faisal Qarah. 2024a. Egybert: A large language model pretrained on egyptian dialect corpora. *arXiv preprint arXiv:2408.03524*.

Faisal Qarah. 2024b. Saudibert: A large language model pretrained on saudi dialect corpora. *arXiv preprint arXiv:2405.06239*.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, et al. 2024. Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect. *arXiv preprint arXiv:2409.17912*.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2023. Understanding calibration for multilingual question answering models. *arXiv preprint arXiv:2311.08669*.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *arXiv preprint arXiv:1808.07674*.

Jihad Zahir. 2022. Iadd: An integrated arabic dialect identification dataset. *Data in Brief*, 40:107777.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. *arXiv preprint arXiv:2106.01494*.