Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus

Helene Bøsei Olsen
University of Oslo
helenbol@ifi.uio.no

Samia Touileb
University of Bergen
samia.touileb@uib.no

Erik Velldal
University of Oslo
erikve@ifi.uio.no

Abstract

This paper provides a systematic analysis and comparison of the performance of state-of-the-art models on the task of fine-grained Arabic dialect identification using the MADAR parallel corpus. We test approaches based on pre-trained transformer language models in addition to Naive Bayes models with a rich set of various features. Through a comprehensive data- and error analysis, we provide valuable insights into the strengths and weaknesses of both approaches. We discuss which dialects are more challenging to differentiate, and identify potential sources of errors. Our analysis reveals an important problem with identical sentences across dialect classes in the test set of the MADAR-26 corpus, which may confuse any classifier. We also show that none of the tested approaches captures the subtle distinctions between closely related dialects.

1 Introduction

Dialect identification (DID) is a task in natural language processing (NLP) aiming to automatically identify a dialect within a pre-determined language. Because dialectal differences tend to be subtle, identifying dialects is considered a more difficult task than language identification (Etman and Beex, 2015). Arabic dialects are considered particularly challenging due to their high level of ambiguity, lack of standardisation, and rich morphology (Diab and Habash, 2007). Most NLP development has focused on Modern Standard Arabic (MSA), the formal and standardised version of Arabic. However, these tools are not always transferable to dialectal Arabic, as dialects differ from each other and MSA in terms of lexicon, phonology, orthography, and morphology (Habash, 2010). A prominent resource for Arabic DID is the MADAR parallel corpus (Bouamor et al., 2018), targeting dialects on the city-level. MADAR has been established as an important corpus for the task, serving as a benchmark for multi-task learning (Seelawi et al., 2021), as well as a Shared Task corpus (Bouamor et al., 2019), and as a subject of independent research (Baimukan et al., 2022). Despite several attempts to develop models using deep neural networks (Lippincott et al., 2019; de Francony et al., 2019) and pre-trained Transformer-based language models (Inoue et al., 2021), the current state-of-the-art approach remains a statistical machine learning model with surface-level feature representation, specifically the Multinomial Naive Bayes (MNB) model introduced by Salameh et al. (2018).

The lack of progress on the task, along with the inability of BERT models to surpass the MNB model, gives rise to several questions that have not yet been thoroughly explored, and on which we focus in the current work. Firstly, do BERT models make the same mistakes as the state-of-the-art MNB model on the dialect identification task? While Salameh et al. (2018) have documented the performance of the MNB model on individual dialects and highlighted the Muscat dialect as the most challenging for the model, there is limited research exploring the misclassifications generated by BERT models. Secondly, if the models make different errors, are these errors centred around the same dialect pairs? Thirdly, we explore if a detailed analysis of the misclassified sentences by both the BERT models and the MNB model can provide deeper insights into the challenges of the task on MADAR-26.

This paper summarises the findings from a comprehensive project on error-analysis on the MADAR parallel corpus conducted by Olsen (2023). We release the code for all experiments and analysis on GitHub.[1]

2 Previous work

Several efforts have focused on building tools and resources to identify Arabic dialects. However,

[1] https://github.com/helenebol/Arabic-dialect-identification

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 370–384
December 7, 2023

the field suffers from fragmented and independent works on different corpora that vary in terms of granularity, size and domain, making it challenging to track the progress of the solutions. Early work focused on binary dialect classification by discriminating one dialect from MSA (Elfardy and Diab, 2013; Tillmann et al., 2014), as well as identifying Arabic dialects at both a region-level (Zaidan and Callison-Burch, 2011, 2014; Elaraby and Abdul-Mageed, 2018; Cotterell and Callison-Burch, 2014) and a country-level (Talafha et al., 2020; Abdelali et al., 2021; AlKhamissi et al., 2021).

In recent years, more efforts have targeted Arabic DID on a more fine-grained level, particularly through shared tasks. The Nuanced Arabic Dialect Identification Shared Tasks (NADI) (Abdul-Mageed et al., 2020, 2021b, 2022) include sub-tasks on country- and province-level on user-generated tweets. Several corpora of written Arabic dialects comprise tweets (Abdelali et al., 2021; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018), others consist of user commentaries (Zaidan and Callison-Burch, 2011), or manually translated sentences (Bouamor et al., 2018, 2014).

For the NADI shared tasks (Abdul-Mageed et al., 2020, 2021b, 2022), all the top performing systems used transformer-based language models pre-trained on dialectal Arabic. However, these models yielded unsatisfactory results and multiple factors were identified, including imbalanced class distribution (AlShenaifi and Azmi, 2020), a significant presence of MSA content in the training data (Touileb, 2020), and the inherent challenges associated with distinguishing between Arabic dialects.

Within the MADAR shared task (Bouamor et al., 2019), the top five performing systems demonstrate that ensemble techniques, n-gram-based features, and traditional machine learning approaches, such as MNB or Support Vector Machines (SVMs), yield the highest levels of performance. While the MADAR corpus proved to be too small for deep learning architectures (Lippincott et al., 2019), the transfer learning ability of BERT-based language models, pre-trained on dialectal Arabic, has shown promising results (Seelawi et al., 2021; Inoue et al., 2021). However, the MNB model introduced by Salameh et al. (2018) is still state-of-the-art with an overall accuracy of 67.9%.

	MADAR-26		MADAR-6	
Sentences	Per dialect	Total	Per dialect	Total
Train	1600	41600	9000	54000
Dev	200	5200	1000	6000
Test	200	5200	-	-

Table 1: Number of sentences per dialect and per split in the MADAR-26 and MADAR-6 corpora.

	Avg.	Min	Max
Tokens	11265.42 (±619)	Basra	MSA
Sent length	5.61 (±0.3)	Basra	MSA
Vocabulary (types)	3273.61 (±204)	Doha	MSA

Table 2: Data statistics for MADAR-26, showing the average number of tokens, average sentence length, and vocabulary size (number of types) across dialects without punctuation. Min and max denote the dialect with the lowest and highest values for each statistic. The numbers in parentheses denote variance.

3 The MADAR corpus

The MADAR corpus is a collection of parallel sentences in the travel domain (Bouamor et al., 2018). The resource contains two corpora with non-overlapping sentences: (1) MADAR-26: covering 25 cities and MSA, and where each dialect is represented with 2000 sentences. (2) MADAR-6: covering the five selected cities Doha, Beirut, Rabat, Cairo, and Tunis, in addition to MSA, each with 12000 sentences. We use the training, development, and test splits from the MADAR shared task 1 (Bouamor et al., 2019) shown in Table 1. As can be seen, all classes are perfectly balanced for each set. In our models, we use MADAR-26 for both training and evaluation, while MADAR-6 is included in the training data of the state-of-the-art system presented by Salameh et al. (2018).

Throughout this work, we define tokens based on white space using the simple word tokeniser from CAMeL Tools[2] to split the sentences. Additionally, all punctuation are removed.

3.1 Corpus statistics

The MADAR-26 training data primarily consists of short sentences, with an average length of 5.6 tokens, as seen in Table 2. Short sentences can be challenging for DID, as they may not encompass enough information to capture the nuances of dialectal variations (Malmasi et al., 2016). The data

[2] https://github.com/CAMeL-Lab/camel_tools/tree/master/camel_tools/tokenizers

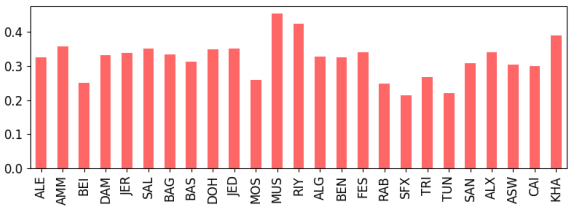


Figure 1: Vocabulary overlap between MSA and the dialects in the training data

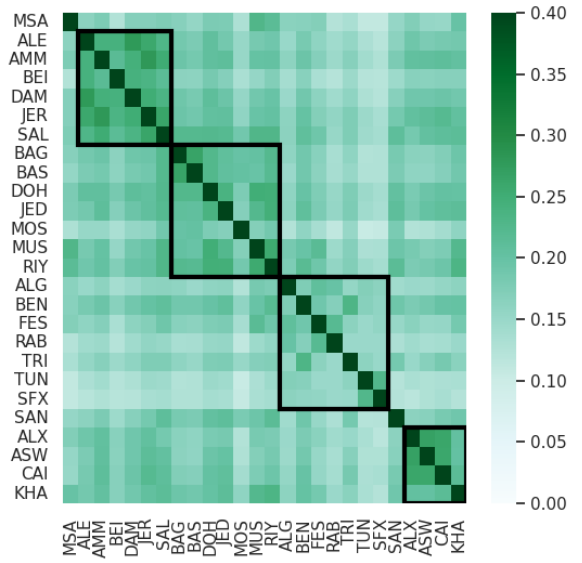


Figure 2: Heatmap of the lexical similarity computed with Overlap coefficient between the dialects in the MADAR-26 training data. The black boarders outline the geographical regions. For a clearer view of nuances, the heatmap threshold is set to 0.40, while some dialects might have a higher score.

also shows variations in vocabulary size across the dialects, where MSA consistently has the largest values. In contrast, the Doha dialect exhibits the smallest vocabulary, and the Basra dialect has the shortest sentences and the lowest number of tokens.

3.2 Lexical overlap

We here explore the degree of lexical overlap between the dialects by analysing the number of common tokens between them. We follow the work of Bouamor et al. (2018) and use the Overlap Coefficient (OC) to measure the degree of similarity between two sets of texts A and B, ranging from 0 (no overlap) to 1 (complete overlap).[3]

Lexical overlap with MSA The diglossic situation of Arabic puts MSA in a distinctive position concerning lexical overlap, given its presence in the daily language use of all dialect users. The source sentences for translation in the MADAR corpus were provided in English and French to minimise the bias of MSA (Bouamor et al., 2018).

As demonstrated in Figure 1, the OC between MSA and each dialect varies and ranges from 0.2 for Sfax and Tunis to over 0.4 for Muscat and Riyadh. While some of the overlap might stem from various bias factors in the translation process, it is also plausible that some of the overlapping vocabulary consists of function words and nouns that are shared with MSA. Rather than considering the vocabulary overlap as noise, it should be a factor when interpreting the results of DID. More specifically, this overlap might suggest that distinguishing MSA from Muscat or Riyadh might be more challenging than from Tunis or Sfax.

Lexical overlap between dialects By calculating the OC for every pair of dialects in the training data, we find the average pairwise similarity between them to be 0.35 with a standard deviation of 0.07. The OC across all dialect pairs can be seen in Figure 2, where the black borders outline the geographical regions Levant, Gulf, Maghreb, and Nile basin. The highest levels of overlap is between dialects within the same geographical region. The most prominent example is the Levant region, where most dialects have a high OC with each other. We find a similar pattern in the Gulf region, except for the Mosul dialect. For the Maghreb region, the overlap is less significant across the region, but higher for dialects within the same country (*e.g.* Rabat and Fes). Interestingly, there seems to be a high level of overlap between the Egyptian city dialects, Cairo, Aswan, and Alexandra, while Khartoum (Sudan), displays a slightly lower overlap. Sanaa is not included in any region, while it seems to have similar vocabulary to both the dialects in the Nile Basin region and the Gulf.

Tunis and Sfax city dialects exhibit relatively low levels of lexical overlap with dialects outside Tunisia, indicating a more distinct vocabulary. A similar pattern is noticeable in the Moroccan city dialects of Fes and Rabat. With an average vocabulary of approximately 3000 tokens, several dialects have fewer than 400 tokens that do not overlap

[3]Defined as: $OC(A,B) = \frac{|A \cap B|}{min(|A|,|B|)}$

with other dialects. This highlights the lack of clear class boundaries and emphasises the challenge of automatically identifying the dialects.

There is a more nuanced distribution of the linguistic features and characteristics of the dialects. There are morphological and lexical differences between the dialects, as well as significant vocabulary similarity within each region. More details about this lexical analysis can be found in Appendix A.1.

4 Models

We here describe our experimental set-up and the tested models.

4.1 Pre-trained Transformer language models

We evaluate three BERT models pre-trained on dialectal Arabic, AraBERTv0.2-Twitter[4](Antoun et al., 2020), MARBERTv2[5](Abdul-Mageed et al., 2021a), and CAMeLBERT-Mix[6] (Inoue et al., 2021). We will refer to them as AraBERT, CAMeLBERT, and MARBERT respectively. There exist several BERT models pre-trained on Arabic dialects. However, to the best of our knowledge, AraBERT and MARBERT have not yet been evaluated on MADAR-26. CAMeLBERT model is considered one of the top-performing models on the task (Inoue et al., 2021), and is therefore included as a baseline.

While all are based on the BERT architecture (Devlin et al., 2019), specifically the "base" version, they differ in terms of their pre-training data, model size, and vocabulary (see details in Table 8 in Appendix A.2). Notably, AraBERT is the smallest model in terms of number of tokens (8.6B), compared to MARBERT (29B) and CAMeLBERT(17.3B). All models are pre-trained on various MSA and dialectal Arabic sources, all including tweets. However, CAMeLBERT has the most diverse dialectal pre-training data, including the MADAR parallel corpus (Inoue et al., 2021).

Experimental setup and data We follow the ALUE benchmark model (Seelawi et al., 2021): the pre-trained BERT encoder takes an Arabic sentence as input and generates contextualised embeddings. The CLS classification token is extracted from the final layer of BERT, passed through a linear layer, before a softmax function computes the predicted classes. Details on implementation and hyperparameter tuning are described in Appendix A.3. We fine-tune all models on the MADAR-26 training set, using the same data splits as supplied for the MADAR shared task (Bouamor et al., 2019). We perform dediacritisation on the data in alignment with the pre-training of the BERT models. Similarly to previous experiments on this corpus (Inoue et al., 2021), no additional pre-processing is done.

[4]https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter

[5]https://huggingface.co/UBC-NLP/MARBERTv2

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

	Accuracy	F_1
CAMeLBERT	63.25 (±0.65)	62.69 (±0.06)
MARBERT	62.36 (±0.05)	61.76 (±0.72)
AraBERT	**65.19** (±1.60)	**65.64** (±0.51)

Table 3: Average results for BERT models after five runs on MADAR-26 development set. Corresponding standard deviation in parentheses. Numbers in bold indicate best results.

4.2 Multinomial Naive Bayes

For the MNB model, we use the CAMeL-tools (Obeid et al., 2020) implementation of Salameh et al. (2018). The system consists of two models, a main MNB model trained on MADAR-26 and a supporting MNB model trained on MADAR-6. This latter classifies each sentence into a dialect from MADAR-6, and then used as a feature in the main model. As the six dialects in MADAR-6 are from different regions in MADAR-26, we consider the supporting classifier a regional classifier.

Both the main and supporting MNB models use similar feature types but from different corpora. The supporting model uses TF-IDF weighted word and character n-grams, in addition to probability scores from 6 dialectal n-gram language models trained on MADAR-6. The main MNB model uses the same feature types but with probability scores from 26 n-gram LMs trained on MADAR-26. Additionally, it takes regional probability scores from the supporting MNB model's predictions.

Note that we also trained a logistic regression model with the same features, but due to its subpar results, we are not including it in the discussion.

4.3 Evaluation

Due to the perfectly balanced classes, we report the overall performance of all models using macro average F1. We also report precision, recall, and F1 for each individual dialect and the average for

Region	Dialect	AraBERT (%)	MNB (%)	Diff
	MSA	75.4	72.3	+3.1
Levant	ALE	63.6	63.8	-0.2
	AMM	52.1	65.5	-13.4
	BEI	64.1	68.8	-4.7
	DAM	50.0	63.4	-13.4
	JER	55.5	61.0	-5.5
	SAL	51.2	56.4	-5.2
Gulf	BAG	64.0	65.9	-1.9
	BAS	60.9	66.3	-5.4
	DOH	61.4	68.9	-7.5
	JED	53.7	59.4	-5.7
	MOS	78.9	86.8	-7.9
	MUS	51.0	48.8	+2.2
	RIY	53.9	57.6	-3.7
Maghreb	ALG	72.7	81.5	-9.0
	BEN	60.5	68.5	-8.0
	FES	65.2	71.3	-6.1
	RAB	67.5	72.3	-4.8
	SFX	69.9	72.9	-3.0
	TRI	70.9	80.0	-9.1
	TUN	64.8	72.3	-7.5
	SAN	68.1	75.0	-6.9
Nile Basin	ALX	74.0	75.9	-1.9
	ASW	63.2	63.8	-0.6
	CAI	53.0	55.8	-2.8
	KHA	66.1	72.5	-6.4
Total		**63.4**	**67.3**	**-3.9**

Table 4: F_1 scores of the AraBERT and MNB models on MADAR-26 test set. Highest scores for each model are in blue, and lowest in red. Green shows where AraBERT has a higher score than MNB. The overall performance of the models is displayed in the final row and marked in bold.

each region for the best-performing model.

Based on the development results in Table 3, we find that AraBERT outperforms the other models, with an average accuracy of 65.19% and a macro-average F1 of 65.64%. These results are interesting, considering AraBERT's smaller pre-training data size compared to the other models. It is also noteworthy that even though CAMeLBERT has MADAR-26 included in the pre-training data, it is outperformed by AraBERT on the development data. We speculate that these outcomes stem from effective filtering and curation of the pre-training data of AraBERT. We inspect the results on the test data in more detail next.

5 Test results

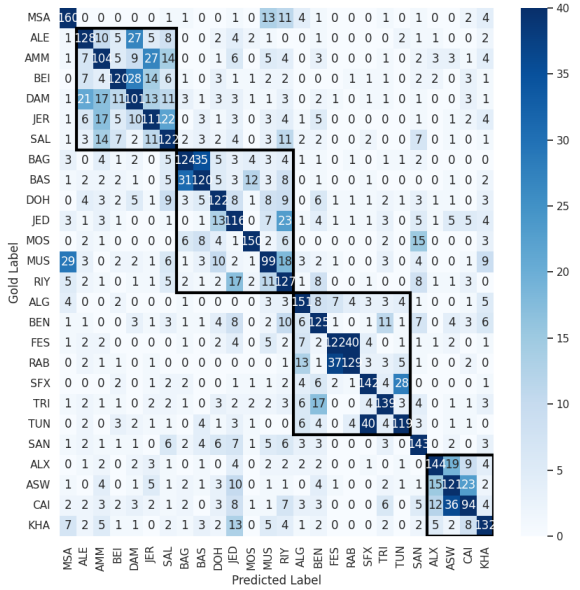
We compare the performance of both selected models, MNB and AraBERT, in terms of F_1 score for each individual dialect in Table 4. The results reveal a notable difference in their overall and individual dialect classification performance, with the MNB model outperforming AraBERT on the majority of dialects. As previously suggested, the results clearly show that the AraBERT model outperforms the MNB model on MSA and the Muscat dialect, with a difference of 3.1 and 2.2 pp, respectively. Interestingly, both models have the lowest performance on the Muscat dialect. We can also observe close performance on the Aleppo and Aswan dialects, while the most significant difference in performance is for the Amman and the Damascus dialects, where the MNB model outperforms the AraBERT model with 13.4 percentage points for both dialects. Due to the high lexical overlap between MSA and Muscat together with the high degree of MSA content in the pre-training data of the AraBERT model, it is likely that the AraBERT model is better at detecting MSA, and thereby not confusing the two dialects to the same degree as the MNB model. More details about the best classifications per dialect and model can be found in Table 11 in Appendix A.4.

6 Error analysis

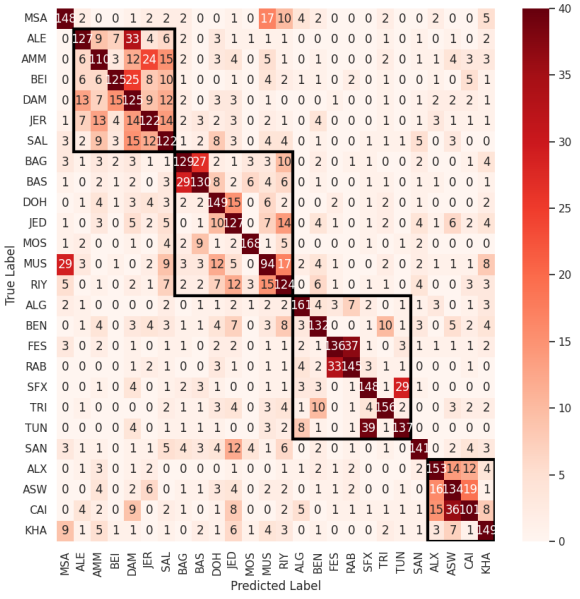
We here provide a systematic analysis of the errors made by the different models.

6.1 Misclassification patterns

Analysing the confusion matrices in Figure 3, which visualises the two models' predictions, reveals distinct similarities in their misclassification patterns. (i) Most errors occur between city dialects from the same geographical regions (outlined with the black borders). For example, in the Levant region, Beirut is misclassified as Damascus, Amman, Aleppo, Jerusalem, and Salt by both models. We can also observe a high density within the dialects in the Nile basin region, while for the Maghreb and Gulf region, the overlap is more spread out. (ii) Both models' most frequent errors occur between city dialects from the same country. Notable examples are the two Moroccan city dialects Fes and Rabat, the Egyptian dialects Aswan, Cairo, and Alexandria, and the Iraqi dialects, Baghdad, and Basra. (iii) When considering the errors occurring outside the regional borders, we find that a significant proportion is associated with Arabic variants that are not attributed to any specific region, namely MSA and Sanaa. Among these outliers, the highest frequency of confusion is between the Muscat



(a) AraBERT model



(b) MNB model

Figure 3: Confusion matrices of (a) AraBERT and (b) MNB predcitions on the MADAR-26 test set.

dialect and MSA for both models.

In comparison to Figure 2 (Section 3.2), illustrating the lexical similarity between the dialects, we can observe some similarities. There is a high lexical similarity between all dialects in the Levant and Nile basin regions. Additionally, errors between dialects from the same country can be inferred from the patterns identified in Figure 2 and the misclassification pattern of MSA. But a few exceptions exist, such as the Mosul dialect having a high lexical similarity with Sanaa, which is only reflected in AraBERT's predictions. Another example is the Sanaa dialect, which displays high lexical similarity with both Doha and Jeddah, which is again evident in AraBERT's predictions, while the MNB model tends to confuse Sanaa with Jeddah.

Despite these exceptions, the high number of similarity patterns implies a positive correlation between high lexical similarity and the misclassifications for both models, but is not a complete explanation. By comparing the distribution of misclassification for each model, we discover that both models struggle with identifying the Muscat, Cairo, and Amman dialects. AraBERT exhibits more errors than the MNB model, particularly concerning the Damascus, Mosul, Tunis, and Doha dialects. However, the MNB model has a higher frequency of misclassifying the MSA, Aleppo, Sanaa, and Riyadh dialects when compared to AraBERT. Furthermore, we find that over 60% of the test data of both the Muscat and Cairo dialects is misclassified by either one or both models.

6.2 Subcategories of misclassified sentences

For our analysis, we are comparing the misclassifications made by the two models on a sentence level, as they can provide a more nuanced understanding of the performance and the difficulty of classifying certain dialects. We will base our error analysis on six categories of misclassifications that provide different insights: (1) Union of misclassification: includes all sentences misclassified by either one of the models or both. (2) Intersecting misclassification (INT): includes the sentences misclassified by both models. This subgroup is partitioned into two subcategories: (i) sentences in which both models have predicted the same incorrect dialect (INT-S), (ii) sentences in which the models have predicted two different dialects (INT-D). (3) Unique-AraBERT and (4) Unique-MNB: sentences misclassified by one model but correctly classified by the other.

The number of sentences for each category (Table 5), reveals interesting insights into the relative difficulty of the task. For example, out of the total of 1227 sentences misclassified by both models, they predicted the same incorrect dialect for 511 of them, while for 716 of the sentences, the models predicted different labels. While the models differ in their respective classification errors for a significant number of sentences, the number of intersecting misclassifications suggests that both models have a similar weakness in predicting a large sub-

	# Sentences	Avg.length
Total test set	5200	5.6 (±2.9)
Union	2415	5.0 (±2.5)
INT-S	511	4.8 (±2.5)
INT-D	716	4.5 (±2.3)
Unique-AraBERT	708	5.6 (±2.8)
Unique-MNB	480	5.5 (±2.9)

Table 5: Overview of number of test sentences and average sentence length for the different categories of misclassification. INT-S and INT-D refer to sentences wrongly classified by both models, where S and D denote whether both models made the same or different predictions. The Unique-AraBERT are the sentences correctly classified by MNB but misclassified by AraBERT, and vice versa for the Unique-MNB category. Union refers to all misclassified sentences regardless of model.

set of the corpus. The high number of sentences in the INT-S category implies that there might be patterns or linguistic features that present challenges for both models, revealing areas where the models have the most difficulty distinguishing between dialects. The INT-D sentences might present insight into particular challenging sentences, as neither model could predict the correct sentence.

6.3 Most frequently confused pair of dialects

We also provide insights into which dialect combinations are most frequently confused. We report on occurrences where a pair of dialects appear together, whether the dialect is a gold or a predicted label, for the same sentence.[7] The two Moroccan city dialects Rabat and Fes are the most frequently confused pair for all categories, except for the INT-D category. In this category, the models' misclassifications are less consistent, leading to less frequent occurrences of dialect pairs.

6.4 Potential sources of error

Table 5 shows the sentence length for each subcategory of misclassified sentences and the total test set. The test data has a similar average length to the training data but with greater variance, and even includes sentences with only one or two tokens, such as the Tripoli sentence فكرة حلوه (*Nice idea*). This may challenge classification, particularly when the tokens are shared among multiple dialects. The shortest sentences are found in the INT-D category, followed by IND-S, which both models misclassified. Interestingly, the unique misclassifications for each model consist, on average, of more tokens compared to the test set average. This suggests that the shorter sentences pose a shared challenge, while the unique misclassified sentences exhibit other challenges particular to each model.

Lexical overlap, the overlap in tokens between two bodies of texts, provides an indication of the extent to which a sentence is a subset of a given dialect's training data. It can also assess the degree to which a misclassified sentence represents the dialect as it appears in the training data. The box plot in Figure 4 illustrates the distribution of the overlap coefficient between sentences in the subcategories of the gold dialects in Figure (a) and between the predicted dialects in Figure (b). The first box in both figures represents the OC between the full training data and the gold dialects vocabulary for comparison purposes.

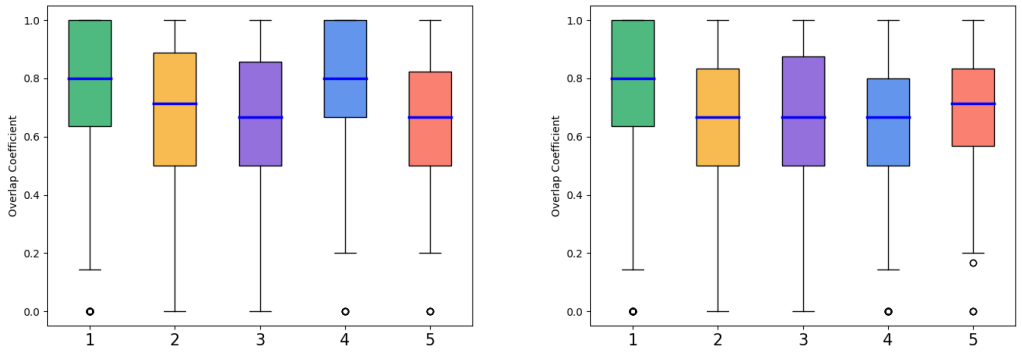
There are three notable observations. Firstly, the OC between the sentences in the test set for both the gold and predicted dialects tends to be high, with an average OC of over 0.5 for all categories in both figures. This trend may imply that certain sentences exhibit a significant vocabulary overlap between multiple dialects, leading to confusion for both models. Secondly, Figure (a) indicates that there are instances in the test data with an OC of 0.0 with the gold dialects, which can also be observed in the OC between the sentences and the predicted dialects in Figure (b), suggesting that lack of vocabulary overlap may be contributing to errors in some cases. Thirdly, box 4, representing the sentences misclassified only by AraBERT, has a higher median OC for the gold dialects compared to the other categories of misclassified sentences in Figure (a). However, in Figure (b), the median for box 4 is lower and more aligned with the other categories. These findings suggest that the AraBERT model tends to prioritise features other than lexical overlap when making predictions.

6.5 Manual example-level analysis

Due to the lack of morphological disambiguators covering all the dialects or regions in MADAR-26, we rely on manual example-level analysis.[8] As part of the comprehensive analysis conducted in Olsen

[7] The top five confused dialect pair for each category is reported in Table 13 in the Appendix.

[8] Since the objective here is to identify sources of misclassification, we will consider the sentences in their original form as input to the models. Consequently, the sentences lack vocalisation, and when analysing specific example sentences, we transcribe them letter-by-letter rather than supplementing the missing characters.



(a) OC with the gold dialects

(b) OC with the predicted dialects

Figure 4: Overlap coefficient (OC) between the test data and the vocabulary in the training data. Figure (a) shows the OC between the sentences and the gold dialects, while Figure (b) shows the OC between the sentences and the predicted dialects. Each box represents the OC for: Box 1: The test data and the gold labels. Box 2: The union of misclassifications of both models. Box 3: The intersection of misclassifications of both models. Box 4: The misclassifications unique to AraBERT. Box 5: The misclassifications unique to MNB.

(2023), the following examples are drawn from the set of 19 cases. We identified various characteristics that present challenges for dialect identification. For instance, there are sentences without dialect-specific features across all categories, as illustrated in Example 1.

(1) وين اتعلم السيد جونز اليابانية ؟
wiin it3lm al+siid jwnz al+iaabaania?
Where did mr. Jones learn Japanese?

In this instance, a sentence from Jerusalem is predicted as Riyadh by AraBERT and as Basra by MNB. Moreover, this sentence contains nouns typically unaffected by dialectal variation, like السيد جونز (*mr. Jones*) and اليابانية (*Japanese*).

For the second example, a Fes sentence is correctly predicted by the MNB model, but confused as Rabat by AraBERT. Rabat and Fes are the most frequently confused dialects, which might be explained by the lack of overlap with dialects outside Morocco (Figure 2), along with the prevalence of linguistic features only observed in the Moroccan dialects training data, such as ديال from Example 2. While these features are distinct enough to exclude the possibility of other dialects, they may not be sufficient to accurately distinguish between closely related dialects such as those from Rabat and Fes.

(2) خليوني ننعس حتا للتاسع ديال الصباح
khliwn+i nn3s h.taa l+ltaas3 diiaal al+s.baah.
Let me sleep until nine in the morning

6.6 Identical sentences in the test data

The MADAR corpus is stated to be created through manual independent translation of sentences in different dialects. However, we identify multiple occurrences of identical sentences in the test data labelled with different dialects. We will refer to these as *duplicates*. There is a total of 522 of duplicate sentences in the test data. 398 such sentences were misclassified by AraBERT, and 393 were misclassified by the MNB model.

Some entries have up to 11 duplicates labelled with different dialects from different regions. An example is the sentence برا, in English *Outside*, which has gold labels from The Levant, Gulf, and Maghreb regions. Most frequent duplicates are very short, some consisting of only one token. The total set of duplicate sentences has an average length of 3.50 tokens with a standard deviation of 1.55, which might explain the high number of identical sentences across multiple regions.

The distribution of the duplicates is skewed, with the highest frequencies among the Levant dialects Jerusalem, Salt, and Damascus, with over 30 sentences each. At the same time, MSA, Mosul, Algiers, Rabat, Sfax, and Sanaa have less than ten each. Furthermore, it appears like dialects with high lexical overlap (see Section 3), have similar amounts of duplicate sentences. See Figure 5 in the Appendix for the distribution across dialects.

Task formulation Because the task of Arabic DID is formulated as a multi-class classification task, many of the sentences in the test data are impossible to identify correctly since they can belong to multiple dialects. The limitations of this task formulation have already been demonstrated (Goutte et al., 2016; Zampieri et al., 2023), suggesting that

	Original	Dedpulicated
Size	5200	4870
Avg. sentences per class	200 (±0)	187.3 (±6.2)
Smallest class	–	Jerusalem (174)
Largest class	–	Sfax (198)

Table 6: Deduplicated MADAR-26 test set compared to the original test set with smallest and largest class.

unless a text belongs to precisely one dialect, the classification task should be approached as a multi-label classification task, rather than a multi-class one (Bernier-colborne et al., 2023).

Deduplication of test data We identify all instances of duplicates and remove them, with only one random instance retained in the test set. The resulting deduplicated test set is presented in Table 6, and consists of 4870 sentences. The result is an imbalanced test set, but, an argument can be made that duplicate sentences in the original test data already imbalanced the test set.

Model evaluation on deduplicated data We evaluate the MNB model on the deduplicated test set and achieve a macro-average F1 score of 70.25%. Compared to the performance on the original test set, evaluation without duplicate sentences across classes results in an increase in performance of 2.95 pp. The presence of duplicate sentences in the data can be viewed as a reflection of natural occurring language use, particularly in the case of short text, where phrases and expressions may be identical across various dialects. Therefore, removing identical sentences may introduce bias in the evaluation process, as it would not reflect the natural occurrence of such duplicates and could lead to an overestimation of a model's performance.

7 Conclusion and future work

This paper investigates the challenging task of fine-grained dialect identification, focusing on the MADAR-26 corpus. By fine-tuning three BERT models pre-trained on dialectal Arabic, we demonstrated that the multinomial naive bayes model introduced by Salameh et al. (2018) remains the state-of-the-art model on this data. However, we identified 480 test sentences that were correctly classified by the best performing BERT model, but were misclassified by the MNB model. A comprehensive error analysis revealed the BERT model exhibits superior performance in predicting sentences in Muscat dialect and MSA, which may be attributed to the amount of MSA content in the pre-training data of the BERT model. We also show that some of the challenges of the task can be attributed to dataset limitations. Particularly the fact that 10% of the sentences in the test set are identical to one or more parallel sentences in the same set but with different labels.

Our analysis of different error types confirms that the MNB and BERT-based model often make different mistakes, but also that a subset of the test data is challenging for both. Notably, we found that the Moroccan city dialects Rabat and Fes are the most confused dialect pair, and show how neither approach is able to capture the subtle distinctions between some of the closely related dialects. Although dataset limitations, such as non-Arabic proper nouns, short sentences without dialect-specific features, and identical sentences across classes, account for some of these errors, the unique errors generated by each model provide evidence that certain sentences can be correctly classified by one model, but not the other. These findings underscore the need to examine model performance beyond simple metric comparison in order to identify new strategies for enhancing Arabic dialect identification.

In the future, we would like to address the formulation of the task, by transforming it into a multi-label classification problem. Instead of simply removing the duplicate sentences from the data, we can combine the labels of duplicate and nearly-duplicate text, converting the single-label dataset into a multi-label dialect classification format.

Another avenue for future research is to evaluate models trained or fine-tuned on MADAR-26 on user-generated data. Due to a lack of annotated data matching the city-levels of MADAR, evaluation on data outside the travel domain has up until recently not been possible. However, the hierarchical mapping schema proposed by Baimukan et al. (2022) can be leveraged for datasets with comparable or more detailed annotations. More specifically, we want to evaluate the performance of the models on the NADI dataset (Abdul-Mageed et al., 2020) by mapping tweets at the province-level to the city-level.

Limitations

Given the scope of this work, we did not conduct an extensive exploration of design choices for the various models, or dedicate considerable time to hyperparameter optimisation and experimentation of the selected models. Nevertheless, we acknowledge that a more rigorous pursuit of hyperparameter tuning may potentially produce different results.

Despite evaluating the Transformer-based models using five different seeds, our error analysis relies solely on the outcomes of a single run. Although the AraBERT model displayed a small degree of instability during the development phase, some of the outcomes used in the error analysis may have varied if a different seed was used. However, due to the extensive nature of the analysis, incorporating outcomes from multiple runs was not a practical option. Therefore, our findings should be considered indicative rather than definitive.

Moreover, the error analysis focused solely on the test set without comparing misclassified and correctly predicted sentences, and thereby limiting our ability to pinpoint the precise factors behind misclassifications. Instead, it offers insights into misclassification categories and variations between types, as well as between the two models.

Because of the wide coverage of the MADAR-26 corpus, some of the dialects in our error analysis are outside our expertise. To mitigate this limitation, we employed the newly publicly released MADAR lexicon (Bouamor et al., 2018) and other resources to aid in analysing these languages. However, inaccuracies may still exist.

Finally, due to the lack of morphological analysers covering all the dialects in MADAR-26, we performed analysis on token-level, where a token is defined by whitespace. This is not optimal for Arabic, as this approach may result in the loss of information conveyed by clitics.

Acknowledgements

The computations were performed on resources provided through Sigma2 – the national research infrastructure provider for High-Performance Computing and large-scale data storage in Norway.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. Qadi: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, page 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. 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)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Badr AlKhamissi, Mohamed Gabr, Muhammad El-Nokrashy, and Khaled Essam. 2021. Adapting MARBERT for improved Arabic dialect identification: Submission to the NADI 2021 shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 260–264, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nouf AlShenaifi and Aqil Azmi. 2020. Faheem at NADI shared task: Identifying the dialect of Arabic tweet. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 282–287, Barcelona, Spain (Online). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic lan-

guage understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

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, Marseille, France. European Language Resources Association.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Florence, Italy. Association for Computational Linguistics.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gael de Francony, Victor Guichard, Praveen Joshi, Haithem Afli, and Abdessalam Bouchekif. 2019. Hierarchical deep learning for arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, page 249–253, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mona Diab and Nizar Habash. 2007. Arabic dialect processing tutorial. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts*, pages 5–6, Rochester, New York. Association for Computational Linguistics.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for Arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Sofia, Bulgaria. Association for Computational Linguistics.

Asma Etman and AA Louis Beex. 2015. Language and dialect identification: A survey. In *2015 SAI intelligent systems conference (IntelliSys)*, pages 220–231. IEEE.

Abbas Ghaddar, Yimeng Wu, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Xinyu Duan, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Phillippe Langlais. 2022. Revisiting pre-trained language models and their evaluation for Arabic natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3135–3151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Nizar Y. Habash. 2010. *Introduction to Arabic natural language processing*, 1 edition, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers.

Salima Harrat, Karima Meftouh, Mourad Abbas, Khaled-Walid Hidouci, and Kamel Smaïli. 2016. An algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications*, 7:384–396.

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. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Tom Lippincott, Pamela Shapiro, Kevin Duh, and Paul McNamee. 2019. Jhu system description for the madar arabic dialect identification shared task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, page 264–268, Florence, Italy. Association for Computational Linguistics.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2016. Arabic dialect identification using a parallel multidialectal corpus. In *Computational Linguistics*, page 35–53, Singapore. Springer Singapore.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Helene Bøsei Olsen. 2023. Fine-grained arabic dialect identification.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-Grained Arabic Dialect Identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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:2007.05612 [cs].

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. 2014. Improved sentence-level arabic dialect classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, page 110–119, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Samia Touileb. 2020. LTG-ST at NADI shared task 1: Arabic dialect identification using a stacking classifier. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 313–319, Barcelona, Spain (Online). Association for Computational Linguistics.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Omar F. Zaidan and Chris Callison-Burch. 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*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels.

A Appendix

A.1 Morphological and lexical differences between dialects

Region	Dialect	Sentence
	English	We want a table close to the stage
	MSA	نريد مائدة بالقرب من المسرح
Gulf	Muscat	نبغا طاولة بالقرب من المسرح
	Doha	بغينا طاولة يم المسرح
	Riyadh	نبغى طاولة قريبة من المسرح
	Jeddah	نبا طاولة جمب المسرح
	Baghdad	نريد طاولة قريبة على الستيج
	Basra	نريد ميز يم الستيج
	Mosul	نغيد ميز قغيب على المسرج
Gulf of Aden	Sanaa	نشتي طاوله قريب من المنصه
Levant	Aleppo	بدنا طاولة جنب المنصة
	Damascus	بدنا طاولة قريبة عالمنصة
	Beirut	بدنا طاولة حد المسرح
	Amman	بدنا طاولة قريبة من المسرح
	Salt	بدنا طاولة قريب من المسرح
	Jerusalem	بدنا طاولة جنب المسرح
Nile Basin	Cairo	عايز ترابيزه جنب المنصة
	Alexandria	عاوزين ترابيزة قريبة من المسرح
	Aswan	أحنا عايزين طربيزة قريبة من المسرح
	Khartoum	دايرين طربيزة جنب المنصة
Maghreb	Tripoli	نبو طاوله جنب المسرح
	Benghazi	نبو طاولة قريبة من المسرح
	Tunis	نحبو طاولة قريبة م الركح
	Sfax	نحبوا طاولة بجنب الواد
	Algiers	رانا حابين طاولة قريبة من منصة العرض
	Rabat	بغينا طاولة قريبة للمسرح
	Fes	بغينا طبلة قريبة للمسرح

Table 7: A sample of a 26-way parallel sentence extracted from MADAR-26 for the English sentence "*We want a table near the stage.*"

To get a more nuanced understanding of the linguistic features and characteristics of the dialects, we analyse the sentence "*We want a table close to the stage*" for all the dialects, see Table 7, as we believe it highlights many of the morphological and lexical differences between the dialects. For example, the English word *table* is translated into مائدة in MSA, while for Basra and Mosul it is ميز, and طربيزة in Aswan. It is translated into طاولة in multiple city dialects in the Gulf, Levant and Maghreb region. Translating the word *table* into طاولة makes sense for many of the dialects in the Levant and in the Gulf, while for others, this translation choice seems to have been influenced by MSA. For instance, in the Algiers dialect, many Algiers dialect speakers view طاولة as a MSA word and prefer the French-derived term طابلة in their daily communication (Harrat et al., 2016).

When examining sentences regionally, we find significant vocabulary similarity within each region. As an example, in the Levant region, all city dialects translate *We want a table* as بدنا طاولة. This contributes to the complexity of DID at a city-level, particularly in distinguishing between cities in the same geographical area.

A.2 Arabic BERT-based models

	Size	#Tokens	pre-training data
AraBERT	541MB	8.6B	77GB+60M Tweets
MARBERT	654MB	29B	167GB
CAMeLBERT	439MB	17.3B	167GB

Table 8: Configuration for AraBERTv0.2-Twitter (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2021a) and CAMeLBERT-mix (Inoue et al., 2021).

A.3 Implementation details

All the reported experiments are run on the high-performance computing resource Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway, made available by the University of Oslo. For replicability, we do not train the BERT models from scratch, relying instead on pre-trained BERT models downloaded directly from Huggingface.

Hyperparameter tuning For all experiments, we are using maximum length of 128 tokens for input sequences, AdamW optimiser with epsilon at 1e-8, and early stopping to determine the optimal number of epochs. To compute the loss, we use Cross entropy and set dropout to 0.1.

We are not experimenting with different hyperparameters for the CAMeL-BERT model, as previous work has made a thorough effort to explore the optimal combination for the model on the task of DID on MADAR-26 (Inoue et al., 2021; Ghaddar et al., 2022). Additionally, we run each model multiple times with different seeds to capture potential deviations in performance (Devlin et al., 2019).

In the case of AraBERT and MARBERT, we base our hyperparameter grid search on previous experiments on earlier versions of the models, namely AraBERTv0.2 (Antoun et al., 2020) and MARBERTv1 (Abdul-Mageed et al., 2021a), on the task of DID on MADAR-26 (Inoue et al., 2021; Ghaddar et al., 2022). Table 9 presents the results from

the hyperparameter grid search, while Table 10 shows the hyperparameters used for evaluation on the development set for all models.

Model	Batch	Learning rate	
		2e-05	1e-4
MARBERT	32	62.56 (±0.63)	62.69 (±0.06)
	16	62.12 (±1.84)	61.76 (±1.72)
AraBERT	32	62.95(±0.57)	65.64 (±0.51)
	16	63.95(±0.40)	64.76 (±1.65)

Table 9: Average results for AraBERT-Twitter and MARBERTv2 on five seeds testing hyperparameters.

Model	Batch	Lr	Epochs
MARBERT	32	1e-4	6
AraBERT	32	1e-4	8
CAMeLBERT	32	2e-05	3

Table 10: Hyperparameters for AraBERTv0.2-Twitter (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2021a) and CAMeLBERT-mix (Inoue et al., 2021).

A.4 Percentage of correctly classified sentences

Table 11 displays the percentage of correctly classified test data for each dialect by each model, focusing on top five best- and top five worst-performing dialects. This summary demonstrates how the MNB model identifies a larger proportion of the sentences compared to the AraBERT model, both for the top five best and weakest results. The table confirms the models' differences in proficiency on various dialects, the most interesting example being the AraBERT model's high accuracy of 80% in predicting the MSA sentences, which is not among the top five results for the MNB model.

	AraBERT	MNB
1.	MSA (80.0%)	MOS (84.0%)
2.	ALG (75.5%)	ALG (80.5%)
3.	MOS (75.0%)	TRI (78.0%)
4.	ALX (72.0%)	ALX (76.5%)
5.	SAN (71.5%)	DOH (74.5%)
22.	JER (55.5%)	SAL (61.0%)
23.	AMM (52.0%)	JER (61.0%)
24.	DAM (50.5%)	AMM(55.0%)
25.	MUS (49.5%)	CAI (50.5%)
26.	CAI (47.0%)	MUS (47.0%)

Table 11: The five top and bottom dialects based on percentage of sentences predicted correctly by each model.

A.5 Cities covered in the MADAR corpus

In Table 12 we give the full list of all cities covered in the MADAR corpus, as well as the abbreviations of their names used throughout the paper.

Dialect city	Abbr.	Country	Region
Damascus	DAM	Syria	Levant
Aleppo	ALE		
Beirut	**BEI**	Lebanon	
Amman	AMM	Jordan	
Salt	SAL		
Jerusalem	JER	Palestine	
Muscat	MUS	Oman	Gulf
Doha	**DOH**	Qatar	
Riyadh	RIY	KSA	
Jeddah	JED		
Baghdad	BAG	Iraq	
Mosul	MOS		
Basra	BAS		
Sanaa	SAN	Yemen	Gulf of Aden
Tripoli	TRI	Libya	Maghreb
Benghazi	BEN		
Tunis	**TUN**	Tunisia	
Sfax	SFX		
Algiers	ALG	Algeria	
Rabat	**RAB**	Morocco	
Fes	FES		
Cairo	**CAI**	Egypt	Nile basin
Alexandria	ALE		
Aswan	ASW		
Khartoum	KHA	Sudan	

Table 12: The cities covered by MADAR-26 with corresponding country and region as defined by Bouamor et al. (2018). The cities included in MADAR-6 are marked with bold.

A.6 Most frequently confused pair of dialects

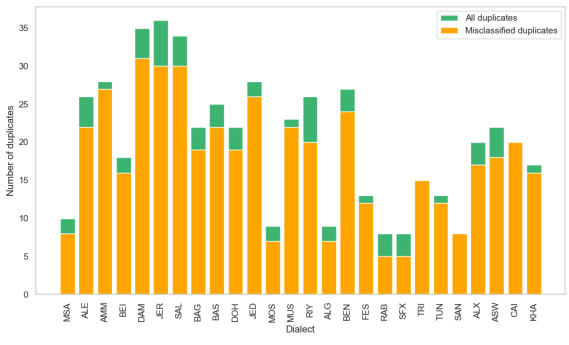


Figure 5: Distribution of duplicate and misclassified duplicate sentences for each dialect in the MADAR-26 test set.

As previously mentioned, the distribution of the duplicates is skewed. As can be seen from Table 5, Jerusalem, Salt, and Damascus (from the Levant region) are the dialects with most duplicates with over 30 sentences each. While MSA, Mo-

Category	Dialect pair	Frequency
Union	RAB, FES	109
	TUN, SFX	100
	BAG, BAS	86
	CAI, ASW	85
	DAM, ALE	61
	MSA, MUS	58
Int-S	RAB, FES	36
	TUN, SFX	30
	DAM, ALE	25
	BAG, BAS	23
	CAI, ASW	22
	DAM, BEI	20
Int-D	RIY, JED	14
	MSA, MUS	13
	BAG, BAS	13
	SAL, AMM	11
	JER, AMM	11
	JER, BEI	11
Unique AraBERT	RAB, FES	39
	TUN, SFX	32
	BAG, BAS	30
	ASW, CAI	30
	JER, AMM	19
Unique MNB	RAB, FES	28
	TUN, SFX	28
	MSA, MUS	26
	BAG, BAS	24
	ASW, CAI	18

Table 13: The five most frequently occurring pairs of dialects in each category. The frequency is based on whether the two dialects occur together, either where d1 is the correct dialect and d2 is the predicted dialect, or where d2 is the correct dialect and d1 is the predicted dialect. Dialect pairs that are not from the same country are marked with bold.

sul, Algiers, Rabat, Sfax, and Sanaa have less than ten each. It is quite clear that having duplicate sentences confuses the models, as the majority of duplicates were actually misclassified.

We report on occurrences where a pair of dialects appear together, either as a gold label or as the predicted label, to inspect which dialect combinations are most frequently confused. The results for each category are presented in Table 13, and show how the most frequently confused dialect pairs are city dialects from the same country. The two Moroccan city dialects Rabat and Fes are the overall most frequently confused dialect in all categories except for the INT-D category. The high frequency between them might be explained by the high lexical overlap in terms of shared tokens in the training data, as reported in Section 3.

The dialect pairs that are not from the same country are highlighted in bold, and they all belong to the Levant region. However, there is one exception - the MSA and Muscat pair, which occur together 58 times. Interestingly, this combination only occurs in the INT-S and the Unique MNB category, in addition to the union of misclassifications, which suggests that the MNB model might contribute more to this confusion than the AraBERT model.

The INT-D category stands out from the others in two ways. Firstly, the frequency of each pair is significantly lower compared to the other categories, suggesting that this subset of misclassifications might have less dialect-specific features. Secondly, it exhibits three dialect pairs that are not located in the same country.