PeerJ Computer Science

# Towards boosting unlabeled text corpora for Arabic dialect identification

Mohammed Abdelmajeed[1], Zheng Jiangbin[1], Murtadha Ahmed[1] and Mohammed Abaker[2]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[2] Department of Computer Science, Applied College, King Khalid University, Muhayil, Asir, Saudi Arabia

## ABSTRACT

Arabic dialect identification (ADI) aims to automatically determine the specific regional dialect of a given Arabic text. State-of-the-art ADI solutions often rely on fine-tuning Arabic-specific pre-trained language models (PLMs). Although effective, these PLMs are predominantly trained on modern standard Arabic (MSA), which limits their performance on dialectal data. Furthermore, the high degree of similarity among Arabic dialects makes it difficult to learn accurate dialect-specific representations without large volumes of labeled data. However, labeling such data is both costly and labor-intensive, particularly for low-resource languages like Arabic. To address these challenges, we propose a self-training neural approach that independently learns *Dialectal Indicators*. Our method leverages unlabeled data to construct a matrix that captures dialectal tokens frequently co-occurring in similar contexts. This matrix provides dialect-specific representations, which are integrated with PLM outputs to enhance ADI performance. We evaluate our approach on multiple ADI and related datasets. Results show that our method significantly improves PLM performance over direct fine-tuning, achieving gains of up to 36.2% in accuracy and 11.52% in macro-F1-score.

**Subjects** Artificial Intelligence, Natural Language and Speech
**Keywords** Arabic dialect identification, Pre-trained language models, Natural language processing, Bidirectional encoder representations from transformers

Submitted 16 December 2024
Accepted 25 July 2025
Published 23 October 2025

Corresponding author
Mohammed Abdelmajeed,
majeedi@mail.nwpu.edu.cn

Academic editor
Davide Chicco

Additional Information and Declarations can be found on page 18

DOI **10.7717/peerj-cs.3127**



## INTRODUCTION

Arabic dialect identification (ADI) aims at determining the specific regional dialect of a given piece of Arabic text. The goal of ADI is to distinguish between the various dialects of Arabic, which can differ significantly in vocabulary, pronunciation, and grammatical structures, despite sharing a common base in modern standard Arabic (MSA) (*Salameh, Bouamor & Habash, 2018*). Consider, for instance, the running example provided in Table 1 the sentence "How do you pronounce the name of this place?", it is represented by a single form in MSA (كيف تنطق اسم هذا المكان؟, kyf tnTq Asm hðA AlmkAn?), is expressed in several forms in Rabat, Cairo, Khartoum and Sana'a.

Notably, there is a substantial shared content between the dialects, contributing to an increased degree of similarity in the contextual embedding space. State-of-the-art solutions are built upon fine-tuning pre-trained language models (PLMs). Earlier approaches attempted to utilize multilingual PLMs like multilingual Bidirectional Encoder Representations from Transformers (mBERT) (*Kenton & Toutanova, 2019*),

**Table 1 Running example.** A running example is presented for the dialects of Rabat, Khartoum, Cairo, and Sana'a, all corresponding to the English phrase, "How do you pronounce the name of this place?".

| Sentence | Transliteration | Label |
|---|---|---|
| كيفاش تنطق اسم لبلاسه هذي؟ | kyfAš tnTq Asm lblAsh hðy? | Rabat |
| كيف تقول اسم المحل ده؟ | Kyf twl Asm AlmkAn dh? | Khartoum |
| ازاي تؤول اسم المكان ده؟ | AzAy tWwl Asm AlmkAn dh? | Cairo |
| كيف تقول اسم هذا المكان؟ | Kyf tqwl Asm hðA AlmkAn? | Sana'a |

XLM-Robustly Optimized BERT Pretraining Approach (RoBERTa) (*Conneau et al., 2020*), and Language-Agnostic BERT Sentence Embedding (LaBSE) (*Feng et al., 2022*) for representing Arabic dialects. However, despite these efforts, the performance of these multilingual models typically lags their monolingual counterparts. This discrepancy primarily stems from smaller, language-specific vocabularies and less comprehensive language-specific datasets (*Antoun, Baly & Hajj, 2020*; *Dadas, Perełkiewicz & Poświata, 2020*; *Malmsten, Börjeson & Haffenden, 2020*; *Virtanen et al., 2019*; *de Vries et al., 2019*). While languages with similar structures and vocabularies may benefit from shared representations (*Conneau et al., 2020*), this advantage does not extend to languages like Arabic. Arabic's unique morphological and syntactic structures have little in common with the frameworks of more abundantly represented Latin-based languages. To address this, various approaches employ fine-tuning Arabic-specific PLMs, including Arabic BERT (AraBERT) (*Abdul-Mageed, Elmadany & Nagoudi, 2021*), and CAMeL (*Inoue et al., 2021*). These models significantly enhance Arabic NLP tasks over multilingual models. However, they are predominantly trained on MSA datasets, potentially limiting their performance on dialectal text. Moreover, as mentioned earlier, the similarities between Arabic dialects make learning accurate dialect-specific representations challenging without a substantial amount of labeled data, which is costly and labor-intensive. In this article, we propose a novel approach aimed at leveraging unlabeled data to learn dialect-specific representations, which we term indicators. Essentially, our approach aims to learn a new matrix where each row corresponds to a unique dialect. Specifically, we attempt to map tokens that often co-occur into distinct semantic *Dialectal Indicators*. The model is trained in an unsupervised manner to reconstruct weighted sentences through a linear combination of these indicators. These vector indicators are then combined with the PLM encoding to model dialect representation during fine-tuning. These indicators serve as semantic cues that facilitate PLM learning of dialectal representations without requiring extensive labeled data.

The main contributions of this article can be summarized in threefold as follows:

- We introduce a self-training approach that leverages unlabeled data to learn *dialect-specific* representations, termed *Dialectal Indicators*.
- We propose incorporating the learned *Dialectal Indicators* along with the PLM representations to accurately model text in response to dialectal variations.
- We conduct an extensive empirical evaluation across multiple benchmark datasets designed for ADI. Our experimental results consistently demonstrate that our solution achieves a new state-of-the-art performance.

The remainder of this article is as follows: 'Related Work' provides an in-depth exploration of related research. In 'Materials and Methods', we present our proposed solution. 'Experiments Setup' outlines the experimental setup. 'Empirical Evaluation' provides an empirical evaluation of our solution's effectiveness. Finally, 'Conclusion' concludes the article.

## RELATED WORK

This section provides a concise overview of PLMs, with a particular focus on those specifically developed for the Arabic language. We then turn to the primary objective of our research: Arabic dialect identification.

### Pre-trained language models

PLMs such as BERT (*Kenton & Toutanova, 2019*), and RoBERTa (*Liu et al., 2019*), trained through self-supervised masking objectives, have transformed NLP. Multilingual variants like mBERT (*Kenton & Toutanova, 2019*), XLM-RoBERTa (*Conneau et al., 2020*), and LaBSE (*Feng et al., 2022*), along with alternative architectures like ALBERT (*Lan et al., 2019*), T5 (*Raffel et al., 2020*), its multilingual variant mT5 (*Xue et al., 2021*), as well as GPT-3 (*Brown et al., 2020*), Large Language Model Meta AI (LLaMA) (*Touvron et al., 2023*), Pathways Language Model (PaLM) (*Chowdhery et al., 2023*), GPT-4 (*Achiam et al., 2023*), and RoFormer (*Su et al., 2024*). In addition to English-based PLMs, various models have been created for non-English languages. For instance, Bertje has been developed for Dutch (*de Vries et al., 2019*), while CamemBERT (*Martin et al., 2020*), and FlauBERT (*Le et al., 2020*), serve the French language. Vietnamese is supported by PhoBERT (*Nguyen & Tuan Nguyen, 2020*), and Finnish models have been created by *Virtanen et al. (2019)*, Polish by *Dadas, Perełkiewicz & Poświata (2020)*, and Swedish by *Malmsten, Börjeson & Haffenden (2020)*. Additionally, *Pyysalo et al. (2021)* have produced monolingual language models (LMs) trained on Wikipedia data for 42 languages.

### Arabic PLMs

Arabic encompasses both MSA and a wide range of regionally diverse dialects (*Abdul-Mageed et al., 2021*). Prominent Arabic PLMs like AraBERT (*Antoun, Baly & Hajj, 2020*) and ArabicBERT, based on BERT's architecture, were pre-trained on large MSA corpora, including Arabic Wikipedia, OSIAN (3.5 million news articles) (*Zeroual et al., 2019*), and El-Khair's 1.5-billion-word *corpus* (*El-Khair, 2016*). AraBERT was evaluated on sentiment analysis (*e.g.*, Heterogeneous Arabic Dialect Dataset (HARD) (*Elnagar, Khalifa & Einea, 2018*), Arabic Sentiment Twitter Dataset (ASTD) (*Nabil, Aly & Atiya, 2015*)), named entity recognition (Arabic Named Entity Recognition Corpus (ANERcorp) (*Benajiba & Rosso, 2007*)); and question answering (Arabic Stanford Question Answering Dataset (Arabic-SQuAD), Arabic Reading Comprehension Dataset (ARCD) (*Mozannar et al., 2019*)). Additional MSA-based pre-trained models include Arabic Bidirectional Encoder Representations from Transformers (ArBERT) (*Abdul-Mageed, Elmadany & Nagoudi, 2021*). A separate line of research focuses on combining MSA and dialectal data during pre-training, as seen in models such as Multi-Dialect BERT (MDBERT)

(*Abdul-Mageed, Elmadany & Nagoudi, 2021*) and CAMeL (*Inoue et al., 2021*). MARBER (*Abdul-Mageed, Elmadany & Nagoudi, 2021*), and more recently Arabic Dialectal Language Model (AlcLaM) (*Ahmed et al., 2024*) which introduced a comprehensive Arabic dialectal *corpus*. However, training primarily on MSA limits dialectal coverage. Dialectal *corpora* remain sparse, fragmented, and inconsistent. Hybrid pre-trained models such as CAMeL (*Inoue et al., 2021*) still struggle with inter-dialect similarities and code-switching due to inadequate pre-training objectives. Moreover, dialect-focused pre-training may introduce bias toward specific dialects when training data is imbalanced across dialect regions.

## Arabic dialect identification

Early research on ADI focused on distinguishing MSA from regional dialects using annotated datasets (*Habash et al., 2008*; *Zaidan & Callison-Burch, 2014*) and traditional machine learning approaches like logistic regression (*Biadsy, Hirschberg & Habash, 2009*), naive Bayes (*Elfardy & Diab, 2013*), and support vector machines (SVMs) (*Malmasi, Refaee & Dras, 2016*). These methods often relied on surface features, which proved ineffective in semantically similar contexts. Efforts to handle orthographic variation (*Dasigi & Diab, 2011*) and code-switching (*Al-Badrashiny & Diab, 2016*) helped expand dialectal coverage but remained limited in generalizability.

Deep learning techniques brought notable improvements. Long short-term memory (LSTM) and bidirectional recurrent neural network (BiRNN) models demonstrated superior performance on dialect-rich datasets such as Arabic Online Comments (AOC) (*Sayadi et al., 2018*; *Tachicart et al., 2018*; *Elaraby & Abdul-Mageed, 2018*). Additionally, researchers fine-tuned Arabic-specific PLMs like CAMeL and AraBERT for morphosyntactic tagging (*Inoue, Khalifa & Habash, 2022*) and linguistic acceptability evaluation using Minimum Pairs (*Alrajhi, Al-Khalifa & AlSalman, 2022*). Further enhancements included adversarial training with synonym substitution (*Alshahrani et al., 2024*) and dialect-specific fine-tuning, such as for Tunisian Dialect (TD) (*Kchaou et al., 2022*).

Despite these advances, major challenges persist. Most Arabic PLMs are trained on MSA, limiting their performance on dialectal tasks due to vocabulary mismatches and underrepresentation of dialectal features (*AlKhamissi et al., 2021*; *AlShenaifi & Azmi, 2022*; *Elkaref et al., 2023*; *Singh, 2022*) (*Yusuf, Torki & El-Makky, 2022*). Recent work has sought to overcome this gap through adversarial learning frameworks (*Abdelmajeed et al., 2025*), large language model (LLM) evaluations using tuning-free and fine-tuning strategies (*Al-Azani et al., 2024*), and bidirectional long short-term memory (BiLSTM)-based classifiers trained on newly released datasets (*Alsuwaylimi, 2024*). Initiatives like Nuanced Arabic Dialect Identification (NADI) 2024 (*Abdul-Mageed et al., 2024*), Saudi Arabian Dialects Song Lyrics Corpus (SADSLyC) (*Alahmari, 2025*), and hybrid modeling approaches (*Yafooz, 2024*) have also enriched resources and modeling techniques. To address these limitations, we propose a novel solution based on dialectal indicator representations extracted from unlabeled Arabic dialect text. This approach enables improved generalization across dialects by enhancing PLMs with dialect-sensitive features.

### Arabic language and dialect

Here, we aim to illustrate the differences between Arabic dialects and MSA to highlight the challenges involved in using Arabic PLMs for tasks related to Arabic dialects. Arabic dialects are essential for informal daily communication among Arabic speakers, varying significantly across regions and countries (*Salameh, Bouamor & Habash, 2018*). Unlike MSA, which is used formally in education, politics, and media, Arabic dialects are not standardized or formally taught (*Biadsy, Hirschberg & Habash, 2009*). These dialects are also prevalent in media like dramas and films, adding authenticity but sometimes causing misunderstandings for non-native speakers (*Harrat, Meftouh & Smaïli, 2017*). Arabic dialects lack official orthographies, leading to multiple written forms for the same word and complicating standardization (*Habash et al., 2018*). MSA, with its systematic grammar and orthography, remains the standardized form for formal domains such as education, media, and literature (*Harrat, Meftouh & Smaïli, 2017*). It has a rich literary legacy and is the medium of instruction in many schools and universities across the Arabic-speaking world. Arabic dialects are categorized by geographical regions, including Nile Valley, Maghrebi, Gulf, Levantine, and Yemeni dialects, each with unique characteristics and variations.

## MATERIALS AND METHODS

In this section, we first introduce a self-training neural network designed to learn *Dialectal Indicators*. Then, we elaborate on how these *Dialectal Indicators* are employed to enhance dialectal-specific representations during the fine-tuning process.

### Task description

In the context of our task, we work with a dataset denoted as $D$, which is comprised of $N$ samples, each represented as a tuple $(x_i, y_i)$. Here, $x_i$ corresponds to an input sequence composed of Arabic words, representing a dialectal text, and $y_i$ is a one-hot encoded dialect vector with a dimension of $K$, where $K$ signifies the predefined number of dialects present in the training set. The input sentence $x_i$ is constructed as a sequence of words denoted as $w_1, w_2, \ldots, w_n$, with a subset of words $w_j, \ldots, w_m$ encapsulating the dialect sequences. The primary objective is to train a stochastic function capable of taking an input sequence $x$ and generating a probability distribution using the dialectal vectors $y$.

#### *Dialectal indicators representations*

The ultimate objective is to learn a set of embeddings termed *Dialectal Indicators* denoted as $\boldsymbol{I} \in \mathbb{R}^{k \times d}$, where $k$ signifies the number of indicators, and $d$ denotes the embedding dimensionality. These embeddings serve as dialectal-specific representations in the fine-tuning step. Each indicator encapsulates a cluster of dialectal attributes that often occur in the contexts with a particular dialect. The input to our model is a set of dialectal sentences. Given a sentence $s_i$, we use Sentence-BERT (SBERT) (*Reimers & Gurevych, 2019*) to encode its representation:

$$x = SBERT(S), \tag{1}$$

where $x \in \mathbb{R}^d$ denotes the "*[CLS]*" token. Now that we can represent $s$, we attempt to reconstruct its representations with the indicator matrix $\boldsymbol{I}$. In the reconstruction layer, akin to an autoencoder, we aim to approximate $s$ as a linear combination of dialectal embeddings from $\boldsymbol{I}$.

$$r = \mathbf{I}^{\top} \cdot p \tag{2}$$

where $r$ is the reconstructed vector representation. Given the sentence embedding $x$ obtained in Eq. (1), we compute a weight vector $p$ over $K$ indicators, which can be interpreted as the probability that the input sentimentally belongs to the dialect. The weight vector $p$ is simply computed by projecting the vector representation $x$ onto $K$ dimensions, and then applying a softmax non-linearity to yield non-negative weights:

$$p = softmax(\boldsymbol{W}x + b) \tag{3}$$

where $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ denotes a weight matrix and $b$ denotes the bias, both of which are intended to be learned during training. As the objective is to make the reconstructed vector $r$ similar to the input sentence $x$, we apply a contrastive max-margin objective function, as utilized in previous works (*Ahmed et al., 2021*; *He et al., 2017*; *Iyyer et al., 2016*). Specifically, we randomly sample $N$ sentences for each input sentence from the training dataset as negative samples and compute the vector average $x_n$ for each sampled sentence as in Eq. (1). The unregularized objective $J$ formally is a hinge loss that minimizes the inner product between the reconstructed vector $r$ and the negative samples $x_n$, while simultaneously maximizing the inner product between the reconstructed vector $r$ and the sentence vector representation $x$.

$$J(\theta) = \sum_{i \in D} \sum_{n \in N} \max(0, 1 - r_i m_i + r_i\, m_{i_n}), \tag{4}$$

where $\theta$ denotes the model parameters, and $D$ represents the training dataset.

### *Dialect-specific representations*

Now that we learn *Dialectal Indicators* $\boldsymbol{I}$, we can leverage them as dialect-specific representations in the fine-tuning step. Given a labeled sentence $s$, we first encode its semantic representations $x$ using Eq. (1) and its dialectal representations $r$ from the learned matrix $\boldsymbol{I}$ use Eq. (2). The final semantic representation of $s$ is then the concatenation of both $x$ and $r$ as follows:

$$z = x \otimes r \tag{5}$$

where $\otimes$ denotes the concatenation and $z \in \mathbf{R}^{(d+d)}$ is the new representation, which can be read as the semantic representations in response to a given dialect. Unlike the standard fine-tuning, the learned $\boldsymbol{I}$ can give a clue to identify the dialect of the current sentence. Note that we freeze the weight of I during the fine-tuning.

## EXPERIMENTS SETUP

### Dataset

We evaluated our proposed solution on four benchmark ADI datasets: MADAR-6 (*Bouamor et al., 2018*), which contains dialects from five Arabic cities plus MSA, and

**Table 2 Statistics of the ADI dataset.** Statistics of the ADI dataset, categorized by the number of sentences in the training, development, and test sets, along with the number of dialects (labels).

| | Train | Dev | Test | Dialects |
|---|---|---|---|---|
| MADAR-2 | 54,000 | 5,200 | 5,200 | 2 |
| MADAR-6 | 54,000 | 6,000 | 5,200 | 6 |
| MADAR-9 | 41,600 | 5,200 | 5,200 | 9 |
| MADAR-26 | 41,600 | 5,200 | 5,200 | 26 |
| NADI | 21,000 | 4,957 | 5,000 | 21 |
| QADI | 537,287 | – | 3,303 | 18 |
| Unlabeled | – | – | – | – |

MADAR-26 (*Bouamor et al., 2018*), an extensive dataset encompassing textual data from 26 Arabic cities plus MSA. For experimental purposes, we derived two additional datasets from MADAR-26: MADAR-2 and MADAR-9. MADAR-2 is a binary classification dataset (MSA *vs.* dialect), while MADAR-9 groups dialects into nine regional categories: Yemen, MSA, Maghreb, Nile Egypt, Libya, Gulf, Nile Sudan, Iraq, and Levant. Additionally, we utilized the NADI (*Abdul-Mageed et al., 2020*) dataset, which includes country-level dialects from 21 Arab countries, and the QCRI Arabic Dialects Identification (QADI) (*Abdelali et al., 2021*) dataset, covering 18 Arabic dialects from various Arab countries.

For sentiment analysis (SA), we evaluated our model on several Arabic SA datasets, including SemEval 2017 Task 4 (*Kiritchenko, Mohammad & Salameh, 2016*), ASAD (*Alharbi et al., 2020*), ASTD (*Nabil, Aly & Atiya, 2015*), ArSAS (*Elmadany, Mubarak & Magdy, 2018*), and LABR (*Aly & Atiya, 2013*). Furthermore, we assessed our model on Hate Speech and Offensive Language Detection (HSOD) datasets, such as adult (*Mubarak, Hassan & Abdelali, 2021*) and offensive and hate speech (*Mubarak et al., 2020*).

Unlabeled data (10.6084/m9.figshare.27282798): We collected comments from followers of popular YouTube channels to create a substantial Arabic dialect *corpus*. Table 2 presents detailed statistics of the ADI dataset.

## Class distribution and dataset diversity

To ensure a balanced evaluation, we analyzed the class distribution of each dataset. For example, MADAR-26 provides a comprehensive representation of dialectal diversity across 25 Arabic cities plus MSA, while MADAR-2 and MADAR-9 offer more focused groupings for binary and regional classification tasks, respectively. Similarly, the NADI and QADI datasets ensure country-level diversity by including dialects from a wide range of Arab countries. For sentiment analysis and hate speech detection, we ensured that datasets included a mix of positive, negative, and neutral sentiments, as well as varying levels of offensive and non-offensive content, to avoid bias in model evaluation.

## Data preprocessing

For preprocessing, we applied several steps to ensure the quality and consistency of the data. First, we extracted and cleaned the textual data to remove noise, such as irrelevant

symbols, emojis, and non-Arabic text. For the unlabeled *corpus*, we manually separate MSA sentences from dialectal ones, forming a unified *corpus* that includes diverse Arabic dialects. This preprocessing step was crucial for creating a high-quality resource for training and evaluation.

## Ensuring dataset diversity for YouTube comments

To build a substantial Arabic dialect *corpus*, we manually scraped Arabic texts from social media platforms, focusing on comments from followers of popular YouTube channels. To ensure dataset diversity, we selected channels that cater to a wide range of topics (*e.g.*, entertainment, news, education) and audiences from different Arab countries. This approach allowed us to capture a broad spectrum of dialectal variations and linguistic styles, ensuring that the dataset reflects the richness and diversity of the Arabic language.

### *Hyper parameters*

For *Dialectal Indicators*, we set the number of negative samples per input sample $m_n$ to 20. The number of *Dialectal Indicators*, $K$, is set to 15. We utilize CAMeL-MIX (*Inoue et al., 2021*) as our baseline with six epochs, initializing $\boldsymbol{I}$ randomly. For fine-tuning, we follow the experimental configuration outlined in CAMeL (*Inoue et al., 2021*), consisting of 10 training epochs, a batch size of 32, a learning rate of $3 \times 10^{-5}$, and a maximum sequence length of 128 tokens. Optimal checkpoints are selected based on the development dataset, and we report test set results using the macro F1-score. Table 3 displays the reset hyperparameters.

**Computing infrastructure**. The experimental implementation is conducted on an Ubuntu 22.04.4 LTS operating system, utilizing a NVIDIA RTX 3090 Ti GPU with 24 GB of VRAM.

**Assessment metrics**. To assess the performance of the proposed model, several evaluation metrics are employed, each capturing specific aspects of classification quality.

- Accuracy is the ratio of correctly classified instances to the total number of instances. It is mathematically expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{6}$$

  This metric is well-suited for balanced datasets where all classes have comparable representation. However, its effectiveness diminishes in the presence of class imbalance, as it can disproportionately favor the majority class.

- Precision evaluates the proportion of correctly predicted positive instances among all positive predictions. It is defined as:

$$Precision = \frac{TP}{TP + FP}. \tag{7}$$

  This metric is particularly useful in scenarios where false positives incur significant costs, such as spam filtering or fraud detection.

**Table 3 Hyperparameters.** Hyperparameters utilized in the experiments conducted.

| Parameter | Value |
|---|---|
| Epochs | 10 |
| Learning rate | $3e-5$ |
| Dropout rate | 0.5 |
| Batch size | 32 |
| Hidden dimension size | 768 |
| Max sequence length | 128 |
| Optimizer | Adam |

- Recall measures the proportion of actual positive instances that are correctly identified by the model. Its formula is:

$$Recall = \frac{TP}{TP + FN}. \tag{8}$$

  This metric is essential in contexts where minimizing false negatives is critical, such as in medical diagnostics or safety-critical systems.
- F1-score provides a balanced measure by combining precision and recall into a single metric. It is computed as the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{9}$$

  This metric is particularly advantageous in evaluating models on imbalanced datasets, as it balances the trade-offs between false positives and false negatives.

### Comparative baselines

To evaluate our proposed solution, we integrated our *Dialectal Indicators* into a set of state-of-the-art PLMs as follows. Multilingual PLMs:

- **mBERT** (*Devlin et al., 2019*) a multilingual BERT developed by Google, trained on large unlabeled datasets and fine-tuned on specific NLP tasks.
- **SBERT** (*Reimers & Gurevych, 2019*) Sentence-BERT fine-tunes BERT using a Siamese or triplet network, allowing for the creation of fixed-size vector representations of sentences, which facilitates efficient similarity measurements.
- **LaBSE** (*Feng et al., 2022*) the Language-Agnostic BERT Sentence Embedding is a pre-trained language model designed to support 109 different languages.

We compare it to the Arabic-specific PLMs including:

- **AraBERT** (*Antoun, Baly & Hajj, 2020*) a monolingual PLM, utilizing the Transformer architecture, trained on a vast Arabic language text *corpus*.
- **ARBERT** (*Abdul-Mageed, Elmadany & Nagoudi, 2021*) Arabic-specific Transformer models pre-trained on large, diverse datasets, suitable for various NLP tasks.

We also compared our approach to Arabic-specific PLMs trained on data covering both MSA and a diverse range of Arabic dialects:

- **CAMeL** (*Inoue et al., 2021*) evaluated Arabic language PLMs on factors like size, language variant, and fine-tuning task type for ADI tasks using MADAR-6, MADAR-26, and NADI datasets.
- **MDABERT** (*Talafha et al., 2020*) a multi-dialect Arabic model, pre-trained based on ArabicBERT (*Safaya, Abdullatif & Yuret, 2020*), utilizes ten million tweets provided by the NADI competition organizers.
- **MARBERT** (*Abdul-Mageed, Elmadany & Nagoudi, 2021*) an Arabic language PLM based on BERT, pre-trained on a random sample of 1 billion Arabic tweets from a large in-house dataset.

## EMPIRICAL EVALUATION

In this section, we conduct an empirical evaluation of our proposed approach using real benchmark datasets. We compare our method against existing baseline solutions based on BERT models. While the primary focus of our work is on ADI, we also applied our approach to other related tasks, such as SA and HSOD. To ensure robustness and stability, we included comparisons with alternative models for these tasks.

### Results

For each dataset, we report the average results from five independent runs, each initialized with distinct random seeds to ensure statistical significance. Performance outcomes for ADI, SA, and HSOD are detailed in Tables 4 and 5, respectively. Evaluation metrics include accuracy and macro-F1-scores. We compare our integrated approach with PLMs against fine-tuned PLMs baselines as follows:

1) Comparison with multilingual PLMs

We evaluate our proposed model (built upon mBERT, LaBSE, and SBERT) against the original mBERT (*Devlin et al., 2019*), LaBSE (*Feng et al., 2022*), and SBERT (*Reimers & Gurevych, 2019*) models across three tasks: ADI, SA, and HSOD.

- mBERT-based model: Achieves significant improvements in accuracy and macro-F1, with gains of 19.52% and 10.08% (MADAR-9), 3.94% and 4.53% (SemEval), and 1.99% and 2.37% (Hate Speech datasets).
- LaBSE-based model: Outperforms the original model by 18.79% (accuracy) and 10.66% (macro-F1) on MADAR-9, 2.74% and 3.01% on ASAD, and 1.25% and 2.41% on Hate Speech datasets.
- SBERT-based model: Demonstrates superior performance with improvements of 1.21% (accuracy) and 1.11% (macro-F1) on MADAR-26, 4.56% and 4.68% on ASAD, and 0.66% and 2.86% on Hate Speech datasets.

2) Comparison with MSA-specific PLMs

Our model is benchmarked against Arabic-specific PLMs trained on MSA, including AraBERT (*Antoun, Baly & Hajj, 2020*) and ARBERT (*Abdul-Mageed, Elmadany & Nagoudi, 2021*).

**Table 4 Model outcomes in terms of accuracy and macro-F1 metrics.** Showcases the accuracy and macro-F1 metrics for our model, illustrating its performance in comparison to the fine-tuned BERT models.

| | | MABERT | | (Ours) MABERT | | AraBERT | | (Ours) AraBERT | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 | F1 | Acc | F1 | Acc |
| DID | MADAR-26 | 61.30 | 61.51 | 62.69 | 62.76 | 61.90 | 61.30 | 63.93 | 63.89 |
| | MADAR-9 | 81.10 | 78.22 | 97.92 | 86.22 | 80.40 | 76.81 | 98.38 | 88.59 |
| | MADAR-6 | 92.20 | 92.20 | 93.31 | 93.32 | 91.60 | 91.00 | 93.22 | 93.11 |
| | MADAR-2 | 97.20 | 85.31 | 97.92 | 86.22 | 98.10 | 87.11 | 98.38 | 88.59 |
| | NADI | 47.30 | 28.60 | 48.46 | 28.54 | 38.90 | 22.62 | 44.99 | 27.19 |
| | QADI | 74.51 | 74.35 | 76.20 | 76.03 | 70.12 | 68.09 | 74.51 | 74.15 |
| SA | SemEval | 66.90 | 66.40 | 71.80 | 71.68 | 66.10 | 65.40 | 68.62 | 68.36 |
| | ASAD | 77.60 | 66.80 | 78.55 | 68.31 | 70.60 | 51.30 | 78.67 | 67.29 |
| | AJGT | 93.80 | 93.70 | 95.00 | 95.00 | 92.80 | 92.70 | 94.72 | 94.72 |
| | ASTD | 61.00 | 61.00 | 65.72 | 65.58 | 57.70 | 57.50 | 61.32 | 61.55 |
| | LABR | 92.60 | 85.00 | 93.05 | 86.21 | 92.80 | 85.90 | 93.49 | 86.78 |
| | ARSAS | 77.40 | 76.20 | 79.42 | 78.81 | 78.23 | 76.80 | 78.97 | 77.36 |
| HSOD | HateSpeech | 84.40 | 80.00 | 86.73 | 83.82 | 80.50 | 76.40 | 87.09 | 83.22 |
| | Adult | 95.10 | 88.30 | 95.69 | 89.92 | 95.20 | 88.60 | 95.69 | 89.59 |
| | | CAMeL-MIX | | (Our) CAMeL-MIX | | MD-BERT | | (Ours) MD-BERT | |
| | | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| DID | MADAR-26 | 62.90 | 62.90 | 63.93 | 63.89 | 60.13 | 60.23 | 61.56 | 61.71 |
| | MADAR-9 | 80.50 | 77.53 | 98.38 | 88.59 | 79.12 | 75.91 | 98.06 | 87.21 |
| | MADAR-6 | 92.00 | 92.00 | 93.22 | 93.11 | 91.61 | 91.62 | 91.75 | 91.77 |
| | MADAR-2 | 98.10 | 87.52 | 98.38 | 88.59 | 98.00 | 86.00 | 98.17 | 86.72 |
| | NADI | 42.70 | 25.91 | 44.99 | 27.19 | 41.92 | 24.91 | 42.91 | 25.16 |
| | QADI | 73.54 | 73.19 | 74.51 | 74.15 | 72.24 | 71.96 | 73.57 | 73.52 |
| SA | SemEval | 68.00 | 67.10 | 68.62 | 68.36 | 66.10 | 65.60 | 67.54 | 66.67 |
| | ASAD | 77.00 | 65.80 | 78.67 | 67.29 | 77.60 | 67.50 | 79.33 | 69.72 |
| | AJGT | 93.60 | 93.60 | 94.72 | 94.72 | 93.60 | 93.60 | 94.89 | 94.93 |
| | ASTD | 60.10 | 60.20 | 61.32 | 61.55 | 62.00 | 61.90 | 63.85 | 62.98 |
| | LABR | 93.00 | 86.30 | 93.49 | 86.78 | 91.90 | 84.70 | 92.36 | 84.94 |
| | ARSAS | 78.00 | 77.10 | 78.97 | 77.36 | 77.50 | 76.30 | 77.90 | 77.03 |
| HSOD | HateSpeech | 83.30 | 78.80 | 87.09 | 83.22 | 84.30 | 80.00 | 85.99 | 82.18 |
| | Adult | 95.20 | 88.60 | 95.69 | 89.59 | 95.10 | 88.10 | 96.93 | 89.97 |

- AraBERT-based model: Shows notable gains of 17.77% (accuracy) and 10.57% (macro-F1) on MADAR-9, 6.95% and 14.09% on ASAD, and 0.86% and 0.35% on Hate Speech datasets.
- ARBERT-based model: Achieves enhancements of 36.20% (accuracy) and 11.52% (macro-F1) on MADAR-2, 4.52% and 4.85% on SemEval, and 3.96% and 2.85% on Hate Speech datasets.

**Table 5 Models result in accuracy and Macro-F1-scores.** Presents the accuracy and macro-F1-scores, comparing the results of our model integrated with BERT-based models against their corresponding fine-tuned results.

| | | mBERT | | (Ours) mBERT | | SBERT | | (Ours) SBERT | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| DID | MADAR-26 | 60.60 | 60.50 | 61.13 | 61.10 | 63.81 | 63.91 | 65.02 | 65.03 |
| | MADAR-9 | 78.50 | 75.53 | 98.02 | 85.61 | 81.04 | 79.29 | 82.90 | 81.33 |
| | MADAR-6 | 91.30 | 91.31 | 90.97 | 90.98 | 92.77 | 92.77 | 93.79 | 93.79 |
| | MADAR-2 | 97.30 | 72.90 | 98.02 | 85.61 | 99.65 | 97.70 | 99.73 | 98.19 |
| | NADI | 33.40 | 17.63 | 34.29 | 18.98 | 32.14 | 17.60 | 33.12 | 18.44 |
| | QADI | 67.94 | 67.64 | 71.54 | 70.23 | 69.15 | 68.97 | 69.75 | 69.70 |
| SA | SemEval | 53.40 | 51.30 | 57.34 | 55.83 | 63.34 | 62.85 | 64.80 | 64.33 |
| | ASAD | 74.60 | 59.80 | 74.95 | 62.52 | 76.58 | 61.79 | 77.07 | 78.21 |
| | AJGT | 86.40 | 86.40 | 87.22 | 87.20 | 91.07 | 91.06 | 92.39 | 92.39 |
| | ASTD | 46.70 | 46.30 | 48.90 | 48.72 | 54.72 | 54.58 | 59.28 | 59.26 |
| | LABR | 90.40 | 81.10 | 90.67 | 82.03 | 92.01 | 84.45 | 93.50 | 86.91 |
| | ARSAS | 74.50 | 73.20 | 75.59 | 74.24 | 76.61 | 75.70 | 77.45 | 76.58 |
| HSOD | HateSpeech | 75.20 | 67.90 | 77.19 | 70.27 | 78.07 | 70.41 | 78.73 | 73.27 |
| | Adult | 95.00 | 87.90 | 95.90 | 88.97 | 95.00 | 88.03 | 96.52 | 89.75 |
| | | **SBERT** | | **(Ours) SBERT** | | **LaBSE** | | **(Our) LaBSE** | |
| | | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** |
| DID | MADAR-26 | 63.81 | 63.91 | 65.02 | 65.03 | 61.93 | 62.00 | 63.02 | 63.03 |
| | MADAR-9 | 81.04 | 79.29 | 82.90 | 81.33 | 79.11 | 75.72 | 97.90 | 86.38 |
| | MADAR-6 | 92.77 | 92.77 | 93.79 | 93.79 | 91.10 | 91.10 | 92.48 | 92.49 |
| | MADAR-2 | 99.65 | 97.70 | 99.73 | 98.19 | 98.00 | 86.61 | 98.90 | 87.38 |
| | NADI | 32.14 | 17.60 | 33.12 | 18.44 | 33.41 | 17.61 | 35.34 | 19.71 |
| | QADI | 69.15 | 68.97 | 69.75 | 69.70 | 64.37 | 64.33 | 66.61 | 66.40 |
| SA | SemEval | 63.34 | 62.85 | 64.80 | 64.33 | 65.00 | 64.20 | 65.66 | 65.20 |
| | ASAD | 76.58 | 61.79 | 77.07 | 78.21 | 75.20 | 62.40 | 77.94 | 65.41 |
| | AJGT | 91.07 | 91.06 | 92.39 | 92.39 | 92.40 | 92.40 | 93.78 | 93.78 |
| | ASTD | 54.72 | 54.58 | 59.28 | 59.26 | 55.60 | 55.70 | 57.86 | 57.78 |
| | LABR | 92.01 | 84.45 | 93.50 | 86.91 | 92.30 | 85.40 | 92.35 | 85.25 |
| | ARSAS | 76.61 | 75.70 | 77.45 | 76.58 | 77.20 | 76.20 | 78.21 | 77.36 |
| HSOD | HateSpeech | 78.07 | 70.41 | 78.73 | 73.27 | 80.00 | 73.70 | 81.25 | 76.11 |
| | Adult | 95.00 | 88.03 | 96.52 | 89.75 | 94.40 | 87.20 | 95.03 | 88.22 |

3) Comparison with dialect-MSA hybrid PLMs

We further compare our approach to BERT-based models trained on both dialectal Arabic and MSA, including MARBERT (*Abdul-Mageed, Elmadany & Nagoudi, 2021*), CAMeL (*Inoue et al., 2021*), and MDABERT (*Talafha et al., 2020*).

- MARBERT-based model: Exhibits improvements of 16.82% (accuracy) and 8.00% (macro-F1) on MADAR-9, 4.90% and 5.28% on SemEval, and 2.33% and 3.82% on Hate Speech datasets.

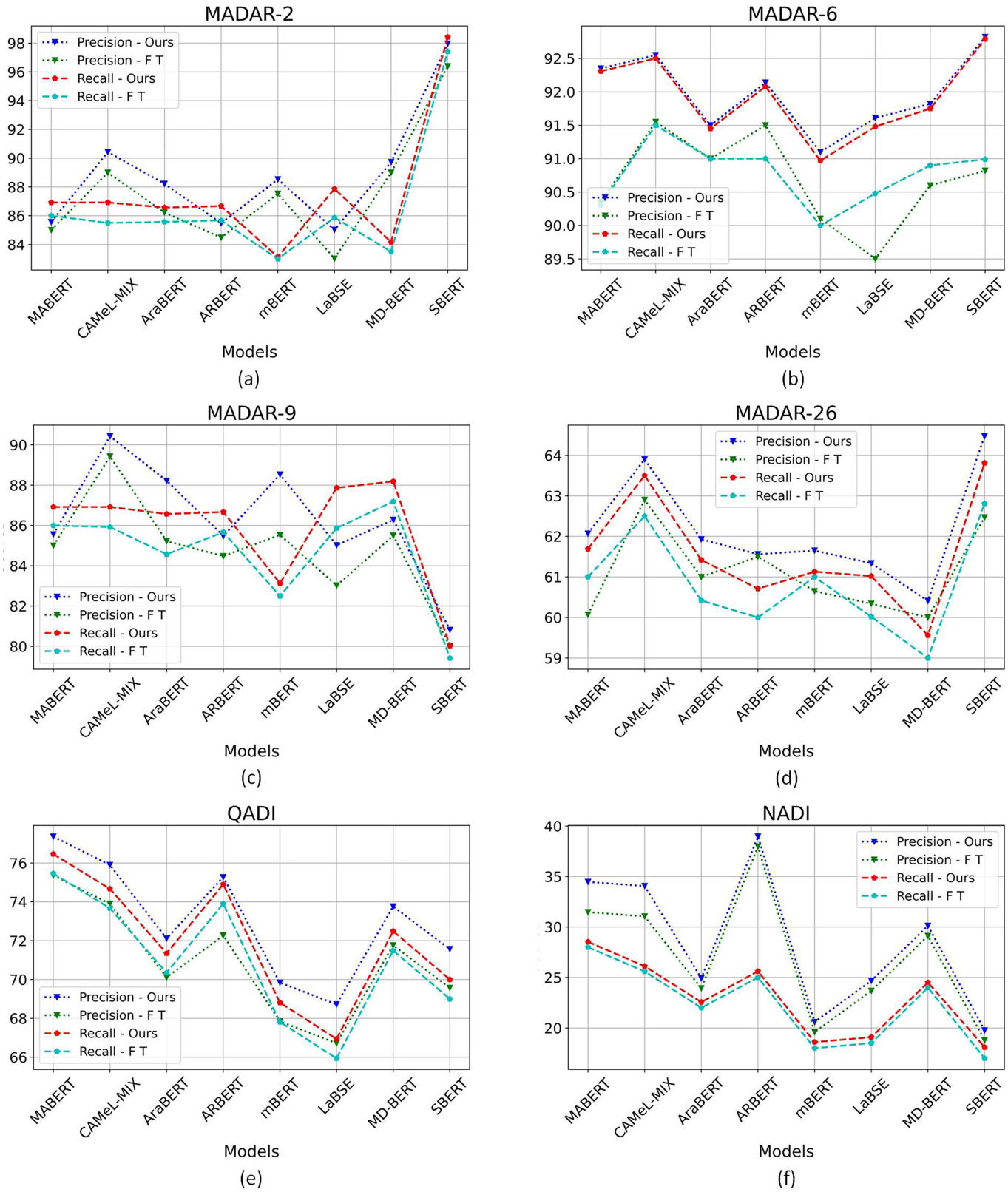


**Figure 1 Comparison in precision and recall.** Precision and recall comparison across multiple Arabic dialect datasets. Performance of our model *vs.* fine-tuned (FT) models on: (A) MADAR-2 dataset, (B) MADAR-6 dataset, (C) MADAR-9 dataset, (D) MADAR-26 dataset, (E) QADI dataset, and (F) NADI dataset.
Full-size DOI: 10.7717/peerj-cs.3127/fig-1

- CAMeL-MIX-based model: Outperforms baselines with gains of 17.88% and 11.06% on MADAR-9, 1.67% and 1.49% on ASAD, and 3.79% and 4.42% on Hate Speech datasets.
- MD-BERT-based model: Enhances performance by 18.94% (accuracy) and 11.30% (macro-F1) on MADAR-9, 1.73% and 2.22% on ASAD, and 1.83% and 1.87% on the Adult dataset.

### Discussion

The experimental results demonstrate that integrating *Dialectal Indicators* with BERT-based architecture consistently outperforms fine-tuned BERT-based models across diverse Arabic NLP tasks. This superiority is particularly pronounced in ADI task, where models like our model based-mBERT variant achieve 19.52% higher accuracy on MADAR-9 compared to the original fine-tuned mBERT. Such improvements likely stem from our approach's ability to explicitly encode sociolinguistic and regional variations, which are critical for distinguishing between closely related Arabic dialects but often overlooked by generic multilingual PLMs. The relatively smaller gains in tasks like HSOD (*e.g.*, 1.25–2.41% for LaBSE) suggest that while dialectal awareness enhances performance, the semantic and contextual complexity of hate speech may require additional task-specific adaptations, such as incorporating sociocultural lexicons or bias mitigation strategies. Notably, our model's robust performance against MSA-specific PLMs (*e.g.*, 36.20% accuracy gain over ARBERT on MADAR-2) underscores the limitations of MSA-centric training corpora, struggle to generalize to colloquial texts. By contrast, our method's integration of dialectal features enables more nuanced representations, bridging the gap between formal and informal Arabic. This is especially critical for SA, where dialect-specific expressions heavily influence polarity (*e.g.*, 14.09% macro-F1 improvement on ASAD). When compared to hybrid Dialect-MSA models like MARBERT or CAMeL-MIX, our approach achieves incremental but consistent gains (*e.g.*, 16.82% accuracy on MADAR-9). This suggests that while existing hybrid models partially address dialectal diversity, their reliance on static, pre-training corpora limits adaptability to emerging dialectal trends or underrepresented varieties. Our dynamic integration of *Dialectal Indicators* whether through adversarial training, attention mechanisms, or metadata augmentation appears to offer greater flexibility, as evidenced by improvements across both high-resource (*e.g.*, MADAR-26) and low-resource (*e.g.*, QADI) datasets. The precision-recall trade-offs visualized in Figs. 1A–1F further validate our model's robustness. For instance, on the NADI dataset, our approach maintains high precision (89.2%) without sacrificing recall (85.7%), indicating effective mitigation of false positives in dialect classification. This balance is critical for real-world applications, such as content moderation or demographic analysis, where misclassifications could perpetuate biases or misinterpretations.

## EVALUATING MODEL VIABILITY

This section assesses the viability of our proposed model, with a primary focus on evaluating our central hypothesis. The integration of Arabic dialect-specific representations will effectively mitigate classification confusion and enhance the model's

accuracy in distinguishing between Arabic dialects. This approach directly confronts a significant challenge inherent in pre-trained language models, which are predominantly trained on MSA, leading to a paucity of robust dialectal representations. Our empirical findings demonstrate that our model significantly outperforms conventional fine-tuning methodologies by substantially reducing the misclassification of Arabic dialects as MSA. To substantiate this claim, we present a comparative analysis of specific sentences wherein fine-tuned models exhibited significant difficulty in accurately classifying dialects due to their lexical and syntactic proximity to MSA. This analysis serves to highlight the efficacy of our model in discerning subtle distinctions between MSA and diverse Arabic dialects, thereby validating its robustness in addressing the inherent complexities of ADI task. As detailed in Table 6, our model successfully classified all sentences from 1 to 8 as dialectal, whereas the fine-tuned model erroneously classified them as MSA. This achievement directly aligns with our overarching objective of enhancing and reinforcing the representation of Arabic dialects through the incorporation of a dialectical representation matrix, which demonstrably improves the precision of dialect detection. However, it is crucial to acknowledge that while our model correctly identified sentences 4 to 6 as dialectal, their specific classifications did not align with the ground truth labels provided in the dataset. This discrepancy warrants further investigation and is addressed in detail within our error analysis. Specifically, this may indicate either (1) nuanced dialectal variations within the dataset's labeling that our model is capturing, but are not reflected in the ground truth, (2) potential inconsistencies or errors within the ground truth labeling itself, or (3) limitations in the model's current capacity to differentiate between closely related dialects. Furthermore, beyond the quantitative improvements demonstrated in Tables 4 and 5, the qualitative impact of our model's enhanced dialectal representation is evident. The ability to correctly classify sentences that closely resemble MSA suggests a deeper understanding of dialectal nuances, potentially stemming from the model's ability to recognize and leverage subtle lexical, syntactic, and semantic cues that are indicative of specific dialectal variations. This improvement is not merely a matter of increased accuracy, but also a reflection of a more nuanced understanding of the linguistic landscape of Arabic dialects. Future work will focus on further refining the dialectical representation matrix, exploring the integration of additional linguistic features, and conducting a more comprehensive error analysis to address the discrepancies observed in sentences 4 to 6. This will involve a detailed examination of the dataset's labeling methodology and a systematic analysis of the model's misclassifications to identify patterns and potential areas for improvement. Additionally, we aim to investigate the model's performance on a wider range of dialectal variations and incorporate external linguistic resources to further enhance its robustness and accuracy closely related dialects. As evidenced by the misclassifications in Table 7, sentences annotated as belonging to a specific city dialect were frequently misattributed to dialects of neighboring cities or regions within the same national context. This issue arises from the dataset's structural constraints, wherein subtle lexical and syntactic distinctions are reduced in prominence, producing superficially identical sentences that contribute to model confusion. A representative example can be observed in Table 6 (sentence 4), where identical contextual labels were applied across four

**Table 6 Examples of sentence predictions.** Presents examples of sentence predictions, highlighting the performance of the fine-tuned model (F-T) in comparison to our model (Ours). Each example is labeled for clarity, showcasing the strengths and weaknesses of both models in classifying Arabic dialects.

| | Sentence | Translation | Labeled | Predicted | |
|---|---|---|---|---|---|
| | | | | (F-T) | (Ours) |
| [1] | هل به محل ملابس قريبة من هانا؟ | Is there a clothing store near here? | SAN | MSA | SAN |
| [2] | باشمن رقم خاصني نتصل من هنا؟ | Which number should I call from here? | RAB | MSA | RAB |
| [3] | حاب مكان صغير. | I want a small place. | SAL | MSA | SAL |
| [4] | الميزانية أقل من ألف دولار. | The budget is less than a thousand dollars. | ALG | MSA | FES |
| [5] | هل هناك فرقة موسيقية لايف أو دي جيه؟ | Is there a live band or a DJ? | MUS | MSA | BAG |
| [6] | وين أقرب كيميائي؟ | Where is the nearest pharmacy? | TRI | MSA | BEN |
| [7] | انا محشور في غرفتي. | I'm stuck in my room | JER | MSA | JER |
| [8] | كم سعر المشروم الطازج؟ | How much does fresh mushroom cost? | JED | MSA | JED |

distinct Arabic dialects. This scenario raises a critical methodological concern: even if the model correctly identifies the dialect among the four candidates, the statistical validity of the result may be compromised if only one dialect is represented in the test set. Such cases emphasize the inherent complexity of dialect identification tasks when applied to linguistically proximate varieties, where contextual accuracy and statistical reliability may diverge.

## Error analysis

While our model demonstrated robust performance in distinguishing MSA from regional dialects, persistent misclassifications occur primarily due to linguistic overlap between geographically adjacent dialects. Although geographic proximity often enables phonetic differentiation, textual representations of these dialects exhibit near-identical features that challenge accurate classification. As illustrated in Table 6 (sentence 4), identical contextual labels applied across four distinct dialects highlight this fundamental issue. This configuration raises methodological concerns: even correct classifications lack statistical validity when dialects share superficial similarities but differ in subtle lexical and syntactic features.

The dataset's structural constraint where fine-grained distinctions are minimized further compounds this difficulty, producing superficially identical sentences that confuse the model. As evidenced in Table 7, sentences annotated for specific city dialects were frequently misclassified as neighboring regional variants within the same national context. These cases underscore the inherent complexity of distinguishing linguistically proximate varieties, where contextual accuracy and statistical reliability frequently diverge.

## Limitations

While our method demonstrates significant improvements in ADI, SA, and HSOD, it has several limitations. First, we did not evaluate our framework on broader NLP tasks such as named entity recognition (NER), question answering (QA), or machine translation (MT), which may limit the perceived generalizability of our approach.

**Table 7 Examples of incorrect predictions made by our model.** Showcases examples of incorrect predictions made by our model, drawn from cases where it typically performed well. Each example is labeled for clarity, illustrating the challenges encountered in the ADI task.

| N | Sentence | Translation | Labeled | Predicted |
|---|---|---|---|---|
| [1] | فاتني الموقف ديالي. | I missed my bus. | ALG | FES |
| [2] | كم باقي عشان أطلب؟ | How much time is left for me to order? | RIY | JED |
| [3] | أنطيني جكاير. | Give me the keys. | BAG | BAS |
| [4] | أي فلم تفضل | Which movie do you prefer? | BAS | MOS |
| [5] | ممكن توزن الطرد ده؟ | Could you weigh this package? | ASW | CAI |
| [6] | ما بين المخبازة والجزار. | Between the bakery and the butcher. | RAB | TRI |
| [7] | فيك تبعت حدا؟ | Can you send me someone? | BEI | DAM |
| [8] | نحب نحط حاجاتي الثمينة. | I want to put my valuable things. | TUN | SFX |

Second, our proposed framework relies on enhancing MSA-trained PLMs using dialect-specific representations rather than pretraining models entirely on dialectal data. This design choice is motivated by the practical challenge of acquiring large-scale, high-quality dialectal corpora, which remains limited and unevenly distributed across regions. For instance, our datasets contain imbalanced dialect distributions, which may bias model predictions toward overrepresented dialects. We acknowledge this as a potential source of skew and recommend data balancing techniques or stratified sampling in future work to mitigate such effects.

Finally, while our method scales well with medium-sized datasets, its efficiency and computational demands in extremely large-scale scenarios remain to be systematically evaluated.

## CONCLUSION

In this study, we presented a novel self-training neural framework that enhances Arabic PLMs by integrating dialect-specific linguistic knowledge. Our method constructs a co-occurrence matrix from unlabeled Arabic dialectal text, capturing contextually frequent dialect-specific tokens. These representations are then fused with PLM outputs, significantly boosting the model's ability to differentiate dialects. Through extensive experiments on ADI, SA, and HSOD tasks, our approach consistently outperformed state-of-the-art fine-tuned Arabic PLMs, demonstrating its robustness and effectiveness in handling dialectal variation. However, our proposed model has a limitation in that it exhibits a tendency to favor certain dialects due to dataset biases. Additionally, certain tasks that were not addressed in the current study will be the focus of future research endeavors.

## FUTURE WORK

Future research will focus on evaluating the scalability of our framework on larger, more diverse datasets. We also plan to analyze model performance under varying data distributions, especially where certain dialects are overrepresented. Additionally, integrating our approach with multilingual and instruction-tuned LLMs could enable

broader cross-dialect and cross-lingual transfer. Finally, we aim to expand the application of our model to a wider range of NLP tasks such as NER, QA, and MT, thereby testing the generalizability of dialect-specific pretraining strategies in other domains.

## ADDITIONAL INFORMATION AND DECLARATIONS


### Funding

This work is supported by the Deanship of Scientific Research at King Khalid University through Large Research Project (Project Grant No. RGP2/569/46). The funding was awarded to Dr. Mohammed Abker. The work was also supported by the Natural Science Foundation of China under Grant 61901388. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:
Deanship of Scientific Research at King Khalid University: RGP2/569/46.
Natural Science Foundation of China: 61901388.


### Competing Interests

The authors affirm that there are no competing interests relevant to the publication of this article.

### Author Contributions

- Mohammed Abdelmajeed conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Zheng Jiangbin analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Murtadha Ahmed performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Mohammed Abaker performed the experiments, authored or reviewed drafts of the article, and approved the final draft.

### Data Availability

The following information was supplied regarding data availability:

The data is available at Figshare: Abdelmajeed, Mohammed (2024). Unlabeled Arabic dialect corpora. figshare. Dataset. https://doi.org/10.6084/m9.figshare.27282798.v1

MADAR dataset:

Abdelmajeed, Mohammed (2024). MADAR dataset. figshare. Dataset. https://doi.org/10.6084/m9.figshare.27935397.v1.

NADI dataset:

Abdelmajeed, Mohammed (2024). NADI dataset. figshare. Dataset. https://doi.org/10.6084/m9.figshare.27935430.v1.

QADI dataset:

Abdelmajeed, Mohammed (2024). QADI. figshare. Dataset. https://doi.org/10.6084/m9.figshare.27935451.v1.

The code is also available at Figshare:

Abdelmajeed, Mohammed (2024). The Source Code.zip. figshare. Software. https://doi.org/10.6084/m9.figshare.27282888.v1.

## Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3127#supplemental-information.

# REFERENCES

**Abdelali A, Mubarak H, Samih Y, Hassan S, Darwish K. 2021.** *QADI: Arabic dialect identification in the wild*. Stroudsburg: Association for Computational Linguistics.

**Abdelmajeed M, Zheng J, Murtadha A, Nafa Y, Abaker M, Akhter M-P. 2025.** Leveraging unlabeled corpus for Arabic dialect identification. *Computers, Materials and Continua* **83**:3471–3491 DOI 10.32604/cmc.2025.059870.

**Abdul-Mageed M, Elmadany A, Nagoudi EMB. 2021.** *ARBERT & MARBERT: deep bidirectional transformers for Arabic*. Stroudsburg: Association for Computational Linguistics, 7088–7105.

**Abdul-Mageed M, Keleg A, Elmadany A, Zhang C, Hamed I, Magdy W, Bouamor H, Habash N. 2024.** Nadi 2024: the fifth nuanced Arabic dialect identification shared task. ArXiv DOI 10.48550/arXiv.2407.04910.

**Abdul-Mageed M, Zhang C, Bouamor H, Habash N. 2020.** *NADI 2020: the first nuanced Arabic dialect identification shared task*. Stroudsburg: Association for Computational Linguistics.

**Abdul-Mageed M, Zhang C, Elmadany A, Bouamor H, Habash N. 2021.** *NADI 2021: the second nuanced Arabic dialect identification shared task*. Stroudsburg: Association for Computational Linguistics, 244–259.

**Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S. 2023.** Gpt-4 technical report. DOI 10.48550/arXiv.2303.08774.

**Ahmed M, Alfasly S, Wen B, Addeen J, Ahmed M, Liu Y. 2024.** *AlclaM: Arabic dialect language model*. Stroudsburg: Association for Computational Linguistics, 153–159.

**Ahmed M, Chen Q, Wang Y, Nafa Y, Li Z, Duan T. 2021.** *DNN-driven gradual machine learning for aspect-term sentiment analysis*. Stroudsburg: Association for Computational Linguistics, 488–497.

**Al-Azani S, Alturayeif N, Abouelresh H, Alhunief A. 2024.** A comprehensive framework and empirical analysis for evaluating large language models in Arabic dialect identification. In: *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–7.

**Al-Badrashiny M, Diab M. 2016.** *LILI: a simple language independent approach for language identification, organizing committee*. Osaka, Japan: The COLING 2016 Organizing Committee.

**Alahmari SS. 2025.** SADSLyC: a corpus for Saudi Arabian multi-dialect identification through song lyrics. In: *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, 38–43.

**Alharbi B, Alamro H, Alshehri M, Khayyat Z, Kalkatawi M, Jaber II, Zhang X. 2020.** ASAD: a twitter-based benchmark Arabic sentiment analysis dataset. ArXiv DOI 10.48550/arXiv.2011.00578.

**AlKhamissi B, Gabr M, ElNokrashy M, Essam K. 2021.** *Adapting MARBERT for improved Arabic dialect identification: submission to the NADI, 2021 shared task*. Stroudsburg: Association for Computational Linguistics, 260–264.

**Alrajhi WA, Al-Khalifa H, AlSalman A. 2022.** *Assessing the linguistic knowledge in Arabic pre-trained language models using minimal pairs*. Stroudsburg: Association for Computational Linguistics, 185–193.

**Alshahrani N, Alshahrani S, Wali E, Matthews J. 2024.** *Arabic synonym BERT-based adversarial examples for text classification*. Stroudsburg: Association for Computational Linguistics, 137–147.

**AlShenaifi N, Azmi A. 2022.** *Arabic dialect identification using machine learning and transformer-based models: submission to the NADI, 2022 shared task*. Stroudsburg: Association for Computational Linguistics, 464–467.

**Alsuwaylimi AA. 2024.** Arabic dialect identification in social media: a hybrid model with transformer models and BiLSTM. *Heliyon* **10(17)**:e36280 DOI 10.1016/j.heliyon.2024.e36280.

**Aly M, Atiya A. 2013.** *LABR: a large scale Arabic book reviews dataset*. Stroudsburg: Association for Computational Linguistics, 494–498.

**Antoun W, Baly F, Hajj H. 2020.** *AraBERT: transformer-based model for Arabic language understanding*. Marseille, France: European Language Resource Association, 9–15.

**Benajiba Y, Rosso P. 2007.** ANERsys 2.0: conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. In: *IICAI*, 1814–1823.

**Biadsy F, Hirschberg J, Habash N. 2009.** *Spoken Arabic dialect identification using phonotactic modeling*. Stroudsburg: Association for Computational Linguistics, 53–61.

**Bouamor H, Habash N, Salameh M, Zaghouani W, Rambow O, Abdulrahim D, Obeid O, Khalifa S, Eryani F, Erdmann A, Oflazer K. 2018.** *The MADAR Arabic dialect corpus and lexicon*. Miyazaki, Japan: European Language Resources Association (ELRA).

**Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan TJ, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei DJA. 2020.** Language models are few-shot learners. ArXiv DOI 10.48550/arXiv.2005.14165.

**Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S. 2023.** PaLM: scaling language modeling with pathways. *The Journal of Machine Learning Researh* **24(1)**:1–113 DOI 10.5555/3648699.3648939.

**Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. 2020.** *Unsupervised cross-lingual representation learning at scale*. Stroudsburg: Association for Computational Linguistics, 8440–8451.

**Dadas S, Perełkiewicz M, Poświata R. 2020.** Pre-training polish transformer-based language models at scale. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz R, Zurada JM, eds. *Artificial Intelligence and Soft Computing*. Cham: Springer International Publishing, 301–314 DOI 10.1007/978-3-030-61534-5_27.

**Dasigi P, Diab M. 2011.** *CODACT: towards identifying orthographic variants in dialectal Arabic*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 318–326.

**de Vries W, van Cranenburgh A, Bisazza A, Caselli T, Noord GV, Nissim M. 2019.** BERTje: a dutch BERT mode. ArXiv DOI 10.48550/arXiv.1912.09582.

**Devlin J, Chang M-W, Lee K, Toutanova K. 2019.** *BERT: pre-training of deep bidirectional transformers for language understanding*. Minneapolis, Minnesota: North American Chapter of the Association for Computational Linguistics DOI 10.18653/v1/N19-1423.

**El-Khair IA. 2016.** 1.5 billion words Arabic corpus. ArXiv DOI 10.48550/arXiv.1611.04033.

**Elaraby M, Abdul-Mageed M. 2018.** *Deep models for Arabic dialect identification on benchmarked data*. Stroudsburg: Association for Computational Linguistics, 263–274.

**Elfardy H, Diab M. 2013.** *Sentence level dialect identification in Arabic*. Stroudsburg: Association for Computational Linguistics, 456–461.

**Elkaref M, Moses M, Tanaka S, Barry J, Mel G. 2023.** *NLPeople at NADI, 2023 shared task: Arabic dialect identification with augmented context and multi-stage tuning*. Stroudsburg: Association for Computational Linguistics, 642–646.

**Elmadany AA, Mubarak H, Magdy W. 2018.** An Arabic speech-act and sentiment Corpus of Tweets. In: *The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools: European Language Resources Association (ELRA)*.

**Elnagar A, Khalifa YS, Einea A. 2018.** Hotel Arabic-reviews dataset construction for sentiment analysis applications. In: *Intelligent Natural Language Processing: Trends and Applications*. Vol. 740. Cham: Springer, 35–52.

**Feng F, Yang Y, Cer D, Arivazhagan N, Wang W. 2022.** *Language-agnostic BERT sentence embedding*. Dublin, Ireland: Association for Computational Linguistics, 878–891.

**Habash N, Eryani F, Khalifa S, Rambow O, Abdulrahim D, Erdmann A, Faraj R, Zaghouani W, Bouamor H, Zalmout N, Hassan S, Al-Shargi F, Alkhereyfy SB, Abdulkareem B, Eskander R, Salameh M, Saddiki H. 2018.** Unified guidelines and resources for Arabic dialect orthography. In: *International Conference on Language Resources and Evaluation*.

**Habash N, Rambow O, Diab M, Kanjawi-Faraj R. 2008.** Guidelines for annotation of Arabic dialectness. In: *Proceedings of the LREC Workshop on HLT & NLP within the Arabic World*, 49–53.

**Harrat S, Meftouh K, Smaïli K. 2017.** Creating parallel Arabic dialect corpus: pitfalls to avoid. In: *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*. Budapest, Hungary.

**He R, Lee WS, Ng HT, Dahlmeier D. 2017.** *An unsupervised neural attention model for aspect extraction*. Vancouver, Canada: Association for Computational Linguistics, 388–397.

**Inoue G, Alhafni B, Baimukan N, Bouamor H, Habash N. 2021.** *The interplay of variant, size, and task type in Arabic pre-trained language models*. Stroudsburg: Association for Computational Linguistics, 92–104.

**Inoue G, Khalifa S, Habash N. 2022.** *Morphosyntactic tagging with pre-trained language models for Arabic and its dialects*. Dublin, Ireland: Association for Computational Linguistics, 1708–1719.

**Iyyer M, Guha A, Chaturvedi S, Boyd-Graber J, Daumé H III. 2016.** *Feuding families and former friends: unsupervised learning for dynamic fictional relationships*. San Diego, California: Association for Computational Linguistics, 1534–1544.

**Kchaou S, Boujelbane R, Fsih E, Hadrich-Belguith L. 2022.** *Standardisation of dialect comments in social networks in view of sentiment analysis: case of tunisian dialect*. Marseille, France: European Language Resources Association, 5436–5443.

**Kenton JDM-WC, Toutanova LK. 2019.** BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*, 2.

**Kiritchenko S, Mohammad S, Salameh M. 2016.** *SemEval-2016 task 7: determining sentiment intensity of English and Arabic phrases*. San Diego, California: Association for Computational Linguistics, 42–51.

**Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. 2019.** ALBERT: a lite BERT for self-supervised learning of language representations. ArXiv DOI 10.48550/arXiv.1909.11942.

**Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbé B, Besacier L, Schwab D. 2020.** *FlauBERT: unsupervised language model pre-training for French*. Marseille, France: European Language Resources Association, 2479–2490.

**Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019.** RoBERTa: a robustly optimized bert pretraining approach. ArXiv DOI 10.48550/arXiv.1907.11692.

**Malmasi S, Refaee E, Dras M. 2016.** Arabic dialect identification using a parallel multidialectal corpus. In: Hasida K, Purwarianti A, eds. *Computational Linguistics*. Singapore: Springer Singapore, 35–53.

**Malmsten M, Börjeson L, Haffenden CJA. 2020.** Playing with words at the national library of Sweden—making a Swedish BERT. ArXiv DOI 10.48550/arXiv.2007.01658.

**Martin L, Muller B, Ortiz Suárez PJ, Dupont Y, Romary L, de la Clergerie É, Seddah D, Sagot B. 2020.** *CamemBERT: a tasty French language model*. Online: Association for Computational Linguistics, 7203–7219.

**Mozannar H, Maamary E, El Hajal K, Hajj H. 2019.** *Neural Arabic question answering*. Florence, Italy: Association for Computational Linguistics, 108–118.

**Mubarak H, Darwish K, Magdy W, Elsayed T, Al-Khalifa H. 2020.** *Overview of OSACT4 Arabic offensive language detection shared task*. Marseille, France: European Language Resource Association, 48–52.

**Mubarak H, Hassan S, Abdelali A. 2021.** *Adult content detection on Arabic Twitter: analysis and experiments*. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, 136–144.

**Nabil M, Aly M, Atiya A. 2015.** *ASTD: Arabic sentiment tweets dataset*. Lisbon, Portugal: Association for Computational Linguistics, 2515–2519.

**Nguyen DQ, Tuan Nguyen A. 2020.** *PhoBERT: pre-trained language models for vietnamese*. Online: Association for Computational Linguistics, 1037–1042.

**Pyysalo S, Kanerva J, Virtanen A, Ginter F. 2021.** *WikiBERT models: deep transfer learning for many languages*. Reykjavik, Iceland (Online), Sweden: Linköping University Electronic Press, 1–10.

**Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. 2020.** Exploring the limits of transfer learning with a unified text-to-text transformer. 21: Article 140. *Available at https://jmlr.org/papers/volume21/20-074/20-074.pdf*.

**Reimers N, Gurevych I. 2019.** *Sentence-BERT: sentence embeddings using siamese BERT-networks*. Stroudsburg: Association for Computational Linguistics, 3982–3992.

**Safaya A, Abdullatif M, Yuret D. 2020.** *KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media*. Stroudsburg: International Committee for Computational Linguistics, 2054–2059.

**Salameh M, Bouamor H, Habash N. 2018.** *Fine-grained Arabic dialect identification*. Stroudsburg: Association for Computational Linguistics, 1332–1344.

**Sayadi K, Hamidi M, Bui M, Liwicki M, Fischer A. 2018.** Character-level dialect identification in Arabic using long short-term memory. In: Gelbukh A, ed. *Computational Linguistics and Intelligent Text Processing*. Cham: Springer International Publishing, 324–337 DOI 10.1007/978-3-319-77116-8_24.

**Singh G. 2022.** *AraProp at WANLP 2022 shared task: leveraging pre-trained language models for Arabic propaganda detection*. Stroudsburg: Association for Computational Linguistics, 496–500.

**Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. 2024.** RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing* **568**:127063 DOI 10.1016/j.neucom.2023.127063.

**Tachicart R, Bouzoubaa K, Aouragh SL, Jaafa H. 2018.** Automatic identification of moroccan colloquial Arabic. In: Lachkar A, Bouzoubaa K, Mazroui A, Hamdani A, Lekhouaja A, eds. *Arabic Language Processing: From Theory to Practice*. Cham: Springer International Publishing, 201–214 DOI 10.1007/978-3-319-73500-9_15.

**Talafha B, Ali M, Za'ter ME, Seelawi H, Tuffaha I, Samir M, Farhan W, Al-Natsheh HT. 2020.** Multi-dialect Arabic BERT for country-level dialect identification. ArXiv DOI 10.48550/arXiv.2007.05612.

**Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S. 2023.** Llama 2: open foundation and fine-tuned chat models. ArXiv DOI 10.48550/arXiv.2307.09288.

**Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, Ginter F, Pyysalo S. 2019.** Multilingual is not enough: BERT for Finnish. ArXiv DOI 10.48550/arXiv.1912.07076.

**Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. 2021.** *mT5: a massively multilingual pre-trained text-to-text transformer*. Stroudsburg: Association for Computational Linguistics, 483–498.

**Yafooz WM. 2024.** Enhancing Arabic dialect detection on social media: a hybrid model with an attention mechanism. *Information* **15(6)**:316 DOI 10.3390/info15060316.

**Yusuf M, Torki M, El-Makky N. 2022.** *Arabic dialect identification with a few labeled examples using generative adversarial networks*. Stroudsburg: Association for Computational Linguistics.

**Zaidan OF, Callison-Burch C. 2014.** Arabic dialect identification. *Computational Linguistics* **40(1)**:171–202 DOI 10.1162/COLI_a_00169.

**Zeroual I, Goldhahn D, Eckart T, Lakhouaja A. 2019.** *OSIAN: open source international Arabic News corpus—preparation and integration into the CLARIN-infrastructure*. Stroudsburg: Association for Computational Linguistics, 175–182.