Applied Computing and Informatics

Comparative analysis of deep learning and transfer learning techniques for dialect identification in the Gulf and Saudi regions

Nouf Alshenaifi, Aqil M. Azmi and Manar Hosny
Department of Computer Science,
College of Computer and Information Sciences, King Saud University,
Riyadh, Saudi Arabia

Received 22 February 2025
Revised 9 June 2025
17 September 2025
Accepted 30 September 2025

Abstract

Purpose – This paper investigates the automatic identification of Arabic dialects within the Gulf Cooperation Council (GCC) countries and regional dialects across Saudi Arabia. Dialect identification (DI) is critical for enhancing natural language processing (NLP) applications in the Arab world, where significant dialectal variations exist.

Design/methodology/approach – Using two newly constructed Twitter corpora, GAC-6 (~ 1.7M tweets across six Gulf dialects) and SDC-5 (~ 790K tweets across five Saudi dialects), we develop a multi-DI framework that leverages deep learning and transfer learning. We train long short-term memory (LSTM), BiLSTM and a fine-tuned Arabic bidirectional encoder representations from transformers (AraBERT) model and benchmark them under a single, fair protocol against strong classical baselines – logistic regression, linear support vector machine, multinomial Naïve Bayes, Random Forest and a soft-voting ensemble – trained on character *n*-gram (2–5) and word *n*-gram (1–3) features using bag-of-words/ term frequency and inverse document frequency (TF-IDF) representations.

Findings – Transformer-based transfer learning yields the best performance: AraBERT achieved superior performance, with macro-F1 score of 90.1% on GAC-6 and 90.6% on SDC-5. This exceeded the strongest non-transformer baselines by +2.5 macro-F1 (vs. the ensemble) and +1.2 (vs. BiLSTM) on the respective datasets and by wider margins against other classical models. To assess robustness under domain shift, we also report results on public benchmarks (NADI, QADI), where the transformer remains state-of-the-art among the compared systems.

Originality/value – The paper introduces a novel approach by focusing on Arabic dialects in a specific region and using advanced NLP techniques for dialect differentiation. It is the first to use the large, specialized datasets GAC-6 and SDC-5 for DI.

Keywords Arabic dialects, Dialect identification, Comparative analysis, Deep learning, Transfer learning, Machine learning

Paper type Research article

1. Introduction

The task of dialect identification (DI) has become increasingly important in the field of language identification and natural language processing (NLP) [1]. DI refers to the process of differentiating between members of the same language family and automatically categorizing the language vocabulary used by a particular group of people into the geographical region to which the native speaker belongs [2]. The rise of social media platforms like Twitter and Facebook has significantly increased the exposure and use of various Arabic dialects in online communications across the Arab world [3].

© Nouf Alshenaifi, Aqil M. Azmi and Manar Hosny. Published in *Applied Computing and Informatics*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at Link to the terms of the CC BY 4.0 licence.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-02-2025-0067

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

Arabic, known for its complex semantics, elaborate grammar and extensive morphological diversity, features a wide range of dialects and significant variability in word interpretation [3]. For example, the word (عَقْد) can mean "necklace," "knots" or "contract" depending on its context. Arabic is divided into two main forms: Modern Standard Arabic (MSA) and Colloquial Arabic [4, 5]. Unlike MSA, these dialects are typically employed in casual settings, including social media and are characterized by their non-standardized and often unstructured spelling [6].

Dialectal Arabic exhibits significant morphological, lexical, phonetic, phonological and syntactic variations from MSA, with differences at both macro (regional) and micro (local) levels [7]. At the macro level, dialects are categorized into Gulf, Egyptian, North African (Maghrebi), Levantine and Iraqi [8]. The Gulf region features distinct dialects within each country, while Saudi Arabia shows micro-level diversity, including Najdi, Hijazi, Southern, Northern and Eastern dialects, each dominant in specific areas. Figures 1a and 1b illustrate these variations, highlighting the nuanced differences within the Gulf and Saudi Arabia.

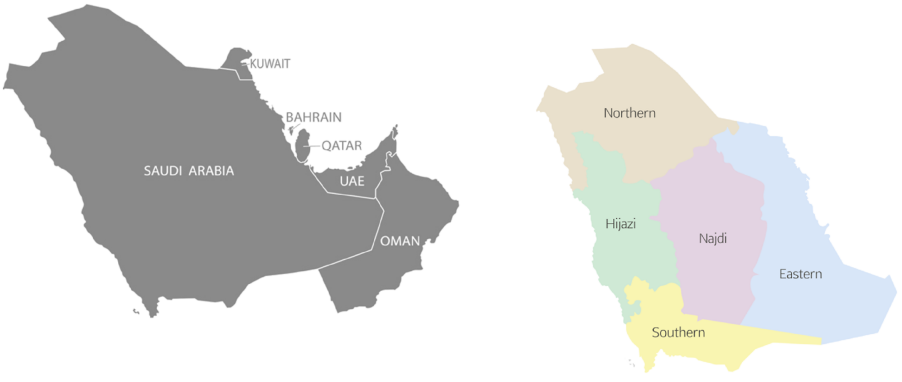
Growing interest in Arabic DI stems from increased online interactions and the rise of written dialects on social media. Research includes binary classification (dialects vs. MSA) and multi-way classification (country-level dialects) [1]. Recognizing dialectal variations is crucial for NLP tasks like sentiment analysis, translation and hate speech detection [9].

This paper focuses on automatically identifying Arabic dialects from social media text using machine learning (ML) and deep learning approaches. We examined Gulf Arabic – spoken in Saudi Arabia, Kuwait, Bahrain, Qatar, the UAE and Oman – and explored Saudi regional variations. Using a dataset from Ref. [10], we developed tweet-level models for DI, distinguishing Gulf dialects effectively. This study included deep learning architectures (LSTM, BiLSTM, fine-tuned AraBERT) and traditional classifiers, with a comparative analysis against benchmark datasets.

The key contributions of this study are summarized as follows:

(1) We used corpora from Ref. [10] to build an advanced Arabic DI framework, combining classical ML with deep learning and transformer-based models.

(2) We compared the proposed models using benchmark datasets, demonstrating their state-of-the-art performance.



Arabian Gulf countries **(a)**

Saudi Arabia **(b)**

Figure 1. The map shows: (a) the six GCC countries and (b) Saudi subregions covered by each dialect. Source: iStock/Getty Images (Standard License). Adapted by the authors

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

This paper is structured as follows: Section 2 reviews Arabic DI literature. Section 3 describes the dataset. Section 4 outlines the identification models. Section 5 presents experimental results and Section 6 concludes with future research directions.

2. Related work

Arabic DI classifies Arabic texts into predefined dialects, a task complicated by shared slang and vocabulary across many dialects and lack of standardized orthography [4, 11]. This area has seen significant research within Arabic NLP [12], employing traditional ML and advanced methods like deep neural networks (DNNs) and pre-trained language models. Research has explored various ML models, analyzing character and word level features for dialect differentiation. Initial studies focused on differentiating specific dialects from MSA, with methods ranging from Naive Bayes classifiers for sentence-level identification [13] to utilizing word and character *n*-gram features in a multinomial Naive Bayes (MNB) model [14]. Other approaches include using Naive Bayes and support vector machines (SVMs) to identify multiple Arabic dialects from online and social media data [15] and rule-based systems combined with SVMs for distinguishing MSA from Moroccan colloquial [5]. Additionally, a rule-based system was developed to improve web search effectiveness for Najdi dialect users [16].

A Random Forest (RF) classifier using word and character *n*-gram features, enhanced with Egyptian lexicon-based elements, was developed to differentiate between dialectal Egyptian and MSA in tweets [17]. ML techniques like Linear SVM, Bernoulli Naive Bayes and MNB classifiers were utilized for Arabic DI at the word and sentence levels [7]. It was demonstrated that linear SVM classifiers and meta-classifiers can significantly enhance the accuracy of Arabic DI, marking a shift from broader studies to more detailed approaches [18]. Extending this trend, models were introduced to classify dialects across 25 cities in the Arab World, providing a more detailed examination of regional variations [19].

Deep learning and transfer learning have significantly impacted NLP tasks, including automatic DI [4]. A dataset of tweets from 18 Arabic dialects was developed to train classifiers, including SVM and fine-tuned models like multilingual bidirectional encoder representations from transformers (mBERT) and Arabic BERT (AraBERT), for country-level DI [9]. Arabic DI was explored through a character-level model targeting Saudi dialects using classical ML and a character-based convolutional neural networks (CNN) combined with bidirectional long short-term memory (BiLSTM) [1]. Similarly, a hybrid CNN-BiLSTM model was employed for this task [2]. Other studies implemented LSTM with pre-trained continuous bag-of-words (CBOW) word embeddings for Arabic DI [20, 21], while methods involving LSTM, Word2Vec embeddings and classifiers based on symmetric Kullback-Leibler divergence and MNB were introduced for DI [22].

A range of deep and traditional ML models, including CNN, LSTM, convolutional LSTM (CLSTM), BiLSTM, bidirectional gated recurrent unit (BiGRU) and BiLSTM with attention, were applied for DI [23]. DNN such as CNN-LSTM, CNN, BiLSTM and CLSTM were also utilized for Arabic dialectal text identification [24]. Earlier methods often treated all words with equal importance, without prioritizing word significance uniformly. In contrast, a three-stage neural model was introduced to emphasize dialectal words, enhancing the model's capacity to identify dialect-specific traits more accurately [4].

3. The dataset

The Arabic tweet dataset was systematically compiled by Ref. [10] through Twitter's Streaming application programming interface (API) using the Tweepy Python library (https://www.tweepy.org/), with collection parameters strictly limited to Arabic-language content ('lang:ar' filter). To ensure dialectal purity and regional specificity, a multi-layered filtration protocol was implemented: First, lexical analysis was conducted using comprehensive, peer-

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

reviewed dialect lexicons – the Mojam dictionary for Gulf variants [25] and another list for the Saudi dialect [26] – enabling precise differentiation between MSA and regional vernacular. Second, user profile metadata was rigorously verified, with particular attention to geolocation markers and self-reported location data. Third, stringent activity thresholds were enforced, requiring minimums of 500 historically accumulated tweets per user account and 3+ words per individual tweet to ensure both data richness and linguistic substance.

Following the established protocols of [10], annotation quality control was implemented through a three-tier system: Two primary annotators with formal linguistic training independently classified all samples, while a third senior linguist resolved discrepancies through consensus-building sessions. This rigorous approach specifically addressed the well-documented challenges of crowdsourcing in classification tasks [27], particularly issues of annotator inconsistency and task comprehension that emerge in fine-grained linguistic categorization tasks.

The following datasets were constructed to represent the linguistic diversity of the Gulf region include:

(1) The Gulf Arabic (GAC-6) corpus, which encompasses dialects from six Gulf countries: Saudi Arabia, Bahrain, Kuwait, Oman, Qatar and the Emirates. This dataset includes approximately 1.7 million tweets.

(2) The Saudi Dialect (SDC-5) corpus, which contains around 790,000 tweets that reflect the five main dialects of Saudi Arabia: Hijazi, Najdi, Southern, Northern and Eastern.

Figure 2 illustrates the dialect distribution across the datasets. Specifically, Figure 2a presents the country-wise breakdown for the GAC-6 Corpus, while Figure 2b shows the dialect distribution in the SDC-5 Corpus. Notably, there is a significant imbalance in the number of tweets per dialect. While this reflects real-world linguistic diversity, it may challenge model performance, particularly for underrepresented classes.

4. Dialect identification models

The objective of the DI task is to assign a specific dialect label to a given text [28–30]. There are two primary levels of Arabic DI [31]:

(1) Coarse-grained identification: This approach develops a model to calculate the percentage of dialectal content within a particular sentence.

(2) Fine-grained identification: This method classifies a sentence according to its dialect.

This study treats Arabic DI as a multi-class classification task, emphasizing fine-grained sentence-level identification using an extensive dataset of Arabic Gulf and Saudi dialects.

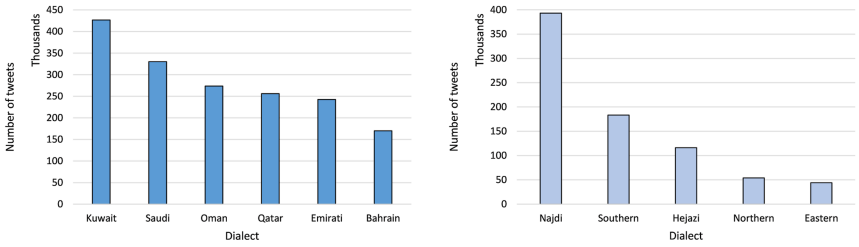


Figure 2. Tweet counts (in thousands) for each dialect in the two corpora. Source: Authors' analysis of GAC-6 and SDC-5 datasets

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

Various ML and deep learning models have been widely employed for DI, each offering distinct advantages and disadvantages [1, 3, 15]. The subsequent section provides an overview of the models employed in the experiments, beginning with classical ML methods and advancing to deep learning approaches and transformer-based algorithms.

4.1 Machine learning-based models

ML is widely applicable across various NLP tasks, showing promise for the DI task [32–34]. To train classifiers, two Python libraries were utilized: Scikit-learn and the Natural Language Toolkit (NLTK). Scikit-learn, an open-source library, offers a vast array of ML algorithms for classification, clustering and regression [35]. NLTK provides comprehensive tools and resources for numerous NLP applications [36].

4.1.1 Learning features. For DI, we employ two lexical feature sets: character *n*-grams (2–5 grams) and word *n*-grams (unigrams to trigrams), denoted as C2–5 and W1–3, respectively. Character *n*-grams capture morphological variations between dialects, such as affixation [33], while word *n*-grams identify dialect-specific words [6, 37]. These features are extracted using Bag-of-Words (BOW) count vectorization and *TF-IDF*. BOW counts *n*-gram frequencies, treating each as an independent feature [38]. *TF-IDF* weights words by their term frequency and inverse document frequency, emphasizing informative but less frequent terms [38]. The *TF-IDF* formula is provided below [38]:

$$W_{x,y} = tf_{x,y} \cdot \log(N/df_x), \tag{1}$$

where $tf_{x,y}$ is the frequency of word x in document y divided by the total words in the corpus, N is the total number of documents and df_x represents the number of documents containing word x.

4.1.2 Training classifiers. We trained supervised models – logistic regression (LR), linear SVM, multinomial Naive Bayes (MNB), RF and an ensemble classifier – on the target dataset (see Section 3) to classify tweets by dialect. LR evaluates word probabilities and uses a logistic function to identify decision boundaries [15, 39], making it effective for text classification [40]. SVM maximizes the margin between labels by finding an optimal hyperplane [15, 41]. MNB calculates label probabilities based on word frequencies, treating features independently [42, 43]. RF builds multiple decision trees to reduce overfitting [44]. The Ensemble model combines classifiers like SGD, MNB and LR using soft voting for superior multi-class performance [7, 18], leveraging individual strengths for better results, as shown in Figure 3.

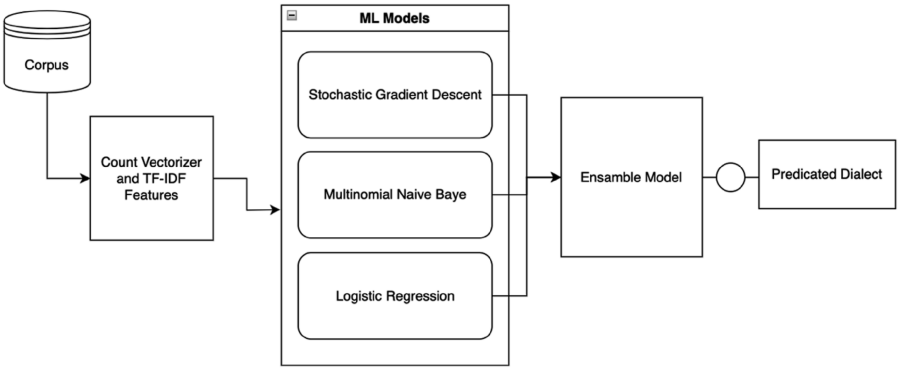


Figure 3. The structure of the ensemble model. Source: Illustration by the authors

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

4.2 Deep learning models

Deep learning models have achieved state-of-the-art results in DI [21, 45]. To identify dialects in tweets, we implemented several deep learning and transformer-based models, including LSTM, BiLSTM and fine-tuned AraBERT. Using the Keras library [46] with TensorFlow [47], we streamlined the development and training of these models, leveraging Keras's intuitive and high-level interface for efficient implementation.

LSTM model, a type of recurrent neural network, excels in handling long-range dependencies and sequential data, making it highly effective for DI [48]. Unlike feed-forward networks, LSTMs preserve sequential integrity, which is crucial for tasks like DI, by using input, forget and output gates to retain relevant information [4, 49]. We employed a word-based LSTM model with an embedding layer using FastText [50] to generate 300-dimensional word embeddings, capturing morphological and semantic variations through CBOW and skip-gram models [51, 52]. The LSTM layer consists of 100 hidden units, a 20% dropout rate and a recurrent dropout of 0.2 to enhance robustness. The model concludes with a softmax layer that classifies inputs into five dialect classes for the Saudi corpus and six for the Gulf corpus, as illustrated in Figure 4.

Bidirectional LSTM (BiLSTM) extends the traditional LSTM by processing sequences in both forward and backward directions, capturing past and future context to improve performance [24]. Building on the word-based LSTM architecture, we used FastText [51] for the embedding layer, leveraging CBOW to capture morphological and semantic nuances. The BiLSTM layer contains 128 hidden units to analyze input sequences bidirectionally, followed by a dense layer with 5,000 units to ensure robust learning and prevent overfitting. The model outputs dialect classifications through a softmax layer.

Fine-tuned AraBERT – BERT [53] and its derivative, AraBERT, have demonstrated exceptional performance in various NLP tasks [54, 55]. AraBERT, pre-trained on Arabic news content, is tailored for Arabic language tasks [56, 57]. For dialect classification, we fine-tuned AraBERT by adding a dense layer and a softmax classifier to predict dialect classes. The model was optimized using the Adam optimizer with a learning rate of 3e−5, trained over five epochs with a batch size of 16. The configuration included a maximum sequence length of 129 tokens and categorical cross-entropy as the loss function.

5. Results analysis and discussion

As detailed in the previous section, the experimental analysis progressed from classic ML methods to the adoption of advanced deep learning models. In this section, we present the performance results of the models discussed. The dataset was partitioned into an 80% training set and a 20% testing set, with the results reported here based on the testing set.

5.1 Evaluation metrics

We evaluated model performance using accuracy, precision, recall and *F*-score metrics, as applied to the corpora in Section 3. Accuracy measures the percentage of correctly classified

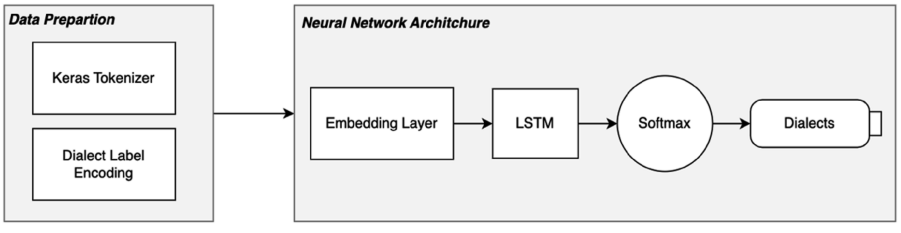


Figure 4. The structure of the LSTM model. Source: Illustration by the authors

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

tweets, while precision, recall and *F*-score provide a comprehensive assessment of dialect prediction effectiveness. In the equations, *TP* denotes true positives, *FP* false positives, *TN* true negatives and *FN* false negatives.

The following standard equations were used to calculate accuracy (*Acc*), precision (*P*), recall (*R*) and *F*-score metrics [58].

$$Acc = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{2}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

$$F\text{-}score = \frac{2P \cdot R}{P + R}. \tag{5}$$

The *F*-score is widely employed as an evaluation metric across multiple NLP tasks including text classification, named entity recognition [59] and extractive summarization [60]. However, for DI tasks, this metric exhibits two well-documented limitations: first, its performance degrades significantly with imbalanced class distributions, which frequently occur when analyzing minority dialects; second, its binary classification heritage makes it unsuitable for complex multi-class dialect scenarios.

These limitations motivate the adoption of the *F*-score's multi-class extensions. The micro-F1 and macro-F1 variants address these issues. Both extensions have proven effective in comparable NLP tasks such as sentiment analysis [61] and question answering systems [62]. For the current DI analysis, we employ macro-F1 due to its demonstrated superiority in handling imbalanced class distributions, a characteristic feature of dialectal datasets.

We adopt macro-F1 as the primary evaluation metric, as it ensures balanced assessment across all dialects. Unlike micro-F1, which is biased toward majority classes, macro-F1 computes the *F*-score for each class independently and averages them, giving equal weight to both rare and dominant dialects. This is particularly critical in DI, where skewed data distributions could otherwise lead to misleading performance estimates.

Consider a scenario where a model is trained to identify two dialects: dialect A, which is well-represented with 950 samples and dialect B, which is comparatively rare, appearing only 50 times. If the model achieves 95% precision and recall for dialect A but only 50% for dialect B, the micro-F1 score remains high at 92.75%, suggesting strong overall performance. However, the macro-F1 score drops significantly to 72.5%, offering a more balanced and revealing assessment. This contrast highlights how micro-F1 can give an inflated sense of accuracy by leaning heavily on majority classes, whereas macro-F1 reflects the true challenges posed by underrepresented dialects in a more equitable manner.

For macro-F1, we calculate the arithmetic mean of the individual *F*-scores of each class (=dialect), as follows:

$$\text{macro-F1} = \frac{\sum F\text{-score of each dialect}}{\text{number of dialects}}. \tag{6}$$

The receiver operating characteristic (ROC) curve plots the *TP* rate against the *FP* rate as the decision threshold varies, making explicit the sensitivity–specificity trade-off; its area under the curve (AUC) condenses performance into a threshold-independent scalar between 0 and 1, where 0.5 indicates random guessing and 1.0 indicates perfect separability. For multi-class DI

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

(e.g. distinguishing Najdi vs. Hijazi Arabic or Kuwaiti vs. Emirati within GAC-6/SDC-5), per-dialect one-vs-rest ROC curves reveal which varieties are most confusable, while macro- and micro-averaged AUC provide complementary corpus-level summaries that are comparatively stable under shifts in class prevalence. High AUC means the model reliably ranks examples from the target dialect above non-target ones, regardless of where the operating threshold is set. ROC analysis also supports principled operating-point selection to reflect application costs (for instance, prioritizing higher recall for minority dialects at acceptable false-alarm rates). Although precision–recall curves can be more informative under extreme imbalance, ROC/AUC remain a rigorous basis for model comparison and ranking, with significance assessable via bootstrap resampling.

5.2 Performance results

Table 1 shows the performance of classical and deep learning models on the GAC-6 and the SDC-5 corpus. Results are reported for the testing set, comprising 20% of the dataset.

We note that among classical models, the ensemble classifier leads with *F*-score of 87.6% on GAC-6 and 88.1% on SDC-5. For deep learning models, fine-tuned AraBERT achieves the highest *F*-score (90.1% on GAC-6 and 90.6% on SDC-5), outperforming BiLSTM. Overall, deep learning models, particularly AraBERT, consistently surpass classical methods across both datasets.

5.3 Comparison with benchmark datasets

To assess the effectiveness of the newly constructed dataset, comparative evaluation was carried out against established benchmark datasets commonly employed in Arabic DI research. The evaluation was based on two publicly available corpora:

(1) The NADI dataset [34], used in the Nuanced Arabic DI shared task, includes 20k training tweets, 4,871 development tweets and covers 18 dialects, with each tweet labeled by dialect.

(2) The QADI dataset [9] is a large, balanced collection of 540k non-genre-specific Arabic dialectal tweets, covering 18 country-level dialects.

Table 2 summarizes the datasets used in the comparative study, while Table 3 evaluates the different models on NADI, QADI and the new corpora using accuracy and macro-F1 metrics.

The corpora achieved higher accuracy and macro-F1 than state-of-the-art datasets like NADI and QADI. The BERT-based model scored macro-F1 of 90.1% and 90.6% on GAC-6 and SDC-5, respectively, but performed lower on NADI (30.8%) and QADI (65.7%). This

Table 1. Performance results of various models on the Gulf Arabic Corpus (GAC-6) and Saudi Dialect Corpus (SDC-5)

Model	GAC-6		SDC-5	
	Acc	macro-F1	*Acc*	macro-F1
Logistic regression	80.8	80.9	86.6	87.9
Support vector classifier	83.6	84.6	85.8	86.7
Multinomial Naive-Bayes	78.8	79.2	82.6	82.4
Random forest	77.5	78.4	80.5	80.8
Ensemble classifier	86.8	87.6	87.7	88.1
LSTM	86.6	88.6	87.4	88.1
BiLSTM	87.8	88.9	88.9	89.4
Fine-tuned AraBERT	89.9	90.1	90.1	90.6

Source(s): Authors' experiment

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

Table 2. Statistics about the datasets used for DI

Datasets	New corpora GAC-6	SDC-5	Benchmark NADI	QADI
Size	1.7M tweets	790K tweets	20K tweets	540K tweets
# of dialects	6	5	18	18
Dialects	Saudi Arabia, Bahrain, Kuwait, Oman, Qatar and United Arab Emirates	Hijazi, Najdi, Southern, Northern and Eastern	Egypt, Iraq, Saudi Arabia, Algeria, Oman, Syria, Libya, Tunisia, Morocco, Lebanon, United Arab Emirates, Kuwait, Jordan, Yemen, Palestine, Bahrain, Bahrain and Sudan	Saudi Arabia, Qatar, Kuwait, Bahrain, Oman, United Arab Emirates, Yemen, Palestine, Egypt, Lebanon, Syria, Jordan, Libya, Iraq, Sudan, Tunisia, Mauritania and Djibouti

Source(s): Authors' compilation of GAC-6, SDC-5, NADI and QADI datasets

Table 3. Results of the comparative evaluation of GAC-6, SDC-5, NADI and QADI datasets

		GAC-6		SDC-5		NADI		QADI	
	Model	*Acc*	macro-F1	*Acc*	macro-F1	*Acc*	macro-F1	*Acc*	macro-F1
Traditional	LR	80.8	80.9	86.6	87.9	36.8	19.1	60.1	56.2
ML-based	SVM	83.6	84.6	85.8	86.7	40.9	19.4	62.5	57.6
(baseline	MNB	78.8	79.2	82.6	82.4	35.3	16.4	59.3	55.8
models)	RF	77.5	78.4	80.5	80.8	33.5	10.1	50.6	49.4
	Ensemble	86.8	87.6	87.7	88.1	45.2	24.4	61.8	58.2
Traditional	LSTM	86.6	88.6	87.4	88.1	47.8	28.8	64.6	64.3
DNN-based	BiLSTM	87.8	88.9	88.9	89.4	46.9	25.3	63.6	63.1
BERT-based	AraBERT	89.9	90.1	90.1	90.6	48.8	30.8	65.8	65.7

Source(s): Authors' experiment

success stems from the dataset's large size and inclusion of smaller labels, enabling the model to generalize more effectively.

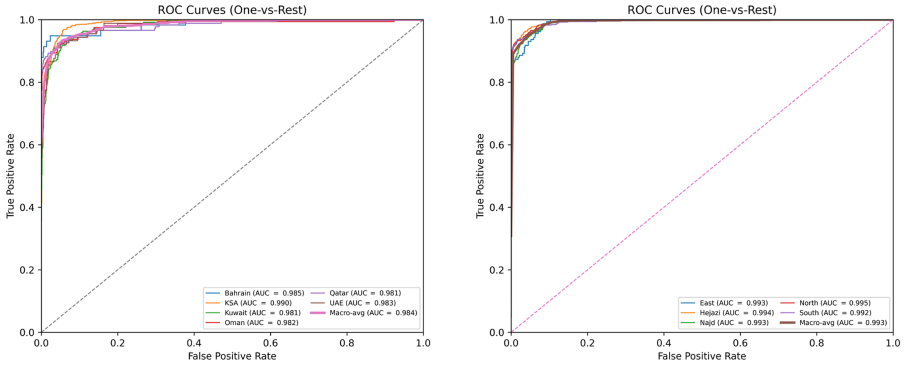
5.4 Discussion

For clarity, we include ROC curves and normalized confusion matrices to illustrate class separability and the specific pairs most often confused. Across both corpora, the ROC curves rise sharply with high AUCs, indicating strong separability between dialect classes: on GAC-6, per-class AUCs range from 0.981 to 0.990; on SDC-5, from 0.992 to 0.995 (Figure 5). The confusion matrices (Figure 6) reveal intuitive, linguistically plausible error patterns. In GAC-6, most classes achieve high recall (KSA ≈ 0.98; Qatar ≈ 0.86; Bahrain ≈ 0.88), while UAE shows the lowest diagonal (≈ 0.75) with its primary confusion toward Kuwait (≈ 0.13). In SDC-5, Hejazi, Najd, North and South exhibit strong diagonals (≈ 0.92-0.95), whereas East is somewhat lower (≈ 0.85) with small spillovers to Hejazi and Najd dialects.

6. Conclusion

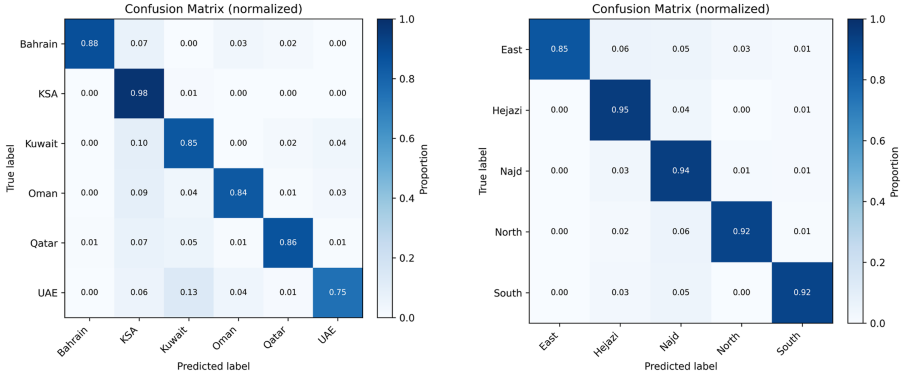
The task of identifying and distinguishing Arabic dialects in text continues to be challenging, primarily due to various factors such as the close relationship between dialects, leading to overlapping vocabularies and a common writing system. In this paper, we presented a framework for identifying Gulf and Saudi dialects by using several ML, deep learning and transformer-based methods. Two datasets were utilized for this task: the Gulf Arabic (GAC-6)

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01



GAC-6 **(a)** SDC-5 **(b)**

Figure 5. The ROC curves for the best model on the dataset: (a) GAC-6 and (b) SDC-5. Source: Authors' experiment



GAC-6 **(a)** SDC-5 **(b)**

Figure 6. The confusion matrix on: (a) GAC-6 and (b) SDC-5, for the best model (transformer). Source: Authors' experiment

corpus with ≈1.7M tweets, covering dialects from Gulf countries and the Saudi Dialect (SDC-5) corpus with ≈790K tweets, focusing on Saudi dialects. Utilizing state-of-the-art models and methodologies, the proposed approach demonstrated strong effectiveness in identifying dialects within the Gulf region and within Saudi Arabia.

Employing a single, fair evaluation protocol, we compared a fine-tuned AraBERT transformer model against representative classical ML baselines – including LR, linear SVM, MNB, RF and a soft-voting ensemble – as well as DNNs (LSTM, BiLSTM). Using the GAC-6 and SDC-5 corpora, AraBERT achieved the strongest results, with macro-F1 scores of 90.1% on GAC-6 and 90.6% on SDC-5. This represents an improvement of +2.5 over the best classical ensemble and +1.2 over the best deep model (BiLSTM), with even wider margins over other classical models: on GAC-6, gains ranged from +5.5 (vs. SVM) to +11.7 (vs. RF) and on SDC-5, from +2.7 (vs. LR) to +9.8 (vs. RF). To assess generalizability, the same models were evaluated on public benchmarks (NADI and QADI), where the transformer continued to achieve state-of-the-art performance.

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

Future work will focus on improving the DI framework's robustness through hyperparameter tuning, alternative weighting strategies and additional transfer learning models. We also plan to explore advanced data balancing techniques (e.g. synthetic minority over-sampling technique (SMOTE), undersampling and XGBoost) and data augmentation to mitigate class imbalance and enhance generalization.

References

1. Alqurashi T. Applying a character-level model to a short Arabic dialect sentence: a Saudi dialect as a case study. Appl Sci. 2022; 12(23): 12435. doi: 10.3390/app122312435.
2. Hedhli M, Kboubi F. CNN-BiLSTM model for Arabic dialect identification. In: International Conference on Computational Collective Intelligence. Springer; 2023. p. 213-25.
3. Shatnawi MQ, Yasin MB, Huq AA. Building a framework for identifying Arabic dialects using deep learning techniques. ACM Trans Asian Low-Resour Lang Inf Process. 2023: 3630632. doi: 10.1145/3630632.
4. Mohammed A, Jiangbin Z, Murtadha A. A three-stage neural model for Arabic dialect identification. Comput Speech Lang. 2023; 80: 101488. doi: 10.1016/j.csl.2023.101488.
5. Tachicart R, Bouzoubaa K, Aouragh SL, Jaafa H. Automatic identification of Moroccan colloquial Arabic. In: Arabic Language Processing: From Theory to Practice: 6th International Conference, ICALP 2017, Fez, Morocco. Springer; 2017. p. 201-14.
6. Wray S. Classification of closely related sub-dialects of Arabic using support-vector machines. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); 2018.
7. Lichouri M, Abbas M, Freihat AA, Megtouf DEH. Word-level vs sentence-level language identification: application to Algerian and Arabic dialects. Procedia Comput Sci. 2018; 142: 246-53. doi: 10.1016/j.procs.2018.10.484.
8. Habash NY. Introduction to Arabic natural language processing. Morgan & Claypool Publishers; 2022.
9. Abdelali A, Mubarak H, Samih Y, Hassan S, Darwish K. QADI: Arabic dialect identification in the wild. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop; 2021. p. 1-10.
10. Al-Shenaifi N, Azmi AM, Hosny M. Advancing AI-driven linguistic analysis: developing and annotating comprehensive Arabic dialect corpora for Gulf countries and Saudi Arabia. Math. 2024; 12(19): 3120. doi: 10.3390/math12193120.
11. Azmi AM, Aljafari EA. Modern information retrieval in Arabic–catering to standard and colloquial Arabic users. J Inf Sci. 2015; 41(4): 506-17. doi: 10.1177/0165551515585720.
12. Obeid O, Salameh M, Bouamor H, Habash N. ADIDA: automatic dialect identification for Arabic. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations); 2019. p. 6-11.
13. Elfardy H, Diab M. Sentence level dialect identification in Arabic. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2. Short Papers; 2013. p. 456-61.
14. Sadat F, Kazemi F, Farzindar A. Automatic identification of Arabic dialects in social media. In: Proceedings of the First International Workshop on Social Media Retrieval and Analysis, in SoMeRA '14, New York, NY. Association for Computing Machinery; 2014. p. 35-40.
15. Cotterell R, Callison-Burch C. A multi-dialect, multi-genre corpus of informal written Arabic. In: LREC; 2014. p. 241-5.
16. Azmi AM, Aljafari EA. Universal web accessibility and the challenge to integrate informal Arabic users: a case study. Universal Access Inf Soc. 2018; 17(1): 131-45. doi: 10.1007/s10209-017-0522-3.

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

17. Darwish K, Sajjad H, Mubarak H. Verifiably effective Arabic dialect identification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1465-8.
18. Malmasi S, Refaee E, Dras M. Arabic dialect identification using a parallel multidialectal corpus. In: International Conference of the Pacific Association for Computational Linguistics. Springer; 2015. p. 35-53.
19. Salameh M, Bouamor H, Habash N. Fine-grained Arabic dialect identification. In: Proceedings of the 27th International Conference on Computational Linguistics; 2018. p. 1332-44.
20. Sayadi K, Hamidi M, Bui M, Liwicki M, Fischer A. Character-level dialect identification in Arabic using long short-term memory. In: Gelbukh A (Ed.) Computational linguistics and intelligent text processing, vol. 10762 of lecture notes in computer science. Cham: Springer International Publishing; 2018. p. 324-37.
21. Issa E, AlShakhori M, Al-Bahrani R, Hahn-Powell G. Country-level Arabic dialect identification using RNNs with and without linguistic features. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop; 2021. p. 276-81.
22. Harrat S, Meftouh K, Abidi K, Smaïli K. Automatic identification methods on a corpus of twenty five fine-grained Arabic dialects. In: Smaïli K (Ed.) Arabic language processing: from theory to practice, vol. 1108 of communications in computer and information science. Cham: Springer International Publishing; 2019. p. 79-92.
23. Elaraby M, Abdul-Mageed M. 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); 2018. p. 263-74.
24. Lulu L, Elnagar A. Automatic Arabic dialect classification using deep learning models. Procedia Comput Sci. 2018; 142: 262-9. doi: 10.1016/j.procs.2018.10.489.
25. Mo3jam: Dictionary of colloquial Arabic/Arabic Slang 2023. Available from: https://en.mo3jam.com/ [accessed 3 September 2023].
26. Aldrsoni S. Dictionary in Saudi local dialects (in Arabic). Available from: https://lahajat.blogspot.com/p/blog-page_7.html
27. Almuzaini HA, Azmi AM. An unsupervised annotation of Arabic texts using multi-label topic modeling and genetic algorithm. Expert Syst Appl. 2022; 203: 117384. doi: 10.1016/j.eswa.2022.117384.
28. Zaidan O, Callison-Burch C. 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; 2011. p. 37-41.
29. AlShenaifi N, Azmi A. Arabic dialect identification using machine learning and transformer-based models. In: The Seventh Arabic Natural Language Processing Workshop (WANLP 2022); 2022.
30. Alshutayri A, Atwell E. Classifying Arabic dialect text in the social media Arabic dialect corpus (SMADC). In: Proceedings of the 3rd Workshop on Arabic Corpus Linguistics; 2019. p. 51-9.
31. Kwaik KA, Saad M, Chatzikyriakidis S, Dobnik S. Shami: a corpus of levantine Arabic dialects. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA); 2018.
32. Mekki AE, Mahdaouy AE, Berrada I, Khoumsi A. AdaSL: an unsupervised domain adaptation framework for Arabic multi-dialectal sequence labeling. Inf Process Manag. 2022; 59(4): 102964. doi: 10.1016/j.ipm.2022.102964.
33. Elnagar A, Yagi SM, Nassif AB, Shahin I, Salloum SA. Systematic literature review of dialectal Arabic: identification and detection. IEEE Access. 2021; 9: 31010-42. doi: 10.1109/access.2021.3059504.
34. Abdul-Mageed M, Zhang C, Bouamor H, Habash N. NADI 2020: the first nuanced Arabic dialect identification shared task, ArXiv Preprint ArXiv:201011334, 2020.
35. Pedregosa F, *et al*. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12: 2825-30.
36. Loper E, Bird S, NLTK: the natural language toolkit, arXiv Preprint arXiv:cs/0205028. 2002; 1: 63-70. doi: 10.3115/1118108.1118117,

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

37. Adouane W, Semmar N, Johansson R, Bobicev V. Automatic detection of Arabicized Berber and Arabic varieties. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3); 2016. p. 63-72.
38. Elmitwally NS, Alsayat A. Classification and construction of Arabic corpus: figurative and literal. J Theor Appl Inf Technol; 98(19): 3252-3264.
39. Shah K, Patel H, Sanghvi D, Shah M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment Hum Res. 2020; 5: 1-16. doi: 10.1007/s41133-020-00032-0.
40. Pranckevičius T, Marcinkevičius V. Comparison of Naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Balt J Mod Comput. 2017; 5(2). doi: 10.22364/bjmc.2017.5.2.05.
41. Tarmom T, Teahan W, Atwell E, Alsalka M. Compression vs traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. J Nat Lang Eng; 6(6): 663-676.
42. Su J, Shirab JS, Matwin S. Large scale text classification using semi-supervised multinomial Naive Bayes. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11); 2011. p. 97-104.
43. Nayel H, Hassan A, Sobhi M, El-Sawy A. Machine learning-based approach for Arabic dialect identification. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop; 2021. p. 287-90.
44. Hussein S, Farouk M, Hemayed E. Gender identification of Egyptian dialect in Twitter, Egypt. Inform J. 2019; 20(2): 109-16. doi: 10.1016/j.eij.2018.12.002.
45. Qwaider C, Chatzikyriakidis S, Dobnik S. Can modern standard Arabic approaches be used for Arabic dialects? Sentiment analysis as a case study. In: Proceedings of the 3rd Workshop on Arabic Corpus Linguistics; 2019. p. 40-50.
46. Gulli A, Pal S. Deep learning with Keras. Packt Publishing; 2017.
47. Abadi M, *et al.* TensorFlow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16); 2016. p. 265-83.
48. Graves A. Long short-term memory. In: Supervised sequence labelling with recurrent neural networks. Berlin, Heidelberg: Springer; 2012. p. 37-45.
49. Luan Y, Lin S. Research on text classification based on CNN and LSTM. In: 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE; 2019. p. 352-5.
50. Joulin A, Grave E, Mikolov PBT. Bag of tricks for efficient text classification. In: EACL 2017; 2017. p. 427-31.
51. Athiwaratkun B, Wilson A, Anandkumar A. Probabilistic FastText for multi-sense word embeddings. In: Gurevych I, Miyao Y (Eds). Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1. Long Papers; 2018.
52. Alghamdi N, Assiri F. A comparison of fastText implementations using Arabic text classification. In: Intelligent systems and applications, vol. 1038 of advances in intelligent systems and computing. Springer International Publishing; 2020. p. 306-11.
53. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding.
54. Alhussain A, Azmi AM. Beyond event-centric narratives: advancing Arabic story generation with large language models and beam search. Math. 2024; 12(10): 1548. doi: 10.3390/math12101548.
55. Talafha B, *et al.*, Multi-dialect Arabic Bert for country-level dialect identification, arXiv Preprint arXiv:200705612, 2020.
56. Antoun W, Baly F, Hajj H. Arabert: transformer-based model for Arabic language understanding, arXiv Preprint arXiv:200300104, 2020.
57. Alammary AS. BERT models for Arabic text classification: a systematic review. Appl Sci. 2022; 12(11): 5720. doi: 10.3390/app12115720.

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01

58. Flach P, Kull M. Precision-recall-gain curves: PR analysis done right. Adv Neural Inf Process Syst; 2015; 28: 838-846.

59. Balla H, Delany SJ. Exploration of approaches to Arabic named entity recognition.

60. Azmi A, Al-Thanyyan S. Ikhtasir–a user selected compression ratio Arabic text summarization system. In: 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering; 2009. p. 1-7. doi: 10.1109/nlpke.2009.5313732.

61. Azmi AM, Alzanin SM. Aara'–a system for mining the polarity of Saudi public opinion through e-newspaper comments. J Inf Sci. 2014; 40(3): 398-410. doi: 10.1177/0165551514524675.

62. Alwaneen TH, Azmi AM. Stacked dynamic memory-coattention network for answering why-questions in Arabic. Neural Comput Appl. 2024; 36(15): 8867-83. doi: 10.1007/s00521-024-09525-0.

Corresponding author
Aqil M. Azmi can be contacted at: aqil@ksu.edu.sa

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

Downloaded from http://ftp.nowpublishers.com/aci/article-pdf/doi/10.1108/ACI-02-2025-0067/10668308/aci-02-2025-0067en.pdf by guest on 01