Arabic Dialect Identification Using Transformer-Based Models

Yousef Adel*, Mostafa Dorrah*, Ahmed Ashraf*, Abdallah ElSaadany*, Mohamed Tarek*
School of Information Technology and Computer Science (ITCS),
Nile University, Giza, Egypt
{Y.Khalil, M.Samer, Ahme.Ashraf, A.Elsaadany, M.TarekSaad}@nu.edu.eg

***Abstract*—The identification of Arabic dialects using deep learning models is a rapidly evolving field in computational linguistics, driven by the increasing digital communication in the Arab world. Arabic dialect identification is crucial due to the significant linguistic variations between dialects and Modern Standard Arabic (MSA). MSA, the formal language of media and education across the Arab world, differs considerably from colloquial dialects in terms of syntax, vocabulary, and pronunciation. These dialects, which are often region-specific, are prevalent in everyday communication and social media, presenting a unique challenge for natural language processing systems. In this paper, we compare 5 different transformer-based models on a specific Arabic dialect corpus and analyze the results of each of them.**

***Index Terms*—Arabic Text, Transformer, Dialectal Text**

I. Introduction

Recent advancements in deep learning have led to the development of sophisticated models for dialect identification. Architectures like BERT (Bidirectional Encoder Representations from Transformers) and its variants have demonstrated remarkable efficacy in understanding and classifying Arabic dialects. These models are pre-trained on large datasets encompassing diverse Arabic texts, enabling them to capture the nuances of different dialects effectively.

The importance of Arabic dialect identification extends beyond linguistic interest. It plays a pivotal role in enhancing communication technologies, from machine translation to voice recognition systems, specifically tailored for the Arab world. Accurate dialect identification improves user experience and facilitates better understanding and engagement with Arabic-speaking audiences, making it a significant area of research in natural language processing and computational linguistics.

II. Related Work

In this section, existing and previous work related to the proposed work in this paper will be mentioned.

The authors of [1] focus on identifying Arabic dialects using NLP techniques. The problem statement revolves around the complexity of dialect identification due to the diversity of Arabic dialects across different countries, cities, and towns. The methodology includes applying multiple algorithms for dialect identification, starting from noise removal to classification, utilizing Naïve Bayes, Logistic Regression, and Decision Tree classifiers combined through voting. They also employed clustering to reduce noise from Modern Standard Arabic tweets in the training phase. The results showed varying f-measure scores for different methodologies, with the highest being 52.38% for the second methodology with clustering. The study acknowledges limitations such as the presence of non-Arabic tweets in the dataset and challenges in distinguishing closely related dialects.

The document [2] addresses the challenge of identifying Arabic dialects from textual data. The problem arises due to the diversity of Arabic dialects and the prevalence of dialectal Arabic in informal online communications. The methodology involves using a dataset known as Arabic Online Commentary (AOC) and implementing four deep neural network models—LSTM, CNN, BLSTM, and CLSTM—for binary and ternary classification of three main Arabic dialects (Egyptian, Levantine, and Gulf). The results showed promising performance, with varying accuracy across different models and dialect pairs. A limitation highlighted is the potential inconsistency in the AOC dataset, emphasizing the need for its thorough revision to enhance the reliability of dialect classification.

The paper [3] explores the challenge of distinguishing between Egyptian Arabic (ARZ) and Modern Standard Arabic (MSA) at the sentence level. It introduces a linear classification model using binary feature functions and a linear support-vector machine (SVM) approach. The dataset used is the Arabic Online Commentary (AOC), and a 10-fold stratified cross-validation method is employed. The system achieved an accuracy of 89.1%, marking a 1.3% absolute improvement over previous results. The research highlights the potential impact of sentence informality on classification accuracy and discusses limitations, such as the decrease in accuracy when informal sentences are involved.

The authors of [4] present an approach for enhancing Arabic dialect classification using semi-supervised learning. It involves training multiple classifiers with weakly, strongly supervised, and unsupervised data. The study demonstrates significant improvement in dialect classification accuracy on two test sets, showing a 5% improvement over the strongly supervised classifier and 20% over the weakly supervised classifier. The research also explores the application of an improved dialect classifier in building a Modern Standard Arabic language model for machine translation, resulting in a model size reduction of 70% and a 0.6 BLEU point increase in translation quality. The paper acknowledges the challenges in

Arabic dialect classification, particularly in social media text, due to informal language use and dialectal variations.

The paper [5] explores Arabic dialect classification using neural networks. The authors test various neural network models, including Multi-Input CNN, CNN-biLSTM, and binary classification CNN-biLSTM, on the task of distinguishing Modern Standard Arabic from Egyptian, Gulf, Levantine, and North African dialects. Their best-performing system is a Multi-Input CNN, yielding a macro-averaged F1 score of 0.5289. They note that acoustic features significantly influence classification accuracy. The study indicates challenges with datasets' limited size for neural networks and the complexity of Arabic dialect identification.

III. Methodology

In this section, we will present our methodology, and the methodology consists of 2 main components: 1- Dataset pre-processing, and 2- few-shot learning using BERT-based models.

A. Data Pre-Processing and Analysis

We selected the [6] dataset, which comprises texts and their corresponding dialects, specifically 18 different country-level dialects as shown in figure 1.

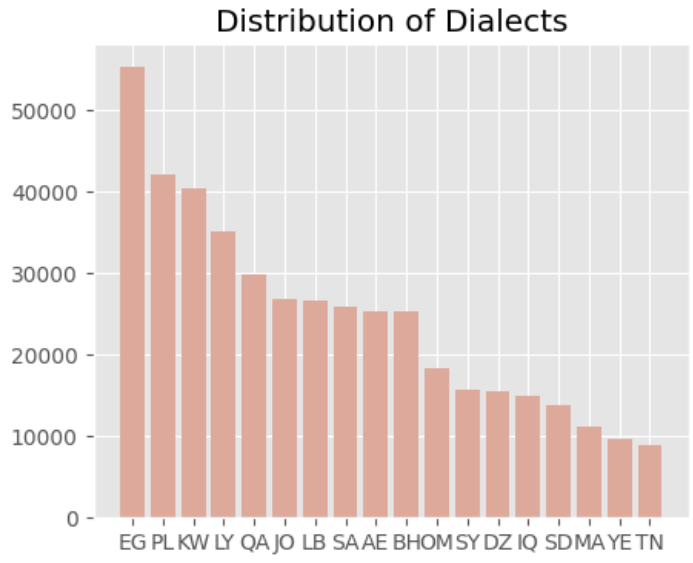

Fig. 1.

To minimize data redundancy, we eliminated duplicate entries. A standardization technique was employed, transforming all instances of ي to ى among other modifications, and we also removed punctuation as well as the hashtags "@" and similar redundant characters. Most importantly, stop-words removal was applied to remove the redundancy of each text record. This preprocessing was applied across the entire dataset, including training, validation, and testing phases. Additionally, we converted labels like "EG" into numerical representations, such as "3," to enhance model prediction accuracy. Moreover, due to the high number of classes, we grouped certain dialects into the same class. The data exhibited class imbalance, as depicted in the subsequent figure 1. We handled the imbalance in the corpus by taking the minimum number of class instances and reducing the number of other class instances to the number of the minimum class instances. Also, we took a sample of only 55% due to computation and time limitations. As shown in Table I, the dialects transformed from 18 classes to 5 labels based on the vocabulary similarity, to minimize the number of distinct classes and make it easier for the model to predict.

Dialect	Group
LB, SY, PL, JO	0
SA, AE, QA, KW, YE, BH, IQ, OM	1
EG, SD	2
MA, DZ, TN	3
LY	4

TABLE I
DIALECT AND ITS CORRESPONDING GROUP

B. Transfomer-based Models

We experimented with various transformer-based models on this corpus to evaluate and compare their performance.

The first model, "MARBERTv2," [7] is an extensive pre-trained masked language model tailored for both Dialectal Arabic (DA) and MSA. It underwent training on a random sample of 1 billion Arabic tweets from a larger dataset containing approximately 6 billion tweets. The architecture of "MARBERTv2" mirrors that of BERT-base, with 12 attention layers, each featuring 12 attention heads and 768 hidden dimensions. It includes a vocabulary of 100,000 WordPieces and comprises roughly 163 million parameters. The maximum text length processed by its tokenizer is 60.

The second model we utilized is "AraBERTv2," [8] an Arabic pre-trained language model developed based on Google's BERT architecture. This model adheres to the BERT-Base configuration, maintaining consistency with the first model in terms of model and tokenizer hyperparameters. This alignment in configuration underscores the systematic approach employed in our comparative analysis of different language models.

The third model in our study is termed "Multi-Dialect-Bert-Base-Arabic." [9] This model commenced with the foundational weights of "Arabic-BERT" and underwent subsequent fine-tuning using a corpus of 10 million Arabic tweets. These tweets were sourced from the unlabeled data segment of the Nuanced Arabic Dialect Identification (NADI) shared task, offering a robust foundation for dialect analysis and model enhancement.

The fourth model utilized in our research is "bert-base-arabic-camelbert-mix," [10] a BERT-based model specifically pre-trained on a diverse array of Arabic texts. This model encompasses a blend of both dialectal Arabic and Modern Standard Arabic (MSA), enabling it to effectively process and understand a wide range of textual sizes and variants within the Arabic language spectrum.

Model	Training Loss	Validation Loss	Accuracy Score	Precision Score	Recall Score	F1 Score
MARBERTv2 [7]	0.436	0.609	**0.794**	**0.730**	**0.778**	**0.749**
AraBERTv2 [8]	0.789	0.750	0.726	0.656	0.695	0.671
Multi-Dialect-Bert-Base-Arabic [9]	0.620	0.693	0.753	0.688	0.725	0.702
bert-base-arabic-camelbert-mix [10]	0.552	0.665	0.763	0.695	0.745	0.714
AraBERTv02-base-twitter [11]	0.606	0.660	0.767	0.697	0.742	0.715

TABLE II
MODELS EVALUATION COMPARISON

The concluding model in our analysis is designated as "AraBERTv02-base-twitter." [11] This model, tailored for Arabic dialects and tweets, underwent an extended pre-training phase using the Masked Language Model (MLM) task on approximately 60 million Arabic tweets, which were selectively curated from a larger pool of 100 million tweets. "AraBERTv02-base-twitter" adheres to the standard BERT-Base architecture and configurations, demonstrating its alignment with established deep learning frameworks for language processing.

IV. EXPERIMENTAL RESULTS

The experiment was done for this study on Kaggle online free computational resources (GPU T4 x2) and applied few-shot learning for five epochs only for each model due to time and computational constraints. The following Table II illustrates the results of the five different models, including training/validation loss of the last training epoch, the accuracy score, precision score, recall score, and F1 score on the held-out-set to fairly evaluate the model's performance on the dataset. We used multiple classification metrics to accurately evaluate the model's performance on the given dataset.

As shown in the Table II above, the five models achieved similar results. However, MARBERTv2 [7] model achieved slightly better results than its competitors. This could be justified because this model is mainly pre-trained on numerous Arabic tweets which contain various types of dialects with the existence of the MSA. So, the model performed better due to the similarity between the data pre-trained on and the data applied few-shot learning on.

V. CONCLUSION AND FUTURE WORK

Our study focuses on identifying Arabic dialects using deep learning. It compares five transformer-based models on an Arabic dialect corpus, evaluating their performance in dialect classification. The study highlights the linguistic diversity of Arabic dialects and their challenge for natural language processing systems. Experimental results show similar performance across models, with the "MARBERTv2" model slightly outperforming others, likely due to its extensive pre-training on diverse Arabic tweets. This study contributes significantly to enhancing Arabic dialect identification, crucial for various communication technologies. In the future, it is expected to compare other models, try other pre-processing techniques, and evaluate more datasets.

ETHICS STATEMENT

There is no ethical conflict in our research.

REFERENCES

[1] A. Aliwy, H. Taher, and Z. AboAltaheen, "Arabic dialects identification for all Arabic countries," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (I. Zitouni, M. Abdul-Mageed, H. Bouamor, F. Bougares, M. El-Haj, N. Tomeh, and W. Zaghouani, eds.), (Barcelona, Spain (Online)), pp. 302–307, Association for Computational Linguistics, Dec. 2020.

[2] L. Lulu and A. Elnagar, "Automatic arabic dialect classification using deep learning models," *Procedia Computer Science*, vol. 142, pp. 262–269, 2018. Arabic Computational Linguistics.

[3] C. Tillmann, S. Mansour, and Y. Al-Onaizan, "Improved sentence-level arabic dialect classification," pp. 110–119, 01 2014.

[4] F. Huang, "Improved arabic dialect classification with social media data," pp. 2118–2126, 01 2015.

[5] E. Michon, M. Q. Pham, J. Crego, and J. Senellart, "Neural network architectures for Arabic dialect identification," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)* (M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, S. Malmasi, and A. Ali, eds.), (Santa Fe, New Mexico, USA), pp. 128–136, Association for Computational Linguistics, Aug. 2018.

[6] A. Rezk, "Arabic dialect identification," 2022. data retrieved from Hugging Face, https://huggingface.co/datasets/Abdelrahman-Rezk/Arabic$_{D}ialect_{I}dentification$.

[7] UBC-NLP, "Marbertv2," 2022. model retrieved from Hugging Face, https://huggingface.co/UBC-NLP/MARBERTv2.

[8] aubmindlab, "bert-base-arabertv2," 2023. model retrieved from Hugging Face, https://huggingface.co/aubmindlab/bert-base-arabertv2.

[9] bashar talafha, "Multi-dialect-bert-base-arabic," 2021. model retrieved from Hugging Face, https://huggingface.co/bashar-talafha/multi-dialect-bert-base-arabic.

[10] CAMeL-Lab, "bert-base-arabic-camelbert-mix," 2021. model retrieved from Hugging Face, https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix.

[11] aubmindlab, "bert-base-arabertv02-twitter," 2023. model retrieved from Hugging Face, https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix.