

Project Title

Arabic Dialect and Country Identification Using Deep Learning and Explainable Artificial Intelligence (XAI)

Background

Arabic is one of the most widely spoken languages in the world, with millions of speakers distributed across different countries and regions. Although Modern Standard Arabic serves as a common written language, people communicate daily using various dialects that differ significantly in vocabulary, pronunciation, and expressions. These dialectal variations create challenges for Natural Language Processing (NLP) applications such as sentiment analysis, machine translation, chatbots, and speech recognition systems.

Automatic dialect identification has become an important research area because accurately identifying the dialect of a text can significantly improve the performance of language technologies. Recent advances in Artificial Intelligence (AI), particularly Deep Learning models such as BERT-based architectures, have shown promising results in dialect classification tasks. However, these models often operate as black boxes, making it difficult to understand the reasons behind their predictions.

To address this challenge, Explainable Artificial Intelligence (XAI) techniques can be employed to provide transparency and interpretability, enabling researchers and users to understand which linguistic features contribute to classification decisions. Therefore, combining AI with XAI offers both high performance and improved trustworthiness in dialect identification systems.

Problem Statement

Arabic dialects vary considerably across countries and regions, making automatic identification a challenging task. Traditional language processing approaches often struggle to distinguish between closely related dialects due to shared vocabulary and linguistic similarities. Furthermore, many deep learning models provide accurate predictions but lack transparency, making it difficult to understand the reasoning behind their decisions. Therefore, there is a need for an intelligent and explainable system capable of accurately identifying Arabic dialects and countries while providing interpretable explanations for its predictions.

Project Objectives

Main Objective

Develop an AI-based system capable of identifying Arabic dialects and countries from textual data and explaining its predictions using Explainable AI techniques.

Specific Objectives

1. Collect and analyze a large-scale Arabic dialect dataset.
2. Perform data preprocessing and text cleaning for Arabic social media content.
3. Build baseline machine learning models such as Logistic Regression, Naïve Bayes, and SVM.
4. Develop advanced deep learning models using MARBERT and AraBERT.
5. Classify Arabic texts according to country-level dialects across 18 Arab countries.
6. Group countries into major dialect categories (Gulf, Levantine, Egyptian, Maghrebi, Iraqi, Yemeni, and Sudanese).
7. Evaluate model performance using Accuracy, Precision, Recall, and F1-score.
8. Apply Explainable AI techniques such as SHAP and LIME to interpret model decisions.
9. Compare traditional machine learning models with transformer-based models.
10. Develop a transparent and reliable dialect identification framework that can support future Arabic NLP applications.

Data Analysis

Dataset Overview

This project utilizes the **QADI (Arabic Dialect Identification Dataset)**, a large-scale Arabic dialect dataset collected from Twitter. The dataset contains Arabic tweets from users across 18 Arab countries, representing different regional dialects. The dataset is designed specifically for Arabic dialect identification and text classification tasks.

Dataset Statistics

The dataset consists of a total of **458,197 Arabic tweets** distributed into three subsets:

Dataset Split Number of Samples

Training Set 440,052

Dataset Split Number of Samples

Validation Set 9,164

Test Set 8,981

Total 458,197

The dataset includes **18 dialect classes**, each corresponding to a specific Arab country.

Dataset Features

The dataset contains the following attributes:

Feature Description

id Unique identifier for each tweet

label Numerical dialect label

text Arabic tweet content

country Country/dialect name (added during preprocessing)

Dialect Classes

The dataset covers the following Arabic dialects:

- Oman
- Sudan
- Saudi Arabia
- Kuwait
- Qatar
- Lebanon
- Jordan
- Syria
- Iraq
- Morocco
- Egypt

- Palestine
- Yemen
- Bahrain
- Algeria
- United Arab Emirates (UAE)
- Tunisia
- Libya

Data Quality Assessment

An initial exploratory analysis was conducted using Python and Jupyter Notebook. The analysis revealed the following findings:

- The dataset contains **18 unique dialect classes**.
- The dataset contains **458,197 Arabic tweets**.
- No missing values were detected in any feature.
- All records contain valid tweet text and corresponding dialect labels.
- The dataset is suitable for machine learning and deep learning applications without requiring major data reconstruction.

Initial Data Exploration Results

The exploratory analysis showed that the dataset is rich in dialectal Arabic text collected from real-world social media interactions. The large number of samples and dialect diversity provide a strong foundation for training robust Artificial Intelligence models for Arabic dialect identification.

The dataset contains informal language, user mentions, URLs, hashtags, emojis, and other social media-specific elements. Therefore, a preprocessing stage will be performed before model training to clean and normalize the text.

Planned Data Preprocessing

Before training the classification models, the following preprocessing steps will be applied:

- Removing URLs and hyperlinks.
- Removing user mentions (@username).

- Removing emojis and special symbols.
- Removing unnecessary punctuation.
- Arabic text normalization.
- Tokenization using an Arabic language model tokenizer.
- Preparing the data for deep learning models such as MARBERT or AraBERT.

Summary

The conducted analysis confirms that the selected dataset is sufficiently large, diverse, and well-structured for Arabic dialect classification. The absence of missing values and the availability of 18 dialect categories make it highly suitable for developing and evaluating Artificial Intelligence and Explainable AI (XAI) approaches for Arabic dialect detection.

Marwa Mohammed AL-Mahmoud

Nouf Ayed Alaklabi

Norah homoud alharbi