Dialect Identification

Meet The Team

Nawal

Hadeer

Nada

Nadin

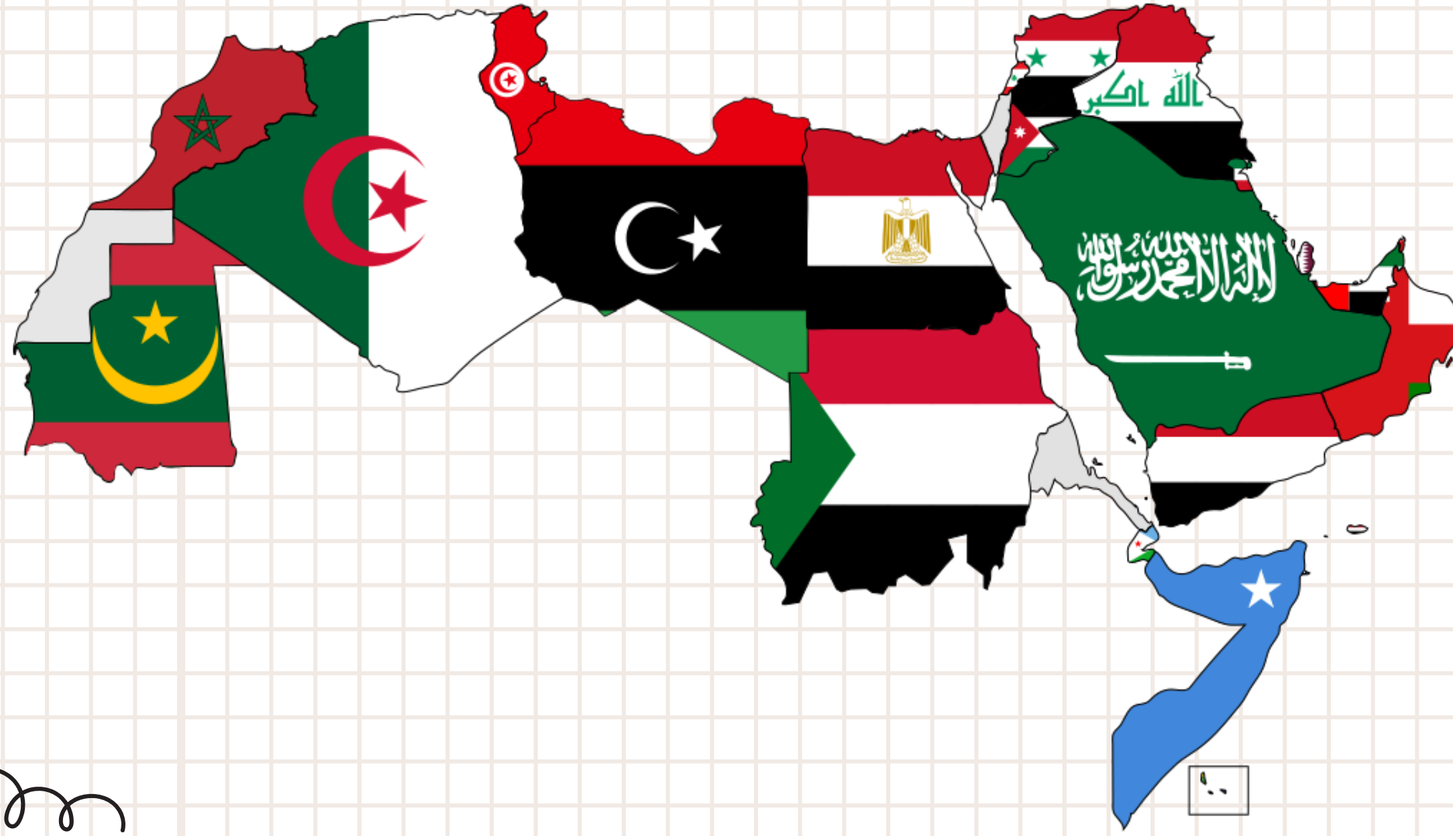
Nagham

Table of content

- Problem
- Cleaning and Preprocessing
- Model Architecture

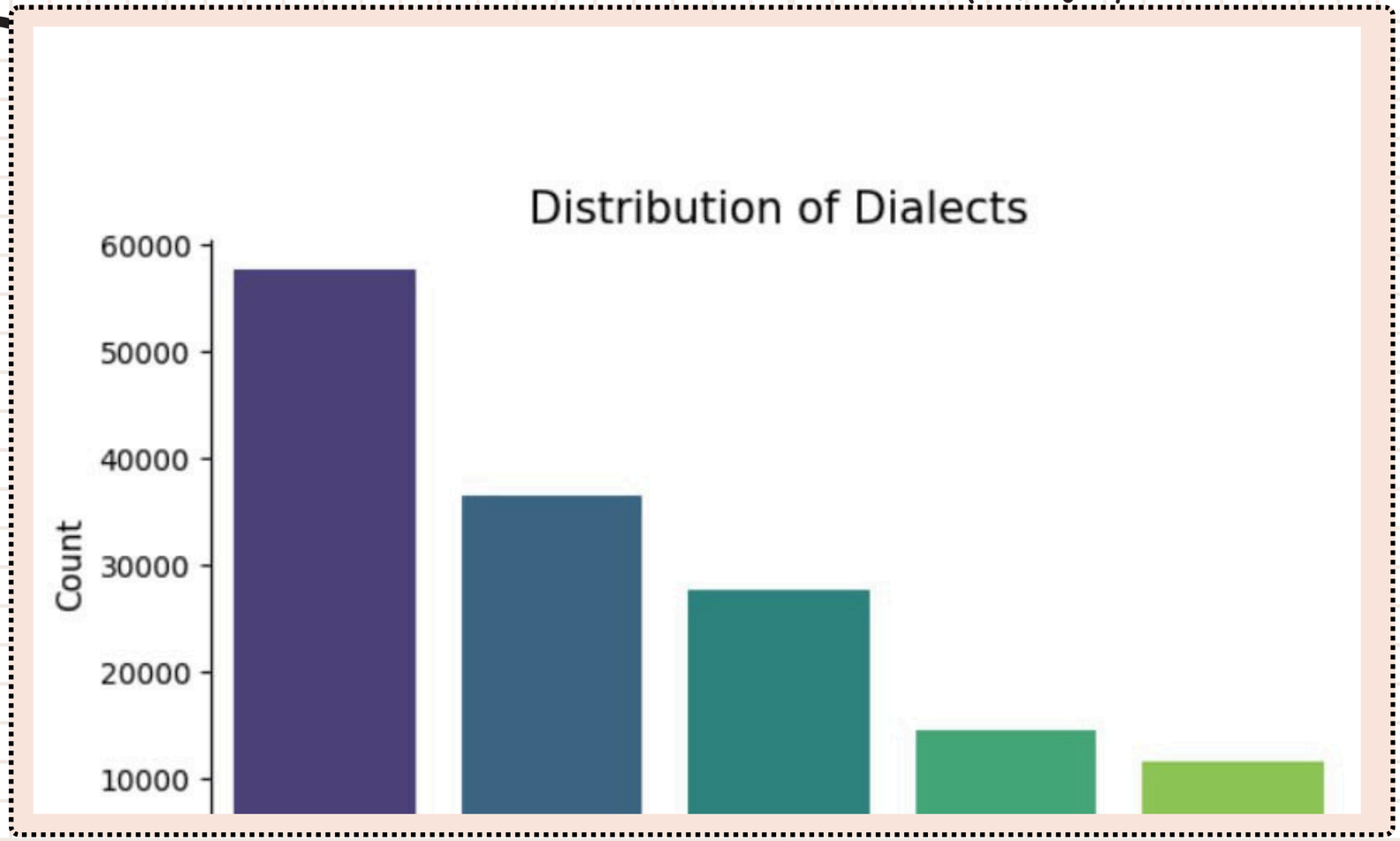
Project Overview

Arabic NLP faces challenges due to the rich variety in word usage and letter formation across 22 Arab countries, each with distinct dialects. This project explores methodologies using machine learning and deep learning models to accurately identify and classify various Arabic dialects, advancing precision in Arabic computational linguistics.



Dataset

The QADI dataset is an automatically collected collection of 540k tweets from 2,525 users, covering 18 country-level Arabic dialects in the Middle East and North Africa. The dataset is curated by filtering user account descriptions and removing tweets in Modern Standard Arabic or with inappropriate language. Evaluation shows 91.5% accuracy in labeling and a macro-averaged F1-score of 60.6% for country-level dialect identification across 18 classes.



Visualize the Distribution of Dialects

Cleaning & Preprocessing

Data is preprocessed by these steps:

- Elements Removed: Emojis ,Links , Mentions, Numbers, Punctuation, Extra Whitespaces, Repeated Characters, English Characters.

- Normalize Characters

text
😂😂 دانتي بتهزري بقى بإست الكل @HnMaroco
هتوجع دماغك بالأونطه اللى حصل انتهى وخلاص هناك اخطاء يجب علاجها @swisii
مع ايلا ما\nالواحد شايل دولارانو وسعر ثابت مافي لا عارف يتخارج منها ويخسر لا عارف يصبر ممكن تاني 😄😄😄 انشاء الله الدولار يصل 3ج\nفيش مستحيل
عشمنا قلوبنا باللي فات ونسينا نعيش !
هي اسودة كده ليه مش باين ان فيها حاجة حلوة دي بقت سوااااد خالص
...
الجزر الثلاثه ماتجيب طاريها ليه ي خيركم في اخوانكم @smralmazrooei1
صار لازم تحط صورتك 🙂 — ما عم افضى اتصوّر. بعدين انتو الحق https://t.co/WVvk5yQipw.
غايتو نو شعار حلوه القطن اللي لعبو بيه #ريال_مدريد اليوم .
الي بمشي معادش بولي باحنه @Asssssss93
شن تقدر ادير بعني السعودية ..نمر من ورق @anwarmalek

ML Model

Logistic regression with TF-IDF vectorizer

	precision	recall	f1-score	support
0	0.84	0.87	0.86	11484
1	0.82	0.84	0.83	5578
2	0.80	0.75	0.77	7268
3	0.69	0.72	0.70	2283
4	0.68	0.67	0.68	2932
accuracy			0.80	29545
macro avg	0.77	0.77	0.77	29545
weighted avg	0.80	0.80	0.80	29545

NN Model

LSTM

```
accr = model.evaluate(X_test,y_test)
print('Test set\n  Loss: {:0.3f}\n  Accuracy: {:0.3f}'.format(accr[0],accr[1]))
```

```
462/462 [==============================] - 30s 64ms/step - loss: 0.5068 - acc: 0.8323
Test set
  Loss: 0.507
  Accuracy: 0.832
```

Thank You