Arabic Dialect Classification

classifying Arabic tweets dialect

Our Team:

Names
Amr Mohamed Abd ALbadee
Amira Hesham Mo’men
Ibrahim Ayman Abu-Shara
Mostafa Mohammed ali
Ziad Mahmoud Mohammed

Git Repo

Introduction

In this presentation, We will discuss project focused on developing a machine learning model that can accurately classify tweets written in different Arabic dialects.

01.

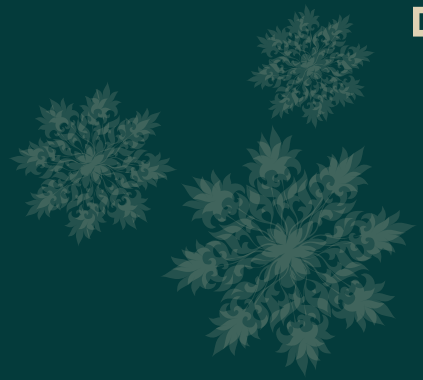
Data & preprocessing

واحد

Retrieve data

We used library named “**sqlite3**” as interface for SQLite databases .

Cleaning data



01. Remove urls

02. Remove user mentions

03. Remove punctuation

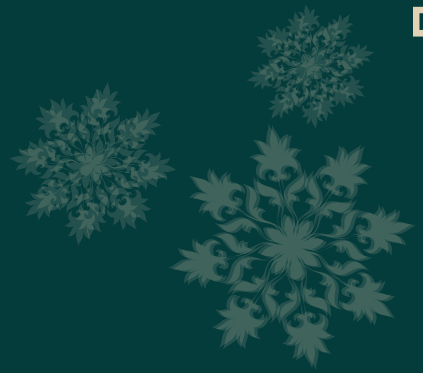
04. Remove numbers,emojis and

Extra Cleaning for ML

01. Remove repeated characters

02. Remove stop words

03. Remove tashkeel ◌ً◌ٌ◌ٍ◌َ◌ُ◌ِ◌ّ◌ْ



02.

Train ML models

اثنان

Train models

We built our pipeline to make the process faster

1. Clean_data
2. Vectorizer_tfidf
3. Evaluate_model
4. Save_model

1. Naive bayern (MultinomialNB)

- Model Evaluation (الحمدلله انه ما غرق)

```
Accuracy: 0.6653799268590004
F1 macro: 0.5056611076652915
              precision    recall  f1-score   support

          EG       0.56      1.00      0.72      5762
          LB       0.96      0.62      0.75      2762
          LY       0.85      0.58      0.69      3648
          MA       1.00      0.16      0.27      1154
          SD       1.00      0.05      0.09      1440

    accuracy                           0.67     14766
   macro avg       0.87      0.48      0.51     14766
weighted avg       0.78      0.67      0.62     14766
```

1. Linear SVC

- Model Evaluation

```
Accuracy: 0.8365840444263849
F1 macro: 0.8047681802874983
              precision    recall  f1-score   support

          EG       0.83      0.93      0.88      5762
          LB       0.86      0.86      0.86      2762
          LY       0.83      0.81      0.82      3648
          MA       0.87      0.68      0.77      1154
          SD       0.82      0.62      0.70      1440

    accuracy                           0.84     14766
   macro avg       0.84      0.78      0.80     14766
weighted avg       0.84      0.84      0.83     14766
```

03.

Train DL models

ثلاثة

1. Vanilla RNN

- Model Evaluation (يخساره وقتي الي مرنتك فيه)

```
462/462 [==============================] - 4s 8ms/step - loss: 1.3998 - accuracy: 0.4395
462/462 [==============================] - 4s 9ms/step
Accuracy: 0.4395180344581604
F1 macro: 0.20083966406445555
              precision    recall  f1-score   support

           0       0.46      0.93      0.61      5764
           1       0.37      0.41      0.39      2762
           2       0.00      0.00      0.00      3650
           3       0.00      0.00      0.00      1154
           4       0.00      0.00      0.00      1443

    accuracy                           0.44     14773
   macro avg       0.17      0.27      0.20     14773
weighted avg       0.25      0.44      0.31     14773
```

1. LSTM

- Model Evaluation

```
462/462 [==============================] - 2s 5ms/step - loss: 0.2927 - accuracy: 0.7867
462/462 [==============================] - 3s 4ms/step
Accuracy: 0.786705493927002
F1 macro: 0.746936223754875
              precision    recall  f1-score   support

           0       0.84      0.87      0.85      5764
           1       0.80      0.81      0.80      2762
           2       0.75      0.76      0.76      3650
           3       0.73      0.64      0.68      1154
           4       0.68      0.61      0.64      1443

    accuracy                           0.79     14773
   macro avg       0.76      0.74      0.75     14773
weighted avg       0.78      0.79      0.79     14773
```