



Arabic Dialect Sentimental Analysis

Prepared by : Hazem Mohamed Hosny

CONTENTS

The background features several wavy, flowing lines in black and red, creating a dynamic and artistic feel. These lines sweep across the page, partially framing the content.

01

Introduction

02

DataSet

03

Proposed Methodology

04

Metrics and Results

05

Deployment

06

Conclusion

01 Introduction

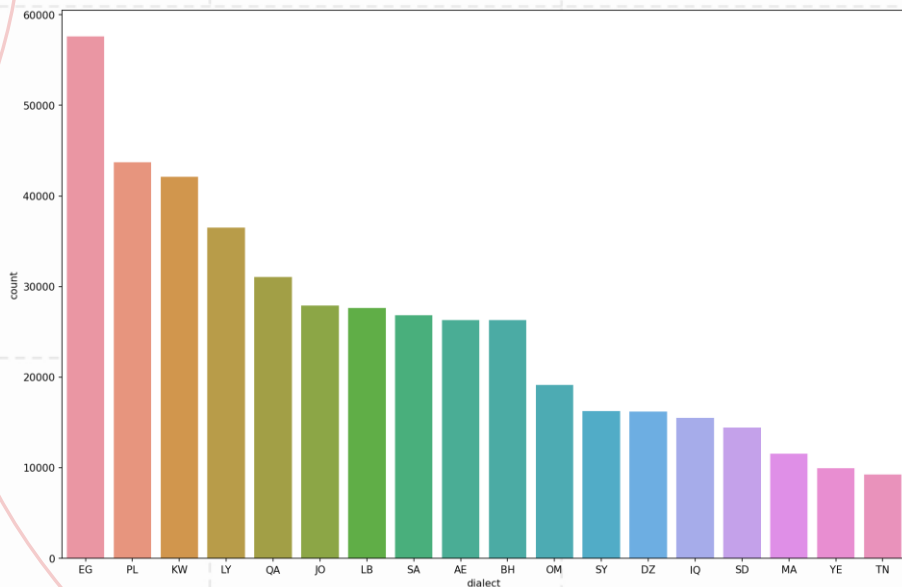
As the official language of 22 countries spread across the Middle-East North Africa (MENA) region, Arabic is the 4th most used language on the Internet. Statistics from 2018 show 164 million internet users in the Middle East and 121 million internet users in North Africa.

As a language, Arabic has complex morphology and various dialects. The complexity increases significantly when considering the informal nature of social-media text and the distinction between Modern Standard Arabic (MSA) and Dialectal Arabic (DA). Further complicating matters, there are numerous dialects, for example the Egyptian dialect is different from the Levantine dialect which is used in Palestine, Jordan, Syria, and Lebanon. Both of these dialects are also distinct from the Gulf dialect used in Kuwait, Bahrain, Qatar, and the United Arab Emirates.



02 DataSet

Our Dataset contains 458197 rows where labels are imbalanced. “EG” is the most major label, Our as “TN” is least label in occurrence. With a distribution as shown in the figure.



	id	dialect	text
0	1175358310087892992	IQ	@Nw8ieIUwaCAArE ٢ يتنفذ .. لكن بالنهاية ..
1	1175416117793349632	IQ	@7zNqXP0yrODdRjK ح ٢ يعني هذا محسوب على البشر ..
2	1175450108898565888	IQ	@KanaanRema مبين من كلامه خليجي
3	1175471073770573824	IQ	@HAIDER76128900 🌸 يسلملي مرورك وروحك الحلوه
4	1175496913145217024	IQ	@hmo2406 🌸 وين هل العيبه اخ محمد
...
458192	1019484980282580992	BH	@Al_mhbaa_7 🌸 ميسوطين منك اللي باسطانا
458193	1021083283709407232	BH	@Zzainabali @P_ameerah والله ماينده ابش يختي
458194	1017477537889431552	BH	@Al_mhbaa_7 ...شو عملنا لك حنا تهربي منا احنا مس
458195	1022430374696239232	BH	@haneenalmwla 🌸 الله يبارك فيها وبالعافيه
458196	1022409931029458944	BH	@jolnar121 🌸 السحله ضيفي ي يتطلع لك سحليه

458197 rows × 3 columns

03 Proposed Methodology

1. PreProcessing “text” which includes:

- Remove URLs
- Remove RT and cc
- Remove @username
- Remove new line in both Windows or Apple
- Clean hashtags
- Clean emojis where translating native emojis (such as: ‘:’) , ‘:(’, ...etc to words), and remove emojis belongs to emojis library or other utf-8 format.
- Remove punctuations
- For Arabic normalizations: remove stopwords, normalize Arabic letters (such as: transform عى to ي, transform ة to ه, etc...), Using pyarabic Package to remove Tashkel, and finally remove longation.

	id	dialect	text	word_count	char_count	avg_char_per_word	stopwords	emoji_count
0	1175358310087892992	IQ	بالنهاية ينتفض يعير	3	19	5.666667	0	0
1	1175416117793349632	IQ	يعني محسوب البشر حيونه ووحشيه وتطلبون العرب يحترمكم ويمن بدينكم ولاينعتكم بالارهاب	12	83	6.000000	0	0
2	1175450108898565888	IQ	مبين كلامه خليجي	3	16	4.666667	0	0
3	1175471073770573824	IQ	يسلملي مروك وروحك الطوه	4	25	5.500000	0	0
4	1175496913145217024	IQ	وين العيبه اخ محمد	4	18	3.750000	0	0

03 Proposed Methodology

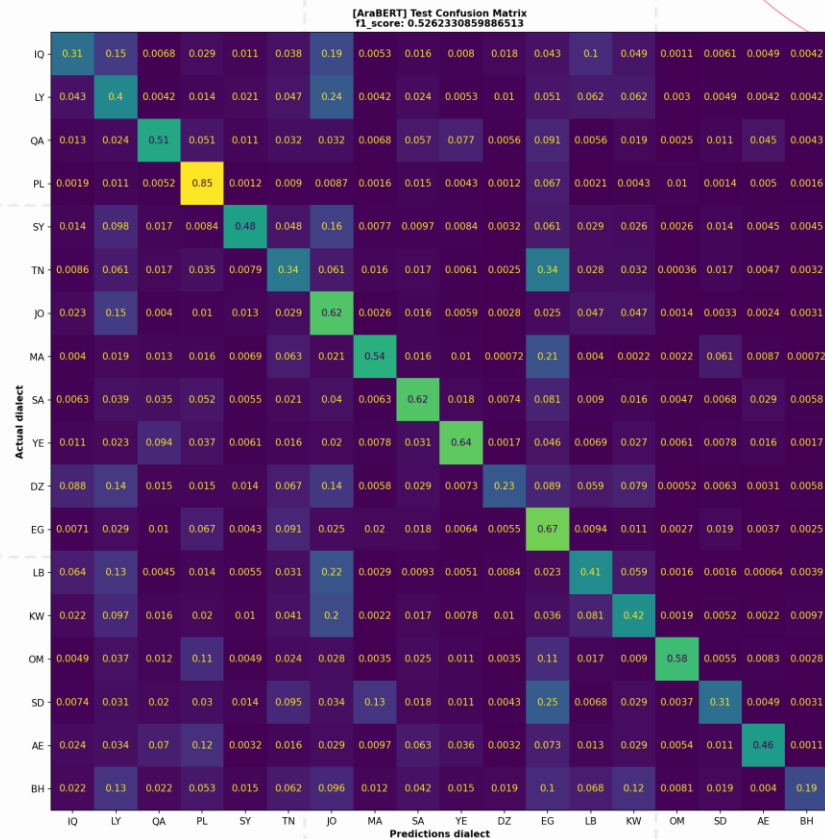
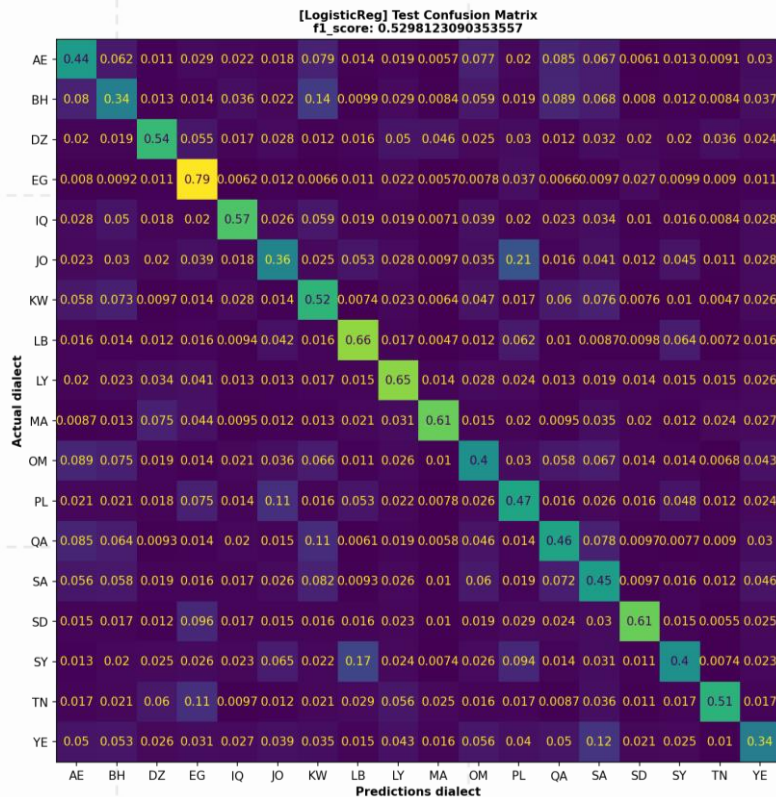
2. For ML Models:

- TF-IDF Vectorizer with unigram word analyzer extraction.
- A Stratified Train test split is considered to be more fit for unbalanced data for more generalized validation score.
- A Lasso Logistic Regression were tested along the TF-IDF Vectorizer on the cleaned data and with class weight to overcome unbalance data
- Complement Naive Bayes classifier were tested too. It is particularly suited for imbalanced data sets.
- Stochastic gradient descent Classifier (SGDClassifier) with Hinge loss as parameter that performs as Linear SVM were tested with class weights setted.

3. For Deep Learning Model:

- An AraBERT version 2 model were tested from “aubmindlab/bert-base-arabertv02” without FarasaSegmenter over the original data with “ArabertPreprocessor()” object, offered Transformers Package over Pytorch.

04 Metrics and Results



05 Deployment

In this section, 5 main libraries were used :

1. scikit-learn for tfidf vectorizer and Logistic regression Model
2. Pyarabic for striping tashkeel
3. Nltk for Arabic stop words
4. Transformers from Hugging Face to add our AraBERT model to the pipeline to predict dialect of text.
5. Fast API used to implement our webserver app



HUGGING FACE



default

GET / Home

POST /Predict_ml Arabic Tweet Dialect Prediction (Logistic Regression Model)

POST /Predict Arabic Tweet Dialect Prediction (AraBERT best label)

POST /Predict_all_labels Arabic Tweet Dialect Prediction with (AraBERT all labels)

05 Deployment

POST /Predict_ml Arabic Tweet Dialect Prediction (Logistic Regression Model)

Parameters

Name	Description
tweet_paragraph * required	عفكره هول المحطين وقت الانتخابات ما تلقنا حشائكم شي .
string (query)	

Execute

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:5000/Predict_ml?tweet_paragraph=%40karmakhayat%20%D8%B9%KD9%81%D9%83%D8%B1%D8%A9%20%D9%87%84%D8%A7%D9%86%D8%AA%D8%AE%D8%A7%D8%A8%D8%A7%D8%AA%20%D9%85%D8%A7%20%D8%A4%D9%81%D9%86%D8%A7%20%D8%B9%D8%A8%D9%85%20%D8%A7%D9%84%D8%A7%D8%B9%D8%A7%D9%8A%D8%A9%20%D8%A7%D9%84%D9%85%D8%A7%D9%81%D9%88%D8%A8%D8%A9%8A%D9%86%20%D8%A6%D8%B9%D8%B1%D8%B3%D9%83%20%D8%A7%D9%86%D8%AA%D9%8A%20%D9%83%D8%A7%D9%86%20%D9%81%D9%8F%99%82'
```

Request URL

```
http://127.0.0.1:5000/Predict_ml?tweet_paragraph=%40karmakhayat%20%D8%B9%KD9%81%D9%83%D8%B1%D8%A9%20%D9%87%D8%84%D8%A7%D9%86%D8%AA%D8%AE%D8%A7%D8%A8%D8%A7%D8%AA%20%D9%85%D8%A7%20%D8%A4%D9%81%D9%86%D8%A7%20%D8%B9%D8%A8%D9%85%20%D8%A7%D9%84%D8%A7%D8%B9%D8%A7%D9%8A%D8%A9%20%D8%A7%D9%84%D9%85%D8%A7%D9%81%D9%88%D8%A8%D8%A9%8A%D9%86%20%D8%A6%D8%B9%D8%B1%D8%B3%D9%83%20%D8%A7%D9%86%D8%AA%D9%8A%20%D9%83%D8%A7%D9%86%20%D9%81%D9%8F%99%82
```

Server response

Code	Details
200	Response body <pre>{ "label": "LB", "score": 0.335316545582381 }</pre>

POST /Predict Arabic Tweet Dialect Prediction (AraBERT best label)

Parameters

Name	Description
tweet_paragraph * required	عفكره هول المحطين وقت الانتخابات ما تلقنا حشائكم شي .
string (query)	

Execute

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:5000/Predict?tweet_paragraph=%40karmakhayat%20%D8%B9%KD9%81%D9%83%D8%B1%D8%A9%20%D9%87%84%D8%A7%D9%86%D8%AA%D8%AE%D8%A7%D8%A8%D8%A7%D8%AA%20%D9%85%D8%A7%20%D8%A4%D9%81%D9%86%D8%A7%20%D8%B9%D8%A8%D9%85%20%D8%A7%D9%84%D8%A7%D8%B9%D8%A7%D9%8A%D8%A9%20%D8%A7%D9%84%D9%85%D8%A7%D9%81%D9%88%D8%A8%D8%A9%8A%D9%86%20%D8%A6%D8%B9%D8%B1%D8%B3%D9%83%20%D8%A7%D9%86%D8%AA%D9%8A%20%D9%83%D8%A7%D9%86%20%D9%81%D9%8F%99%82'
```

Request URL

```
http://127.0.0.1:5000/Predict?tweet_paragraph=%40karmakhayat%20%D8%B9%KD9%81%D9%83%D8%B1%D8%A9%20%D9%87%D8%84%D8%A7%D9%86%D8%AA%D8%AE%D8%A7%D8%A8%D8%A7%D8%AA%20%D9%85%D8%A7%20%D8%A4%D9%81%D9%86%D8%A7%20%D8%B9%D8%A8%D9%85%20%D8%A7%D9%84%D8%A7%D8%B9%D8%A7%D9%8A%D8%A9%20%D8%A7%D9%84%D9%85%D8%A7%D9%81%D9%88%D8%A8%D8%A9%8A%D9%86%20%D8%A6%D8%B9%D8%B1%D8%B3%D9%83%20%D8%A7%D9%86%D8%AA%D9%8A%20%D9%83%D8%A7%D9%86%20%D9%81%D9%8F%99%82
```

Server response

Code	Details
200	Response body <pre>{ "label": "LB", "score": 0.280286664867481 }</pre> Response headers <pre>content-length: 43 content-type: application/json date: Tue, 18 Mar 2022 00:31:42 GMT server: uvicorn</pre>



THANKS!

Do you have any
questions?