Arabic Dialect Sentimental Analysis

Prepared by :   Hazem Mohamed Hosny

CONTENTS

01

Introduction

02

DataSet

03

Proposed Methodology

04

Metrics and Results

05

Deployment

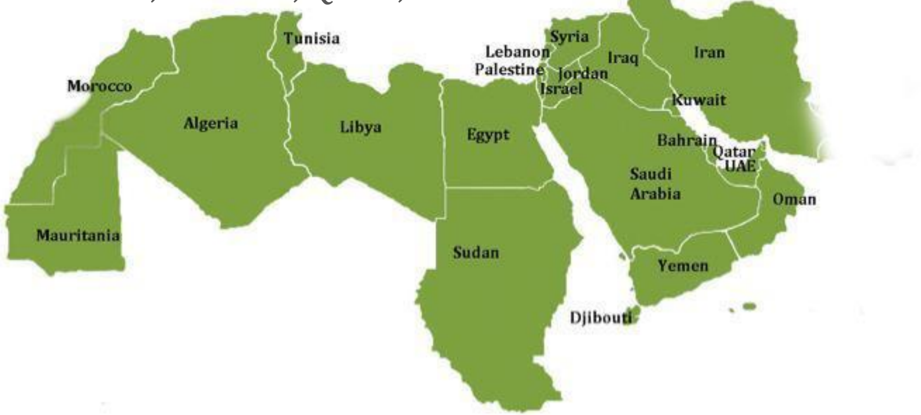
06

Conclusion

01 Introduction

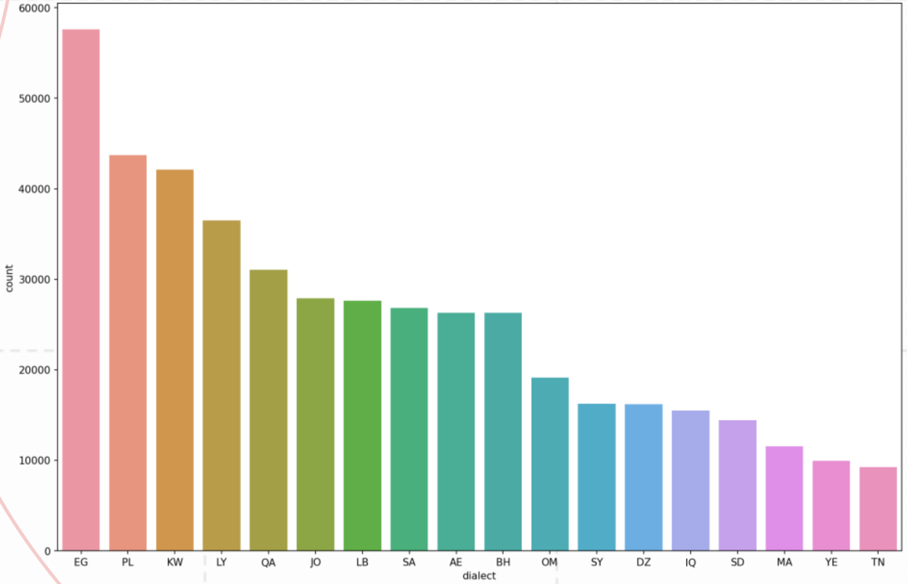
As the official language of 22 countries spread across the Middle-East North Africa (MENA) region, Arabic is the 4th most used language on the Internet. Statistics from 2018 show 164 million internet users in the Middle East and 121 million internet users in North Africa.

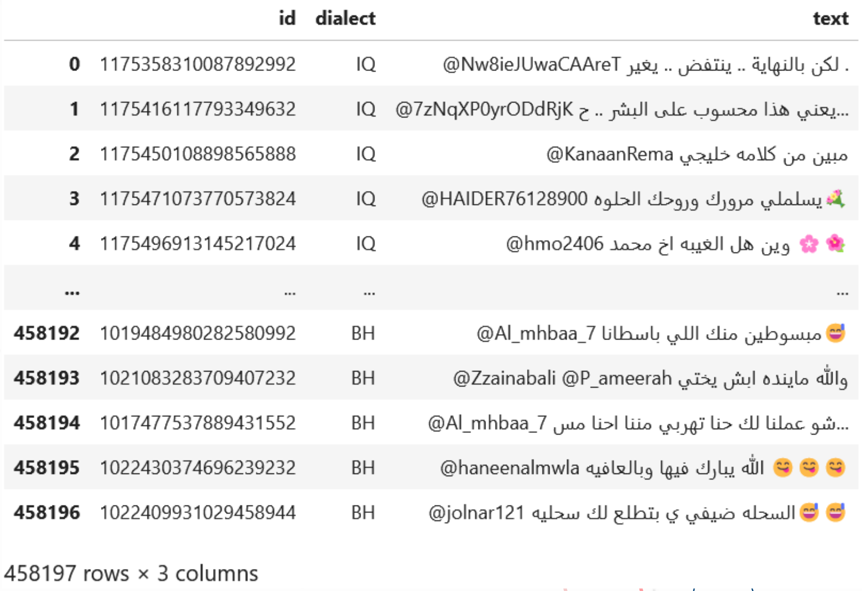
As a language, Arabic has complex morphology and various dialects. The complexity increases significantly when considering the informal nature of social-media text and the distinction between Modern Standard Arabic (MSA) and Dialectical Arabic (DA). Further complicating matters, there are numerous dialects, 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.





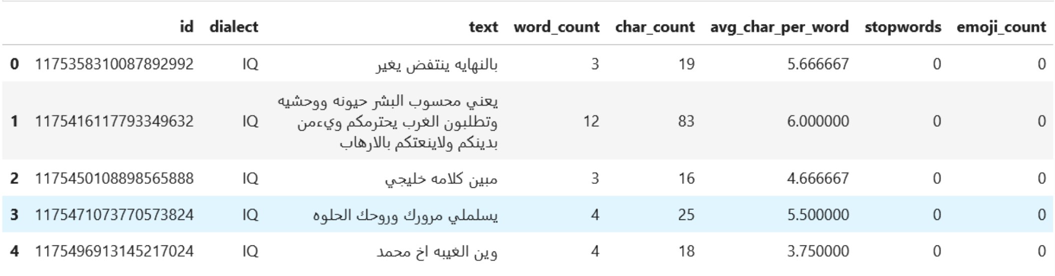
	id	dialect	text
0	1175358310087892992	IQ	@Nw8ieJUwaCAAreT لكن بالنهاية .. ينتفض .. يغير .
1	1175416117793349632	IQ	@7zNqXP0yrODdRjK يعني هذا محسوب على البشر .. ح...
2	1175450108898565888	IQ	@KanaanRema مبين من كلامه خليجي
3	1175471073770573824	IQ	@HAIDER76128900 💐يسلملي مرورك وروحك الحلوه
4	1175496913145217024	IQ	@hmo2406 🌺 🌸 وين هل الغيبه اخ محمد
...	...	...	...
458192	1019484980282580992	BH	@Al_mhbaa_7 😅مبسوطين منك اللي باسطانا
458193	1021083283709407232	BH	@Zzainabali @P_ameerah والله ماينده ابش يختي
458194	1017477537889431552	BH	@Al_mhbaa_7 شو عملنا لك حنا تهربي منتا احنا مس...
458195	1022430374696239232	BH	@haneenalmwla 😋 😋 😋 الله يبارك فيها وبالعافيه
458196	1022409931029458944	BH	@jolnar121 😅 😅السحله ضيفي ي بتطلع 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 ى 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	بالنهايه ينتفض يغير	3	19	5.666667	0	0
1	1175416117793349632	IQ	يعني محسوب البشر حيونه ووحشيه وتطلبون الغرب يحترمكم ويءمن 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 مرورك وروحك الحلوه	4	25	5.500000	0	0
4	1175496913145217024	IQ	وين الغيبه اخ محمد	4	18	3.750000	0	0

03 Proposed Methodology

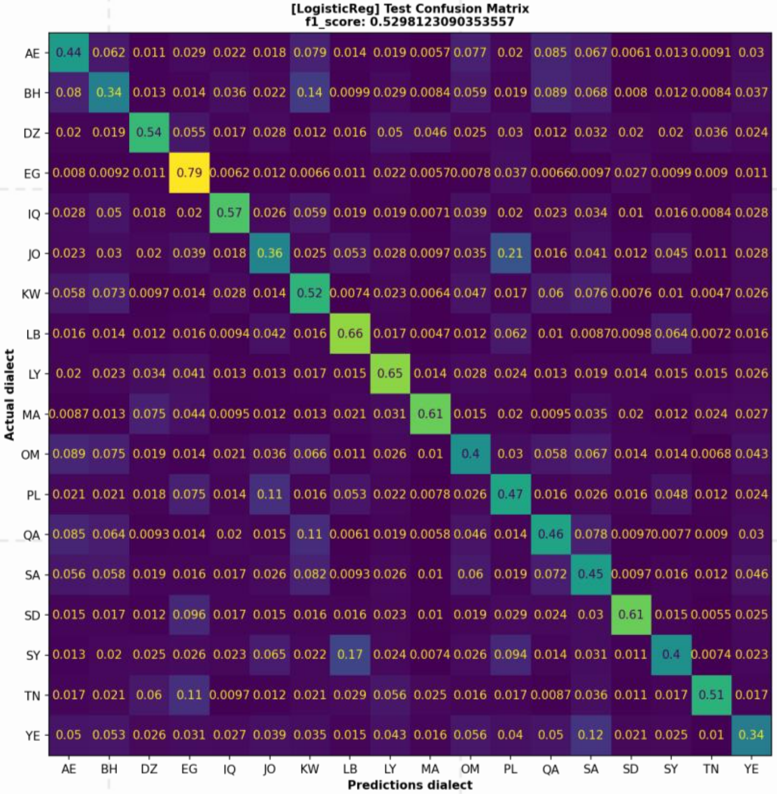
2. For ML Models:

- 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:

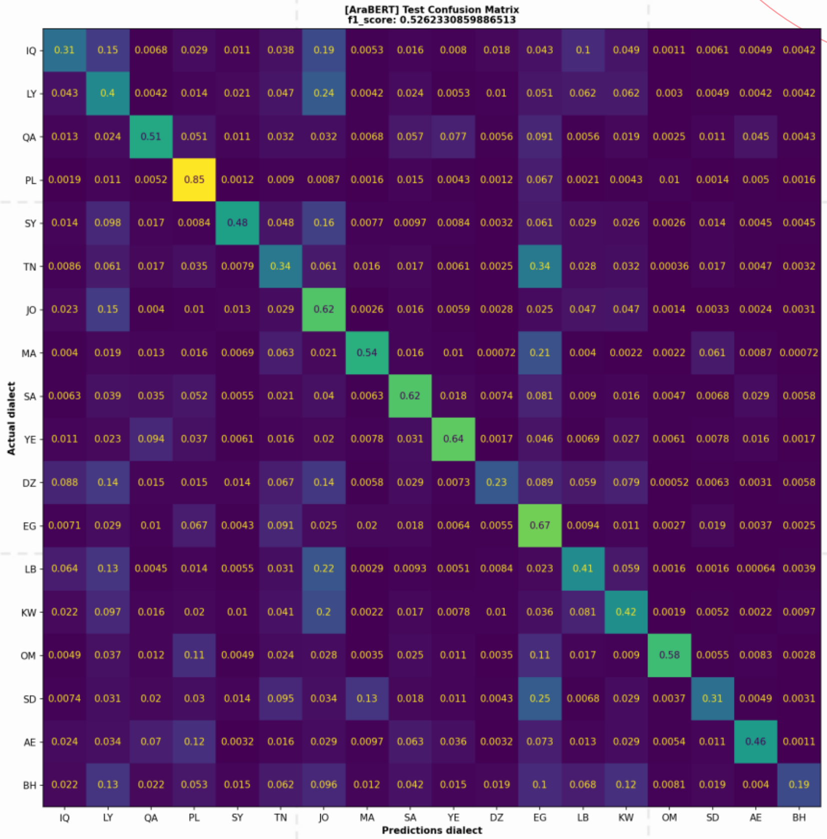
- An AraBERT version 2 model were tested from "aubmindlab/bert-base-arabertv02" without FarasaSegmenter over the original data with "ArabertPreprocessor()" object, offered Transformers Package over Pytorch.

04 Metrics and Results



[LogisticReg] Test Confusion Matrix
f1_score: 0.5298123090353557

	AE	BH	DZ	EG	IQ	JO	KW	LB	LY	MA	OM	PL	QA	SA	SD	SY	TN	YE
AE	0.44	0.062	0.011	0.029	0.022	0.018	0.079	0.014	0.019	0.0057	0.077	0.02	0.085	0.067	0.0061	0.013	0.0091	0.03
BH	0.08	0.34	0.013	0.014	0.036	0.022	0.14	0.0099	0.029	0.0084	0.059	0.019	0.089	0.068	0.008	0.012	0.0084	0.037
DZ	0.02	0.019	0.54	0.055	0.017	0.028	0.012	0.016	0.05	0.046	0.025	0.03	0.012	0.032	0.02	0.02	0.036	0.024
EG	0.008	0.0092	0.011	0.79	0.0062	0.012	0.0066	0.011	0.022	0.0057	0.0078	0.037	0.0066	0.0097	0.027	0.0099	0.009	0.011
IQ	0.028	0.05	0.018	0.02	0.57	0.026	0.059	0.019	0.019	0.0071	0.039	0.02	0.023	0.034	0.01	0.016	0.0084	0.028
JO	0.023	0.03	0.02	0.039	0.018	0.36	0.025	0.053	0.028	0.0097	0.035	0.21	0.016	0.041	0.012	0.045	0.011	0.028
KW	0.058	0.073	0.0097	0.014	0.028	0.014	0.52	0.0074	0.023	0.0064	0.047	0.017	0.06	0.076	0.0076	0.01	0.0047	0.026
LB	0.016	0.014	0.012	0.016	0.0094	0.042	0.016	0.66	0.017	0.0047	0.012	0.062	0.01	0.0087	0.0098	0.064	0.0072	0.016
LY	0.02	0.023	0.034	0.041	0.013	0.013	0.017	0.015	0.65	0.014	0.028	0.024	0.013	0.019	0.014	0.015	0.015	0.026
MA	0.0087	0.013	0.075	0.044	0.0095	0.012	0.013	0.021	0.031	0.61	0.015	0.02	0.0095	0.035	0.02	0.012	0.024	0.027
OM	0.089	0.075	0.019	0.014	0.021	0.036	0.066	0.011	0.026	0.01	0.4	0.03	0.058	0.067	0.014	0.014	0.0068	0.043
PL	0.021	0.021	0.018	0.075	0.014	0.11	0.016	0.053	0.022	0.0078	0.026	0.47	0.016	0.026	0.016	0.048	0.012	0.024
QA	0.085	0.064	0.0093	0.014	0.02	0.015	0.11	0.0061	0.019	0.0058	0.046	0.014	0.46	0.078	0.0097	0.0077	0.009	0.03
SA	0.056	0.058	0.019	0.016	0.017	0.026	0.082	0.0093	0.026	0.01	0.06	0.019	0.072	0.45	0.0097	0.016	0.012	0.046
SD	0.015	0.017	0.012	0.096	0.017	0.015	0.016	0.016	0.023	0.01	0.019	0.029	0.024	0.03	0.61	0.015	0.0055	0.025
SY	0.013	0.02	0.025	0.026	0.023	0.065	0.022	0.17	0.024	0.0074	0.026	0.094	0.014	0.031	0.011	0.4	0.0074	0.023
TN	0.017	0.021	0.06	0.11	0.0097	0.012	0.021	0.029	0.056	0.025	0.016	0.017	0.0087	0.036	0.011	0.017	0.51	0.017
YE	0.05	0.053	0.026	0.031	0.027	0.039	0.035	0.015	0.043	0.016	0.056	0.04	0.05	0.12	0.021	0.025	0.01	0.34



[AraBERT] Test Confusion Matrix
f1_score: 0.5262330859886513

	IQ	LY	QA	PL	SY	TN	JO	MA	SA	YE	DZ	EG	LB	KW	OM	SD	AE	BH
IQ	0.31	0.15	0.0068	0.029	0.011	0.038	0.19	0.0053	0.016	0.008	0.018	0.043	0.1	0.049	0.0011	0.0061	0.0049	0.0042
LY	0.043	0.4	0.0042	0.014	0.021	0.047	0.24	0.0042	0.024	0.0053	0.01	0.051	0.062	0.062	0.003	0.0049	0.0042	0.0042
QA	0.013	0.024	0.51	0.051	0.011	0.032	0.032	0.0068	0.057	0.077	0.0056	0.091	0.0056	0.019	0.0025	0.011	0.045	0.0043
PL	0.0019	0.011	0.0052	0.85	0.0012	0.009	0.0087	0.0016	0.015	0.0043	0.0012	0.067	0.0021	0.0043	0.01	0.0014	0.005	0.0016
SY	0.014	0.098	0.017	0.0084	0.48	0.048	0.16	0.0077	0.0097	0.0084	0.0032	0.061	0.029	0.026	0.0026	0.014	0.0045	0.0045
TN	0.0086	0.061	0.017	0.035	0.0079	0.34	0.061	0.016	0.017	0.0061	0.0025	0.34	0.028	0.032	0.00036	0.017	0.0047	0.0032
JO	0.023	0.15	0.004	0.01	0.013	0.029	0.62	0.0026	0.016	0.0059	0.0028	0.025	0.047	0.047	0.0014	0.0033	0.0024	0.0031
MA	0.004	0.019	0.013	0.016	0.0069	0.063	0.021	0.54	0.016	0.01	0.00072	0.21	0.004	0.0022	0.0022	0.061	0.0087	0.00072
SA	0.0063	0.039	0.035	0.052	0.0055	0.021	0.04	0.0063	0.62	0.018	0.0074	0.081	0.009	0.016	0.0047	0.0068	0.029	0.0058
YE	0.011	0.023	0.094	0.037	0.0061	0.016	0.02	0.0078	0.031	0.64	0.0017	0.046	0.0069	0.027	0.0061	0.0078	0.016	0.0017
DZ	0.088	0.14	0.015	0.015	0.014	0.067	0.14	0.0058	0.029	0.0073	0.23	0.089	0.059	0.079	0.00052	0.0063	0.0031	0.0058
EG	0.0071	0.029	0.01	0.067	0.0043	0.091	0.025	0.02	0.018	0.0064	0.0055	0.67	0.0094	0.011	0.0027	0.019	0.0037	0.0025
LB	0.064	0.13	0.0045	0.014	0.0055	0.031	0.22	0.0029	0.0093	0.0051	0.0084	0.023	0.41	0.059	0.0016	0.0016	0.00064	0.0039
KW	0.022	0.097	0.016	0.02	0.01	0.041	0.2	0.0022	0.017	0.0078	0.01	0.036	0.081	0.42	0.0019	0.0052	0.0022	0.0097
OM	0.0049	0.037	0.012	0.11	0.0049	0.024	0.028	0.0035	0.025	0.011	0.0035	0.11	0.017	0.009	0.58	0.0055	0.0083	0.0028
SD	0.0074	0.031	0.02	0.03	0.014	0.095	0.034	0.13	0.018	0.011	0.0043	0.25	0.0068	0.029	0.0037	0.31	0.0049	0.0031
AE	0.024	0.034	0.07	0.12	0.0032	0.016	0.029	0.0097	0.063	0.036	0.0032	0.073	0.013	0.029	0.0054	0.011	0.46	0.0011
BH	0.022	0.13	0.022	0.053	0.015	0.062	0.096	0.012	0.042	0.015	0.019	0.1	0.068	0.12	0.0081	0.019	0.004	0.19

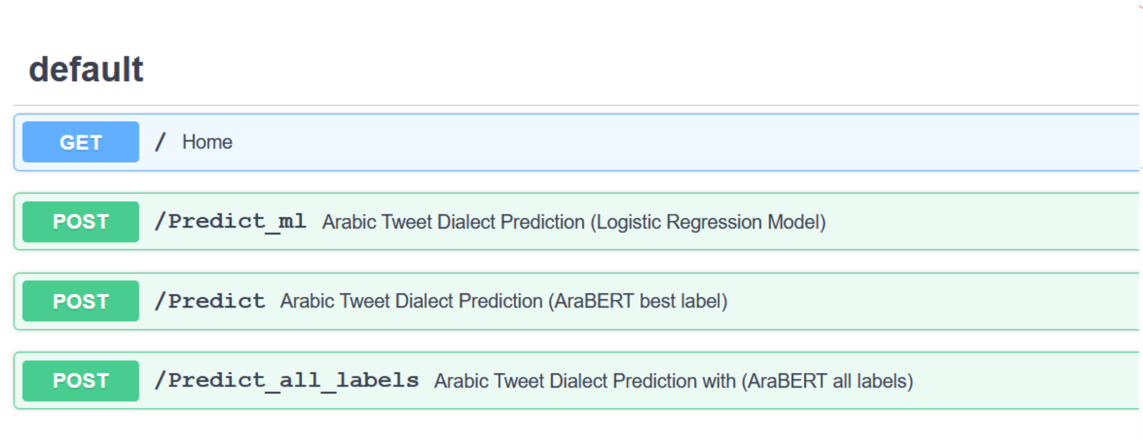
05 Deployment

FastAPI

HUGGING FACE

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. Flask / Fast API it’s a two technologies that used to implement our webserver app



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)

THANKS!

Do you have any questions?