



PROJET NLP

Offensiveness detection in Arabic with geographic disaggregation (Offensiveness + Dialect detection)

Réalisé par :

- Mustapha Ahricha

Encadré par :

- Pr : Jihad Zahir

CONTENT

- 01** INTRODUCTION
- 02** DESCRIPTION DU PROJET
- 03** OFFENSIVE DETECTION
- 04** DETECTION DE DIALECTE
- 05** VISUALISATION & TRAITEMENT DES DONNÉES
- 06** MODELING
- 07** DEPLOIEMENT

INTRODUCTION



NLP est une branche d'IA qui se concentre sur l'interaction entre les ordinateurs et le langage humain. Elle vise à permettre aux machines de comprendre, d'interpréter et de générer un langage naturel de manière semblable à celle des humains.



Notre projet se concentre sur deux aspects essentiels: Detection d'offensive et classification du dialecte arabe



DESCRIPTION DU PROJET

Détection de l'Offensivité en Arabe



- Se concentre sur la création d'outils et de modèles qui peuvent automatiquement repérer et classifier les contenus offensants ou inappropriés dans des textes écrits en arabe. En utilisant des techniques de NLP,
- l'objectif est de développer des systèmes capables de filtrer ces contenus dans divers contextes en ligne, comme les réseaux sociaux ou les plateformes de messagerie,

- consiste à élaborer des outils et des modèles de traitement du langage permettant d'identifier et de distinguer les différents dialectes arabes.
- vise à développer des algorithmes capables de reconnaître et de classifier automatiquement ces variations linguistiques.

Detection de dialecte arabe



DÉTECTION DE L'OFFENSIVITÉ EN ARABE

DATASET DESCRIPTION

Le dataset sur laquelle on a travailler c'est une dataset arriver à la fusion du 3 dataset principale :

syriens/libanais

3 catégories:

- Les tweets normaux
- Les tweets abusifs
- Les tweets haineux
- regroupe 5 846 tweets

L-HSAB

Tweets de tunisian
Contient 2 catégories
Offensives et
Non-Offensive

TUNISIAN-HATE-
SPEECH-AND-ABUSIVE

2 catégories : Offensive
et Non-Offensive
4000 Tweets

ARABIC_OFFENSIVE_COM
MENT_DETECTION_ANNO
TATION_4000_SELECTED

DÉTECTION DE L'OFFENSIVITÉ EN ARABE

DATASET DESCRIPTION

Le dataset sur laquelle on a travailler c'est une dataset arriver à la fusion du 3 dataset principale :

- Puisque 20000 Sentences pour chaque classe de notre dataset

NOTRE DATASET



Preprocessed_arabic_dialect

Qadi dataset

IADD

dialectid_data.tsv

VISUALISATION & TRAITEMENT DES DONNÉES

OFFENSIVE DETECTION

Le dataset sur laquelle on a travailler c'est une dataset arriver à la fusion du 3 classes principale :

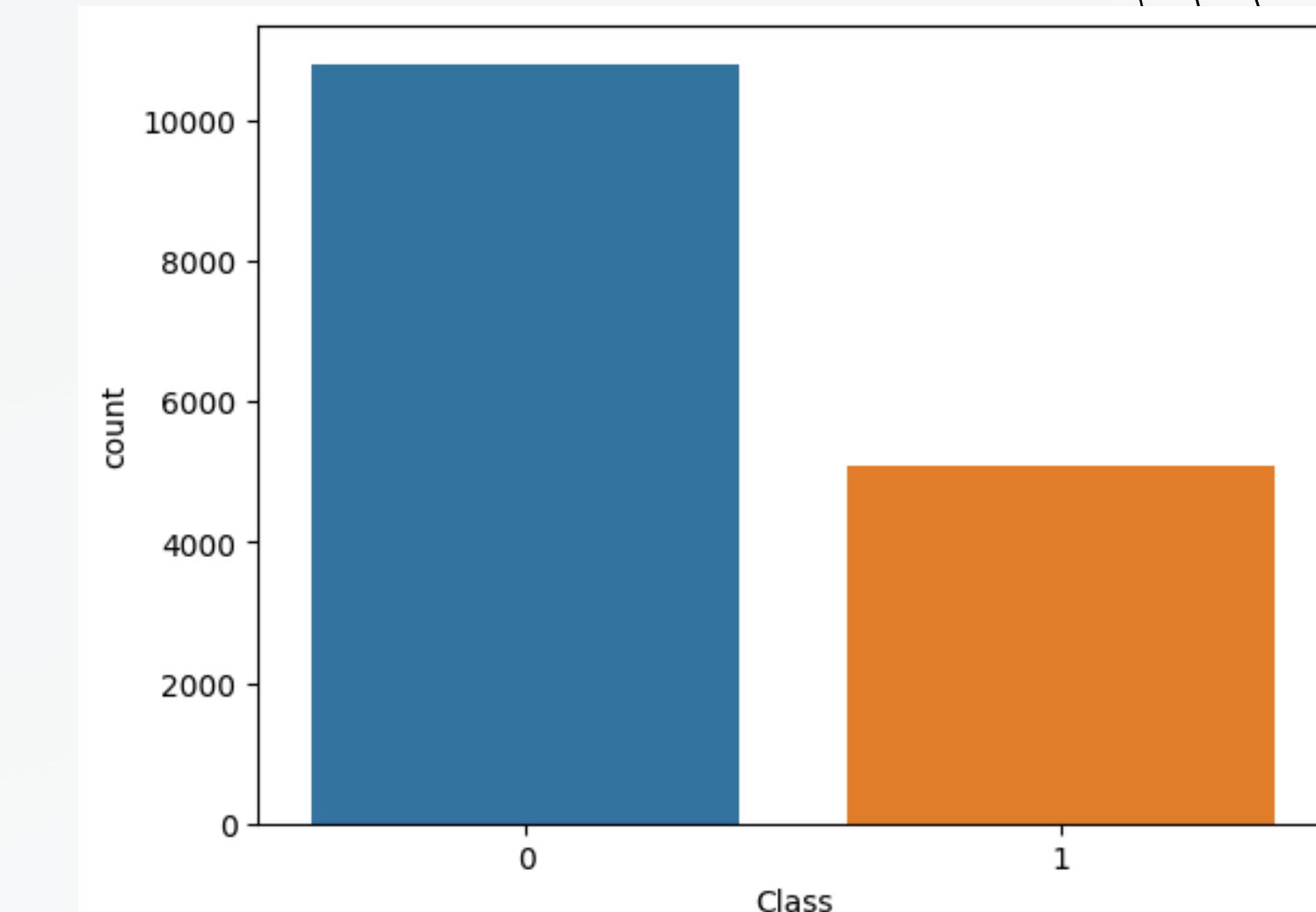
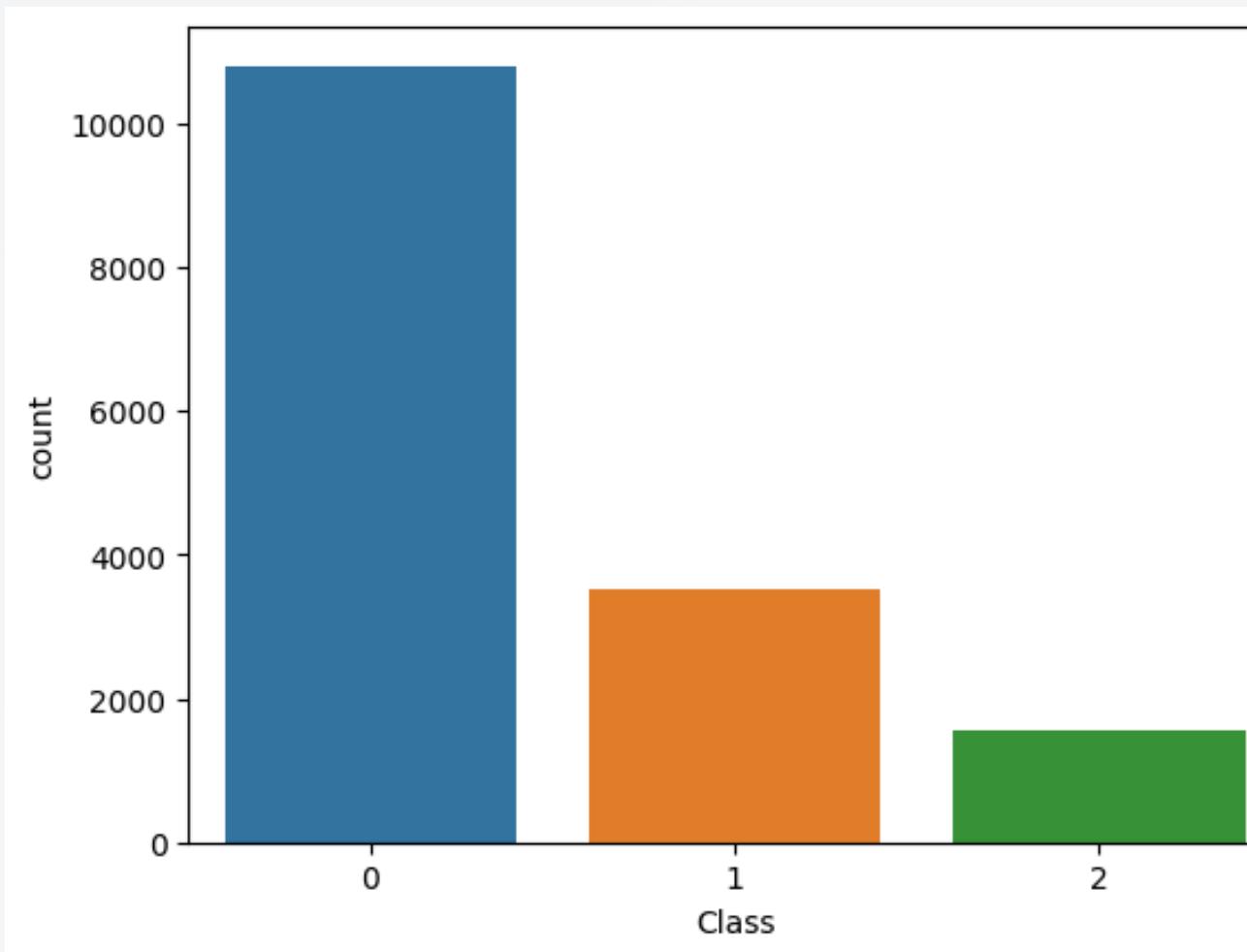
- Suppression des colonnes qui n'ont pas besoin
- Rendre les colonnes de même nom
- Renommer les classes
- Teste des valeurs manquantes

	Tweet	Class
0	الوزير جبران باسيل تاج راسك يا جربان ممنوع بعد	Offensive
1	صديقني انت ابن جامعه اللعبه اكبر من داعش اللعبه	Non-Offensive
2	و مصلحة لبنان تبدأ باستخراج النفط و الغاز لوقف	Non-Offensive
3	وليد جنبلاط كاتب الحكمة يا قذر	Offensive
4	شو بتلبيلاك كلمة خنزير بتجس مفضلة على قياسك وشك	Offensive
...
15865	رحماك رب رحماك رب التوانسة ولات تناقض القرآن ت	Offensive
15866	إنسان تافه وكلام فارغ تفوروو كلب	Offensive
15867	صرير معجبوك من تحتن عينى قناة عادة مكروها ونط	Non-Offensive
15868	نكرة امهما	Offensive
15869	سي لطفي فلفل ليس مريول تقوا سانية فلفول	Non-Offensive

VISUALISATION & TRAITEMENT DES DONNÉES

OFFENSIVE DETECTION

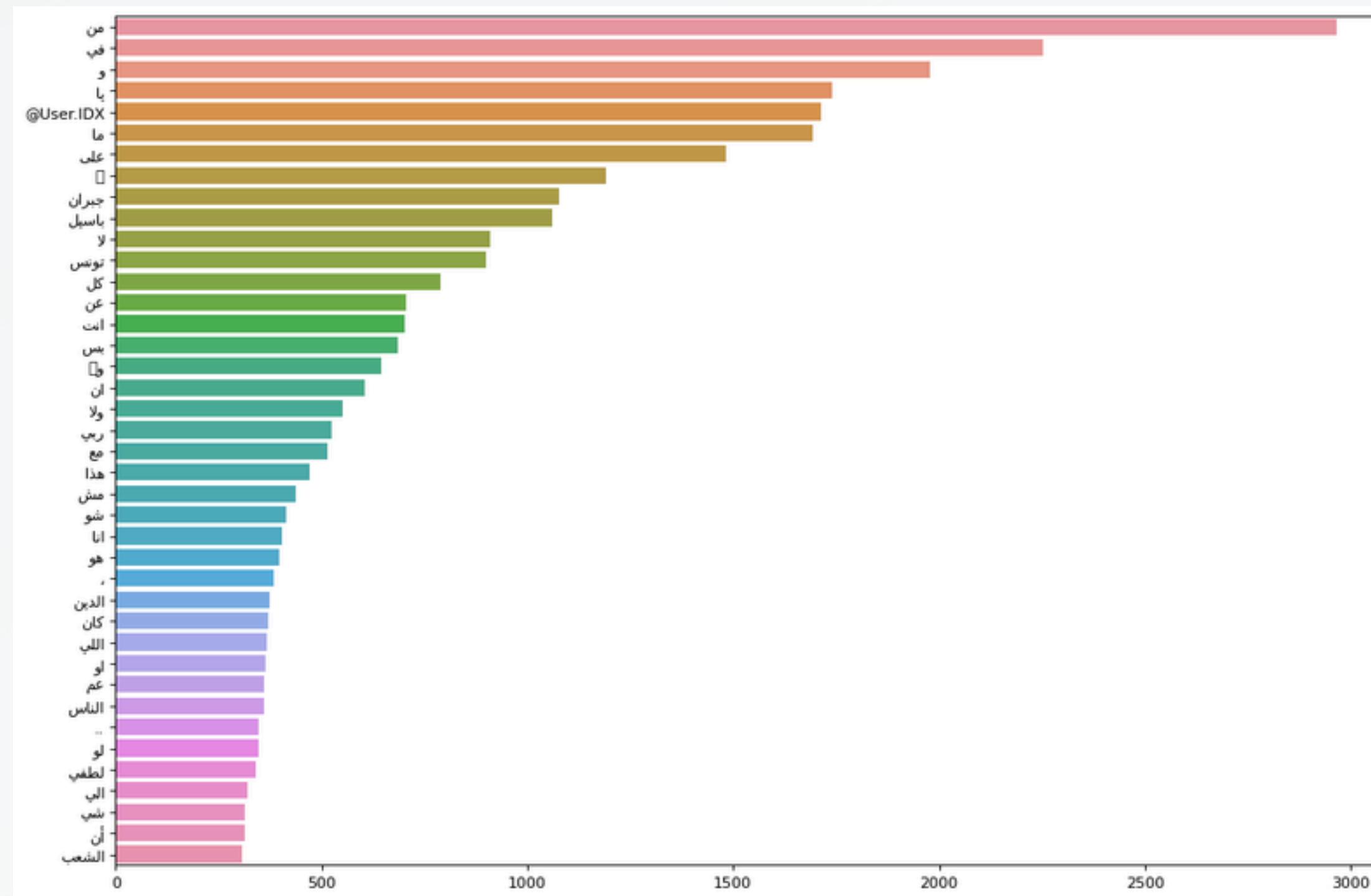
Le dataset sur laquelle on a travailler c'est une dataset arriver à la fusion du 3 classes principale :



Class	
Non-Offensive	10795
Offensive	5075
Name: count, dtype: int64	

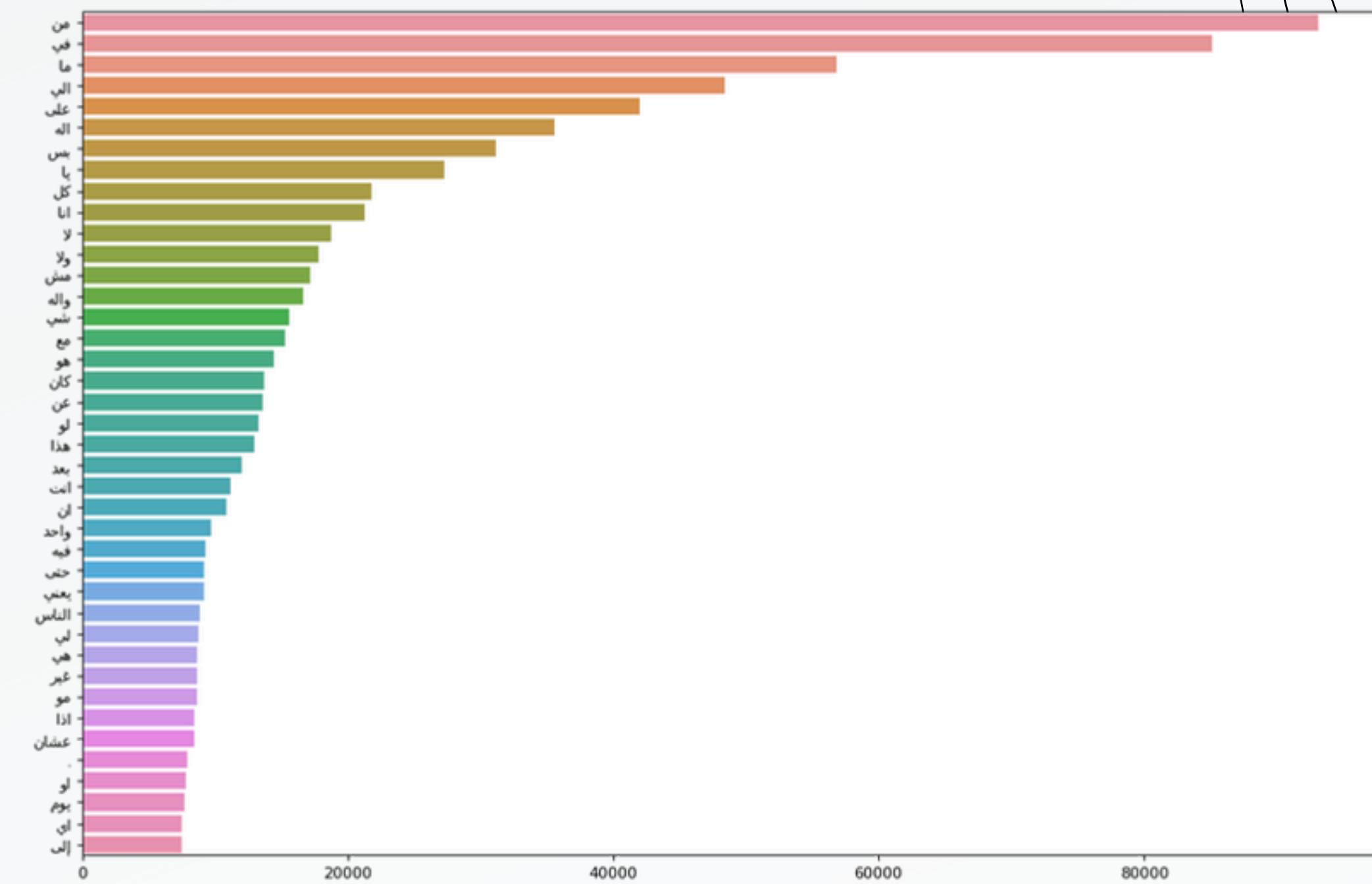
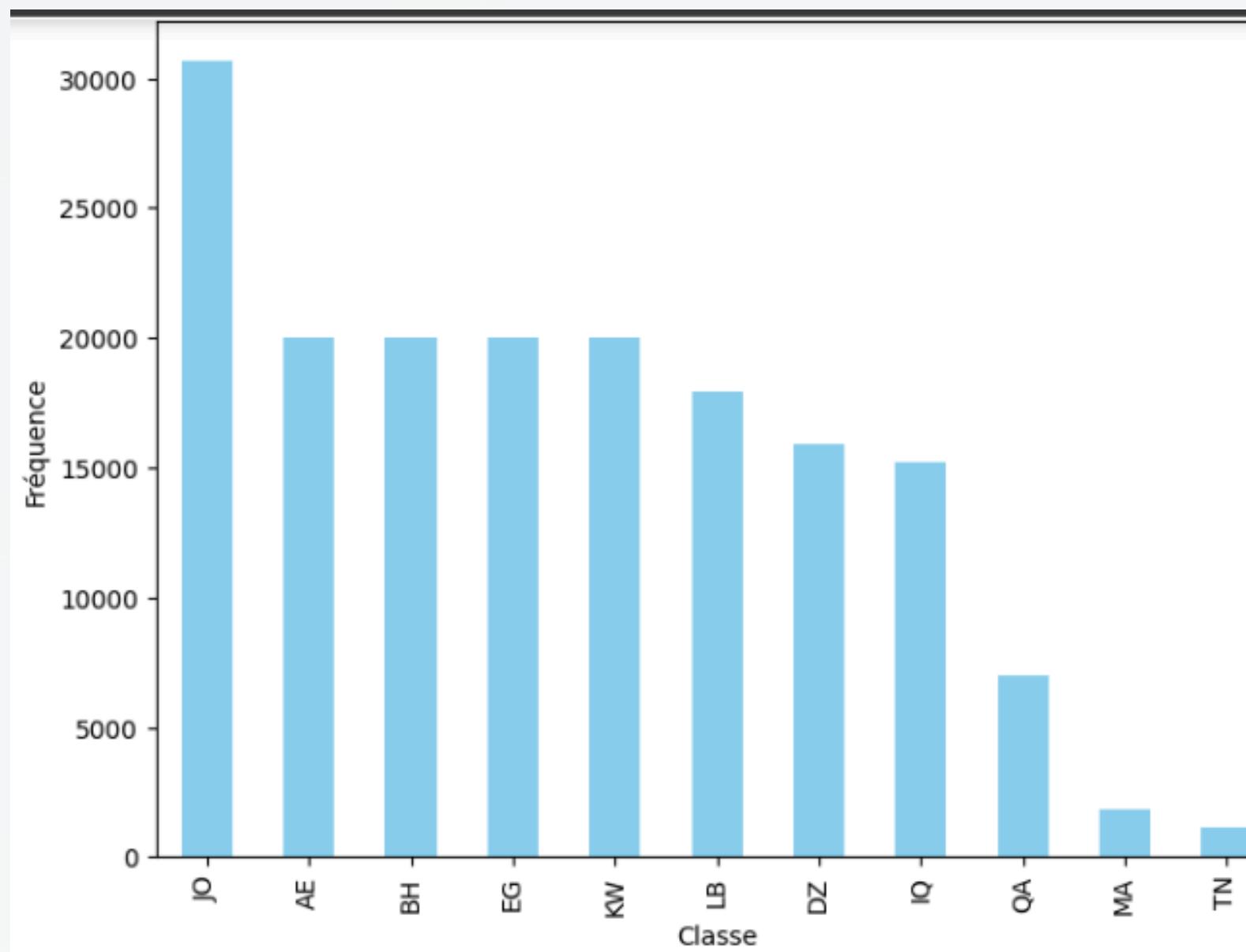
VISUALISATION & TRAITEMENT DES DONNÉES

OFFENSIVE DETECTION



VISUALISATION & TRAITEMENT DES DONNÉES

DIALECT DETECTION



VISUALISATION & TRAITEMENT DES DONNÉES

1. Nettoyage du Texte :

- Suppression des caractères spéciaux, des symboles de ponctuation, et des chiffres pour garantir une cohérence dans les données.

2. Tokenisation :

- Division des phrases en mots individuels (tokens) pour faciliter l'analyse du texte.

3. Lemmatisation :

- Réduction des mots à leur forme de base (lemme) pour standardiser le vocabulaire et améliorer la généralisation du modèle.

4. Suppression des Mots Vides :

- Élimination des mots couramment utilisés mais peu informatifs (mots vides) pour réduire la dimensionnalité du modèle.

5. Encodage du Texte :

- Transformation des mots en représentations numériques pour être utilisées comme entrées pour le modèle de détection de dialecte.



MODELING

On a appliqué les méthodes de machine et deep learning pour trouver le meilleur modèles qui est capable de détecter l'offensive dans les texte d'arabe et aussi de détecter le dialect arabic :



La régression logistique est un algorithme d'apprentissage supervisé utilisé pour la classification.

LOGISTIC REGRESSION



Le "Naive Bayes" est un algorithme d'e ML utilisé pour la classification. Il se base sur le théorème de Bayes en supposant l'indépendance conditionnelle entre chaque paire de caractéristiques et la supposition).

NAIVE BAYES



une méthode d'apprentissage automatique qui combine plusieurs arbres de décision pour prendre des décisions plus précises et robustes. Chaque arbre de décision est formé sur un sous-ensemble aléatoire des données d'entraînement).

RANDOM FOREST

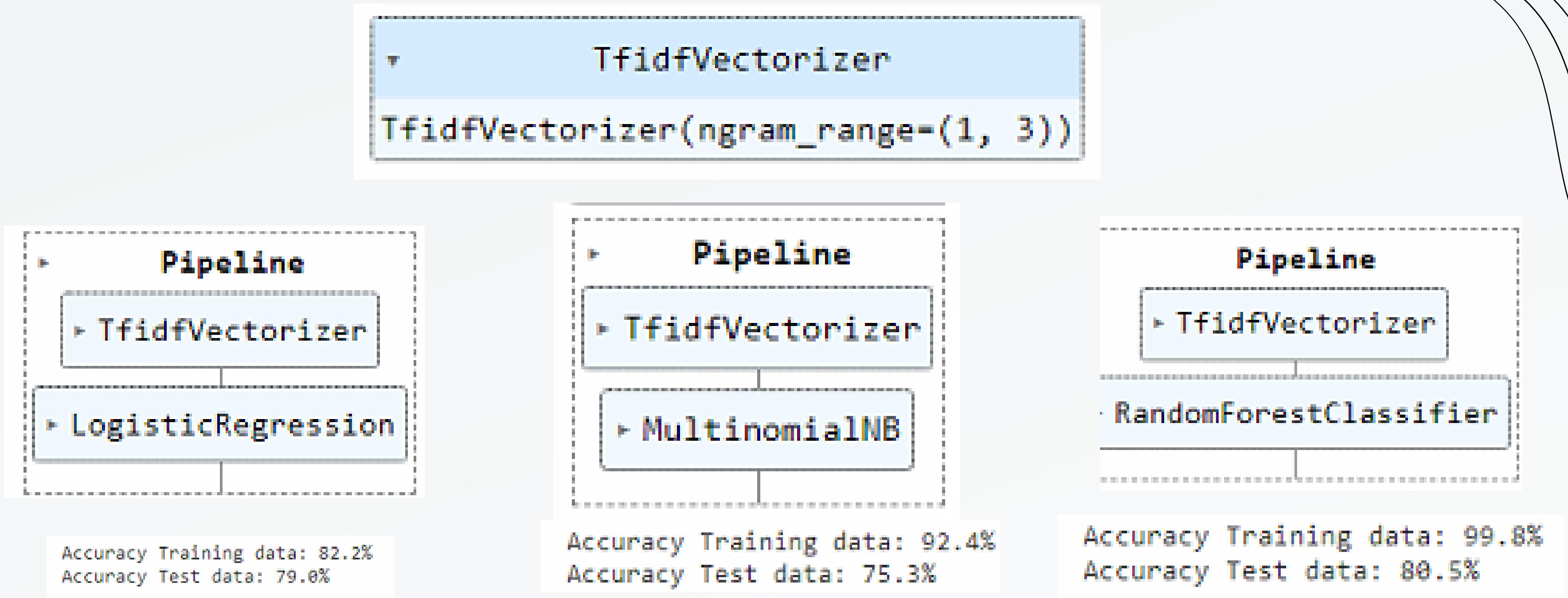
TF / IDF

- est un outil utilisé pour convertir du texte en représentations numériques exploitables par les algorithmes d'apprentissage automatique..
- **TF (Fréquence des Termes)** : Mesure à quel point un terme est fréquent dans un document spécifique. Un terme fréquent aura un score TF plus élevé.
- **IDF (Fréquence Inverse des Documents)** : Mesure à quel point ce terme est rare dans l'ensemble des documents.



METHODOLOGIES

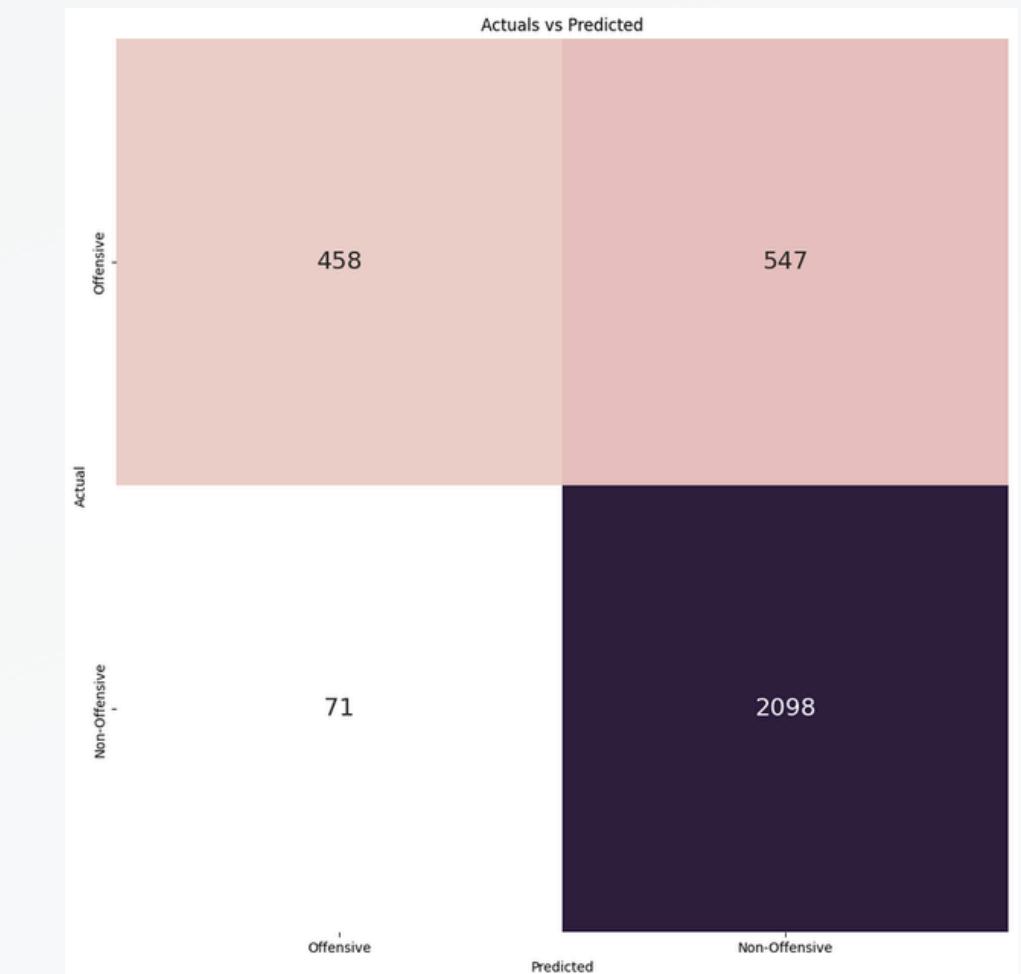
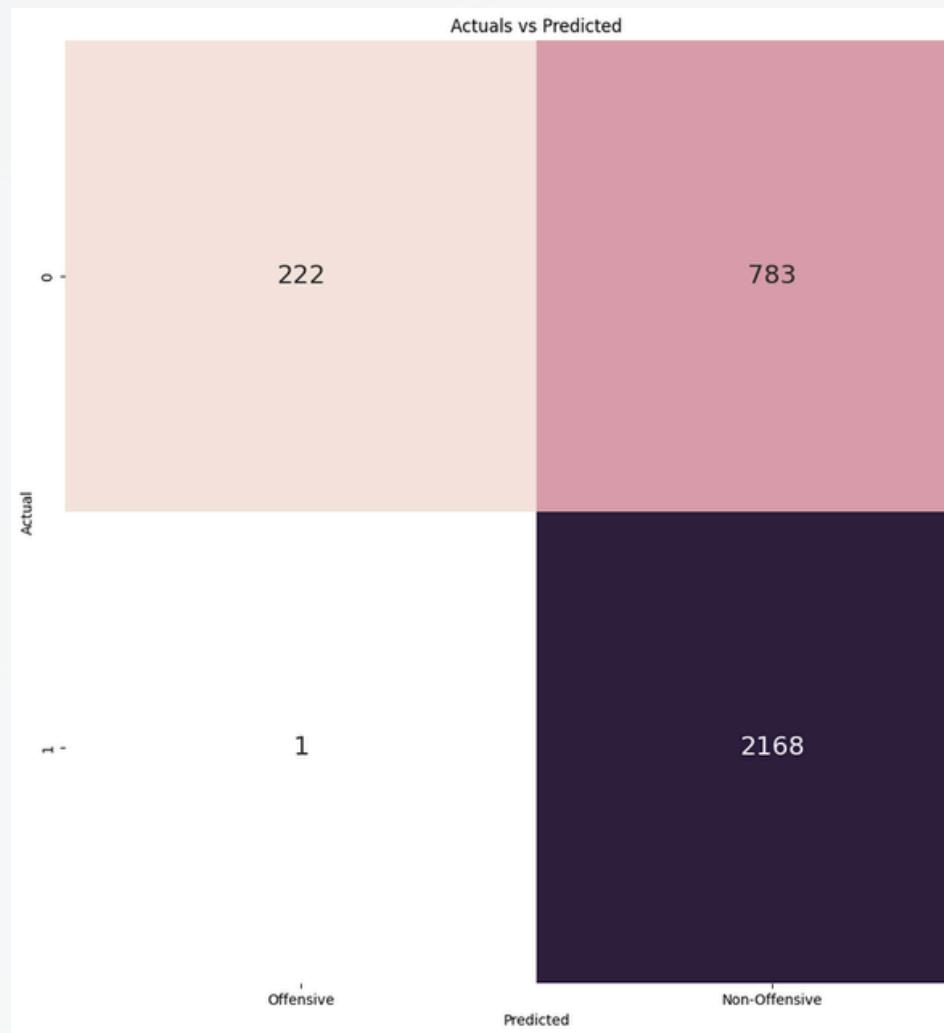
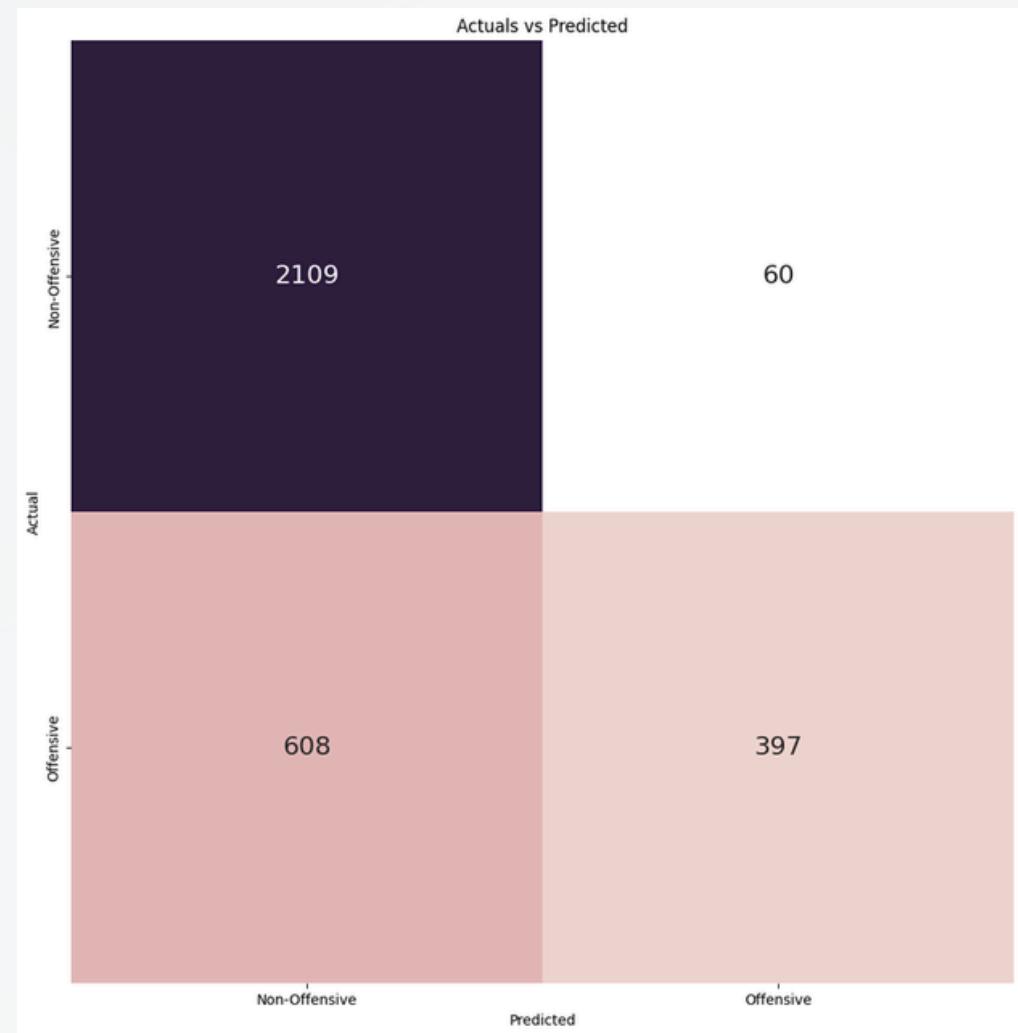
On a appliqué les méthodes de machine et deep learning pour trouver le meilleur modèles qui est capable de detecter l'offensive dans les texte d'arabe et aussi de detecter le dialect arabic :



DEPLOIEMENT

COMPARAISON

A partir la matrice de confusion (Risque / Erreur) pour la detection d'offensive



LogesticR

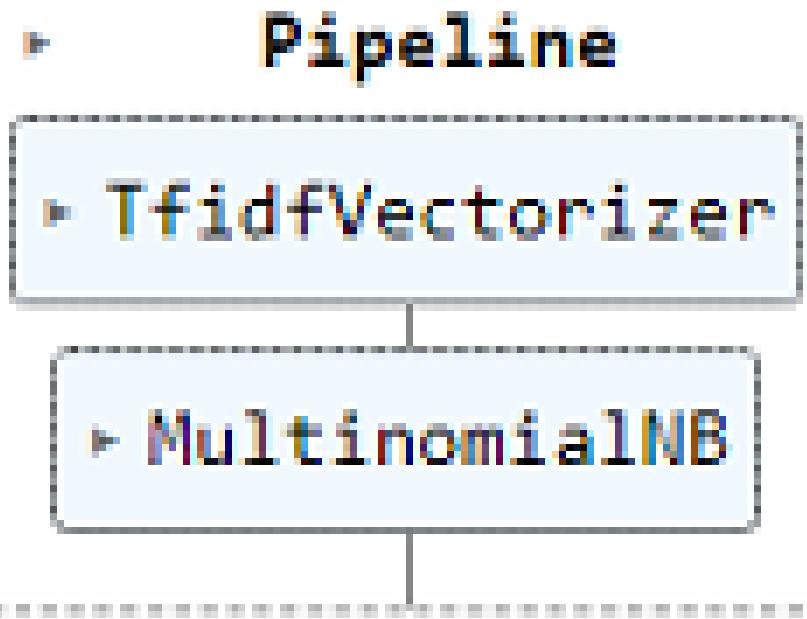
NB

RL

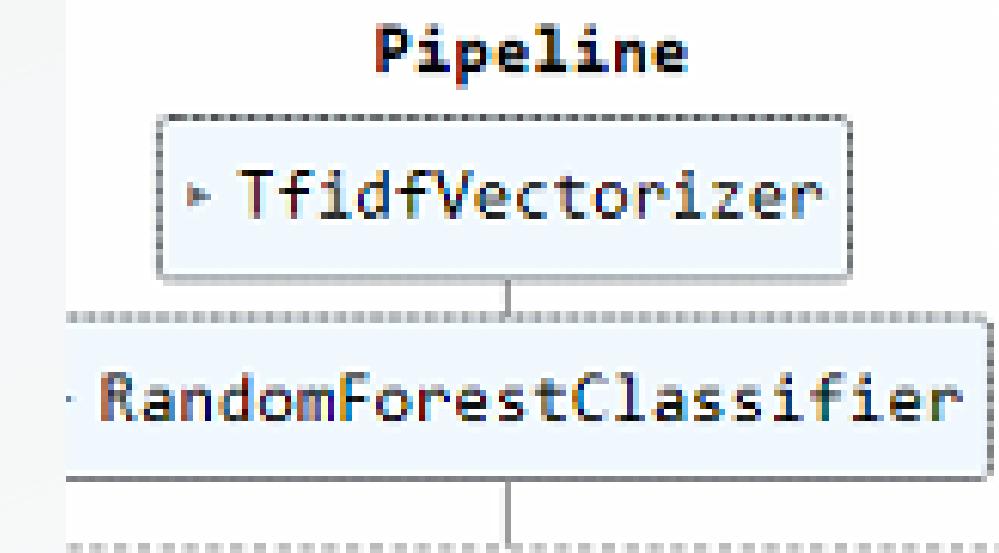
METHODOLOGIES

Pour detection du dialect:

```
TfidfVectorizer  
TfidfVectorizer(ngram_range=(1, 3))
```



Accuracy Training data: 94.6%
Accuracy Test data: 75.9%



Accuracy Training data: 99.8%
Accuracy Test data: 74.7%

DEPLOIEMENT

COMPARAISON

A partir la matrice de confusion (Risque / Erreur) pour la detection d'offensive

		Actuals vs Predicted																		
		Predicted																		
Actual	Predicted	MA	DZ	TN	LY	EG	SD	LB	PL	SY	JO	OM	BH	KW	SA	AE	QA	IQ	YE	MSA
		160	4	0	3	5	0	0	1	0	0	0	1	0	2	2	0	0	0	0
MA	MA	160	4	0	3	5	0	0	1	0	0	0	1	0	2	2	0	0	0	0
DZ	DZ	1	164	0	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0
TN	TN	1	10	111	6	6	0	2	2	0	0	1	0	4	0	2	3	1	0	5
LY	LY	0	6	0	144	3	0	1	2	1	2	0	1	1	1	3	2	1	0	1
EG	EG	2	2	1	2	177	0	2	5	1	1	0	1	1	1	0	2	0	0	2
SD	SD	0	3	0	1	9	165	2	1	0	0	1	0	0	2	2	2	0	0	0
LB	LB	0	1	1	3	2	0	171	4	5	2	0	0	0	2	0	2	1	0	0
PL	PL	0	3	1	7	7	0	9	101	7	24	0	0	0	0	6	4	2	0	2
SY	SY	0	0	0	1	0	0	16	4	154	5	1	1	0	5	1	4	2	0	0
JO	JO	0	4	1	1	2	0	5	12	7	126	3	1	1	2	3	8	1	0	3
OM	OM	0	2	0	0	1	0	1	1	0	0	148	3	2	4	3	1	3	0	0
BH	BH	1	2	0	1	0	0	1	1	0	1	2	134	5	5	12	16	1	0	2
KW	KW	0	1	1	1	2	0	1	1	2	2	2	21	112	11	11	16	2	0	4
SA	SA	0	2	0	2	2	0	1	1	1	4	5	5	11	140	11	11	1	0	2
AE	AE	0	1	1	1	3	0	2	1	0	1	4	5	2	3	156	9	0	0	3
QA	QA	0	1	0	0	1	0	2	1	1	4	2	6	8	7	7	156	0	0	2
IQ	IQ	0	1	1	1	0	0	0	1	0	1	2	1	1	1	0	3	164	0	1
YE	YE	1	5	1	5	14	0	2	10	7	6	12	4	11	26	14	30	4	29	12
MSA	MSA	0	2	1	1	5	0	1	1	0	2	10	6	7	11	4	1	1	0	147

		Actuals vs Predicted																		
		Predicted																		
Actual	Predicted	MA	DZ	TN	LY	EG	SD	LB	PL	SY	JO	OM	BH	KW	SA	AE	QA	IQ	YE	MSA
MA	MA	165	1	3	0	2	1	0	2	0	1	0	0	0	1	1	1	0	0	0
DZ	DZ	2	164	1	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0
TN	TN	4	5	128	5	3	0	0	0	1	1	0	0	0	0	1	1	1	3	0
LY	LY	1	4	3	123	3	0	4	3	3	8	6	0	1	2	4	1	2	1	0
EG	EG	4	2	7	8	142	8	4	3	2	7	2	0	1	2	2	0	1	3	0
SD	SD	0	0	0	1	4	183	0	0	0	0	0	0	0	0	0	0	0	0	0
LB	LB	2	2	2	0	2	0	150	9	10	8	3	1	0	0	0	1	3	0	1
PL	PL	2	3	2	5	6	2	15	91	15	18	3	2	1	2	1	0	3	0	2
SY	SY	3	0	0	0	1	1	12	2	162	4	1	0	2	1	1	3	0	0	0
JO	JO	0	4	3	0	1	1	5	15	5	123	5	3	3	2	3	1	3	1	2
OM	OM	0	2	1	1	2	0	0	1	2	3	144	2	3	2	3	1	1	0	1
BH	BH	5	3	8	1	0	0	1	1	1	2	7	133	3	2	2	6	5	1	3
KW	KW	3	6	3	7	2	0	1	4	2	4	6	10	99	8	14	5	14	0	2
SA	SA	2	7	5	3	2	0	1	4	0	11	9	3	7	124	5	8	5	2	1
AE	AE	4	1	1	0	5	2	2	3	4	7	6	1	7	142	0	3	1	1	1
QA	QA	2	1	5	0	1	0	1	2	1	6	12	13	7	6	13	120	2	3	3
IQ	IQ	1	0	1	1	0	2	0	1	2	1	1	0	2	0	0	163	1	1	1
YE	YE	0	2	3	4	3	1	0	0	4	5	6	3	4	5	4	2	3	143	1
MSA	MSA	2	4	6	4	7	2	1	6	2	8	10	6	6	2	7	0	6	4	117

LogesticR

NB

RL

MODELING

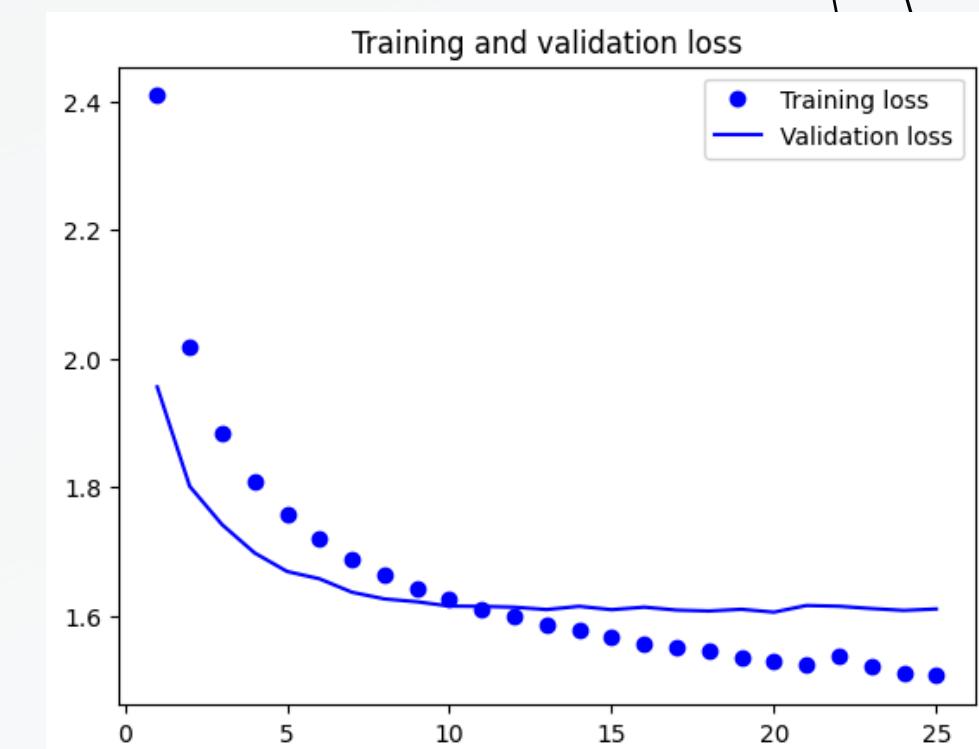
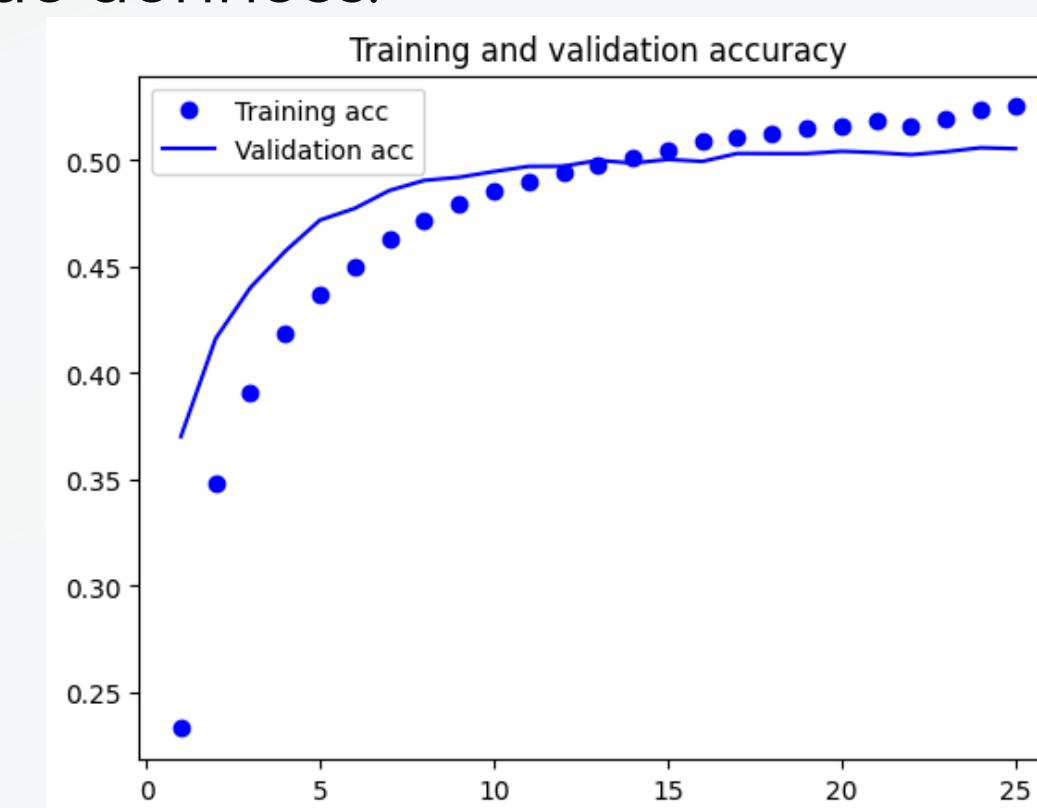
LSTM

sont une forme spécifique de RNN utilisés dans le domaine du NLP et de la séquence de données. Contrairement aux RNN traditionnels, les LSTM sont conçus pour mieux gérer les dépendances à long terme dans les séquences de données.

Pour dialecte

```
epochs = 30  
emb_dim = 128  
batch_size = 128
```

```
model = Sequential()  
model.add(Embedding(n_most_common_words, emb_dim, input_length=X.shape[1]))  
model.add(SpatialDropout1D(0.7))  
model.add(LSTM(64, dropout=0.7, recurrent_dropout=0.7))  
model.add(Dense(18, activation='softmax'))  
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['acc'])
```



Test set
Loss: 1.563
Accuracy: 0.531



Loss: 1.563
Accuracy: 0.531

MODELING

TEST LSTM

```
txt = ["عذان تقبل لازم تعمل كوبس مش تبقي تايم كده"]
seq = tokenizer.texts_to_sequences(txt)
padded = pad_sequences(seq, maxlen=max_len) # Assuming you have defined `max_len`
pred = modelA.predict(padded)
labels = ['SA', 'QA', 'KW', 'AE', 'OM', 'JO', 'PL', 'BH', 'LY', 'EG', 'SD', 'IQ', 'LB', 'SY', 'TN', 'DZ', 'MA', 'YE']
predicted_label = labels[np.argmax(pred)]
print(pred, predicted_label)
```

Pyth

```
1/1 [=====] - 0s 44ms/step
[[6.2127196e-04 6.2448857e-04 4.7383294e-04 3.0227038e-03 3.6598285e-04
 6.8539870e-03 4.4943579e-02 3.7773870e-04 8.4192511e-03 8.1925243e-01
 3.5766996e-02 1.6358263e-03 1.9873355e-03 1.8251145e-03 2.6246684e-02
 4.2933721e-02 2.4335973e-03 2.2955793e-03]] EG
```

```
txt = ["ما نوصل 4 د الليل وباف مانعستيش"]
seq = tokenizer.texts_to_sequences(txt)
padded = pad_sequences(seq, maxlen=max_len)
pred = model.predict(padded)
labels = ['SA', 'QA', 'KW', 'AE', 'OM', 'JO', 'PL', 'BH', 'LY', 'EG', 'SD', 'IQ', 'LB', 'SY',
          'TN', 'DZ', 'MA', 'YE']
print(pred, labels[np.argmax(pred)])
```

```
1/1 [=====] - 0s 315ms/step
[[1.2893110e-04 3.0861534e-05 8.5972686e-05 7.9781683e-05 8.5899657e-05
 3.3717573e-05 1.2692495e-04 2.3031171e-04 6.5637017e-03 1.3960707e-04
 7.6095774e-03 1.2437424e-04 1.3462806e-04 2.0933572e-04 8.2762793e-02
 1.4684219e-02 8.8691097e-01 5.8377042e-05]] MA
```

DEPLOIEMENT

INTERFACE

Voici l'interface qu'on a choisi pour déployer notre modèle. Alors on choisit de déployer LSTM pour la détection de dialecte et Random Forest pour la détection d'offensive.

The screenshot shows a web application interface. On the left, a sidebar titled "Choisissez un projet" contains a dropdown menu with two options: "Test dialecte" (selected) and "Test d'offensive". On the right, the main content area has a title "Test de dialecte et test d'offensive". It includes a welcome message "Bienvenue dans le test de dialecte!", a text input field "Entrez du texte pour le test de dialecte", and a file upload section "Choisissez un fichier CSV pour le test de dialecte" with a "Drag and drop file here" button and a "Browse files" button. A "Limit 200MB per file" note is also present. At the bottom, there is a button "Afficher les prédictions pour le test de dialecte". In the top right corner of the main content area, there are "Deploy" and three-dot buttons.

DEPLOIEMENT

INTERFACE

Exemple d'entré un texte et l'importer un csv

Choisissez un projet
Test d'offensive

Test de dialecte et test d'offensive

Bienvenue dans le test d'offensive!

Entrez du texte pour le test d'offensive

Choisissez un fichier CSV pour le test d'offensive

Drag and drop file here
Limit 200MB per file

Browse files

 output1 (1).csv 11.3MB X

Afficher les résultats pour le test d'offensive

Choisissez un projet
Test dialecte

Test de dialecte et test d'offensive

Bienvenue dans le test de dialecte!

Entrez du texte pour le test de dialecte

أوالماش نيل اليوم واعر اصحابي عرقني البارحة لعبات مزيان

Test de dialecte et test d'offensive

Bienvenue dans le test de dialecte!

Entrez du texte pour le test de dialecte

أوالماش نيل اليوم واعر اصحابي عرقني البارحة لعبات مزيان

Choisissez un fichier CSV pour le test de dialecte

Drag and drop file here
Limit 200MB per file

Browse files

Afficher les prédictions pour le test de dialecte

X

DEPLOIEMENT

DIALECT DETECTION

Exemple test par un texte :

```
df['Country'] = df['Country'].map({
    'EG': 'Egypt', 'SA': 'Saudi Arabia', 'MA': 'Morocco', 'DZ': 'Algeria', 'SY': 'Syria',
    'QA': 'Qatar', 'LB': 'Lebanon', 'YE': 'Yemen', 'AE': 'United Arab Emirates', 'KW': 'Kuwait',
    'SD': 'Sudan', 'BH': 'Bahrain', 'JO': 'Jordan', 'IQ': 'Iraq', 'PL': 'Palestine', 'OM': 'Oman',
    'LY': 'Libya', 'TN': 'Tunisia'
})
```

Choisissez un projet

Test dialecte

Bienvenue dans le test de dialecte!

Entrez du texte pour le test de dialecte

فتش كفر مصل ٤ د الليل وبقى مانعطفين

Choisissez un fichier CSV pour le test de dialecte

Drag and drop file here
Limit 200MB per file

Browse files

Afficher les prédictions pour le test de dialecte

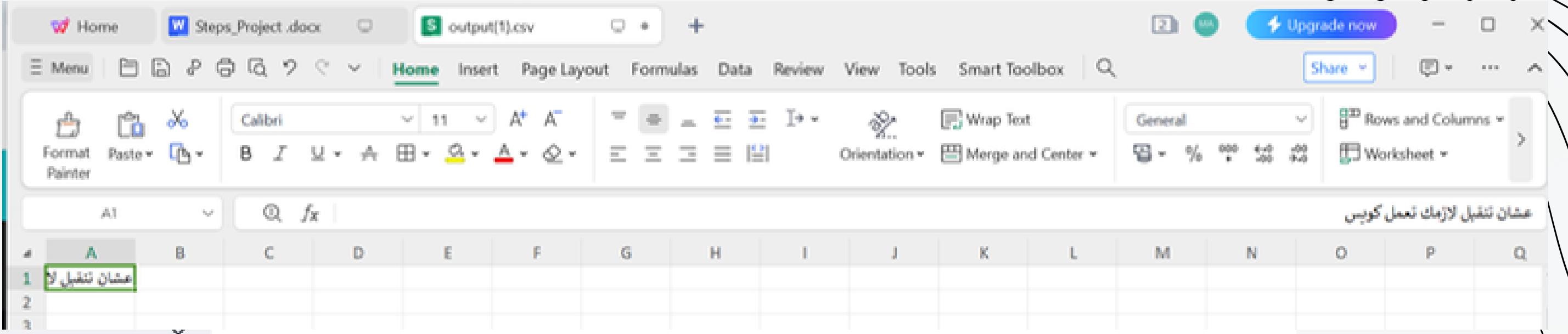
Prédictions pour le test de dialecte

Morocco

DEPLOIEMENT

DIALECT

Exemple de déploiement d'un fichier csv contenant une sentence :



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	A															
2	عنوان تنقل لازمك تعمل كوبس															

Test de dialecte et test d'offensive

Choisissez un projet

Test dialecte

Bienvenue dans le test de dialecte!

Entrez du texte pour le test de dialecte

Choisissez un fichier CSV pour le test de dialecte

Drag and drop file here
Limit 200MB per file

Browse files

output1 (1).csv 11.3MB

Afficher les prédictions pour le test de dialecte

Prédictions pour le test de dialecte

Egypt

DEPLOIEMENT

```
text_input("Entrez du texte pour le test d'offensive")
file = st.file_uploader("Choisissez un fichier CSV pour les résultats d'offensive")
st.write("Résultats pour le test d'offensive:")
# Prediction avec le modèle Random Forest
defensive = random_forest_model.predict([text])
# Résultat de la prédiction
if defensive == 1:
    "Ce texte contient l'offensive."
else:
    "Ce texte est normal."
```

Exemple d'entrer un texte d'offensive

Choisissez un projet

Test d'offensive

Test de dialecte et test d'offensive

Bienvenue dans le test d'offensive!

Entrez du texte pour le test d'offensive

نفر عليك يا الكلب و الله انت ماسكتهاش الغير لي درت فيهاك. النسان زيافه و حمار

Choisissez un fichier CSV pour le test d'offensive

Drag and drop file here
Limit 200MB per file

Browse files

Afficher les résultats pour le test d'offensive

Résultats pour le test d'offensive

Ce texte contient l'offensive ↗

DEPLOIEMENT

DETECTION D'OFFENSIVE

Exemple d'entrée un texte d'offensive

The screenshot shows a web-based application for testing offensive language detection. On the left, a sidebar displays a dropdown menu titled "Choisissez un projet" with "Test d'offensive" selected. The main content area has a large title "Test de dialecte et test d'offensive". Below it, a welcome message "Bienvenue dans le test d'offensive!" is followed by a text input field containing Arabic text: "محظى تأهيل مسكن أين اختنى هذا الفنان لماذا لم بعد يشارك في التصفيق ولو انه فنان سيرجي الأول في زمـن الإيجـاع الفـارقـي". A CSV file upload section follows, with instructions "Choisissez un fichier CSV pour le test d'offensive", a "Drag and drop file here" button featuring a cloud icon, a "Limit 200MB per file" note, and a "Browse files" button. At the bottom, a red-bordered button labeled "Afficher les résultats pour le test d'offensive" is visible, along with the text "Résultats pour le test d'offensive" and "Ce texte est normal".

Choisissez un projet

Test d'offensive

Test de dialecte et test d'offensive

Bienvenue dans le test d'offensive!

Entrez du texte pour le test d'offensive

محظى تأهيل مسكن أين اختنى هذا الفنان لماذا لم بعد يشارك في التصفيق ولو انه فنان سيرجي الأول في زمـن الإيجـاع الفـارقـي

Choisissez un fichier CSV pour le test d'offensive

Drag and drop file here

Limit 200MB per file

Afficher les résultats pour le test d'offensive

Résultats pour le test d'offensive

Ce texte est normal

Browse files

FIN

Merci !