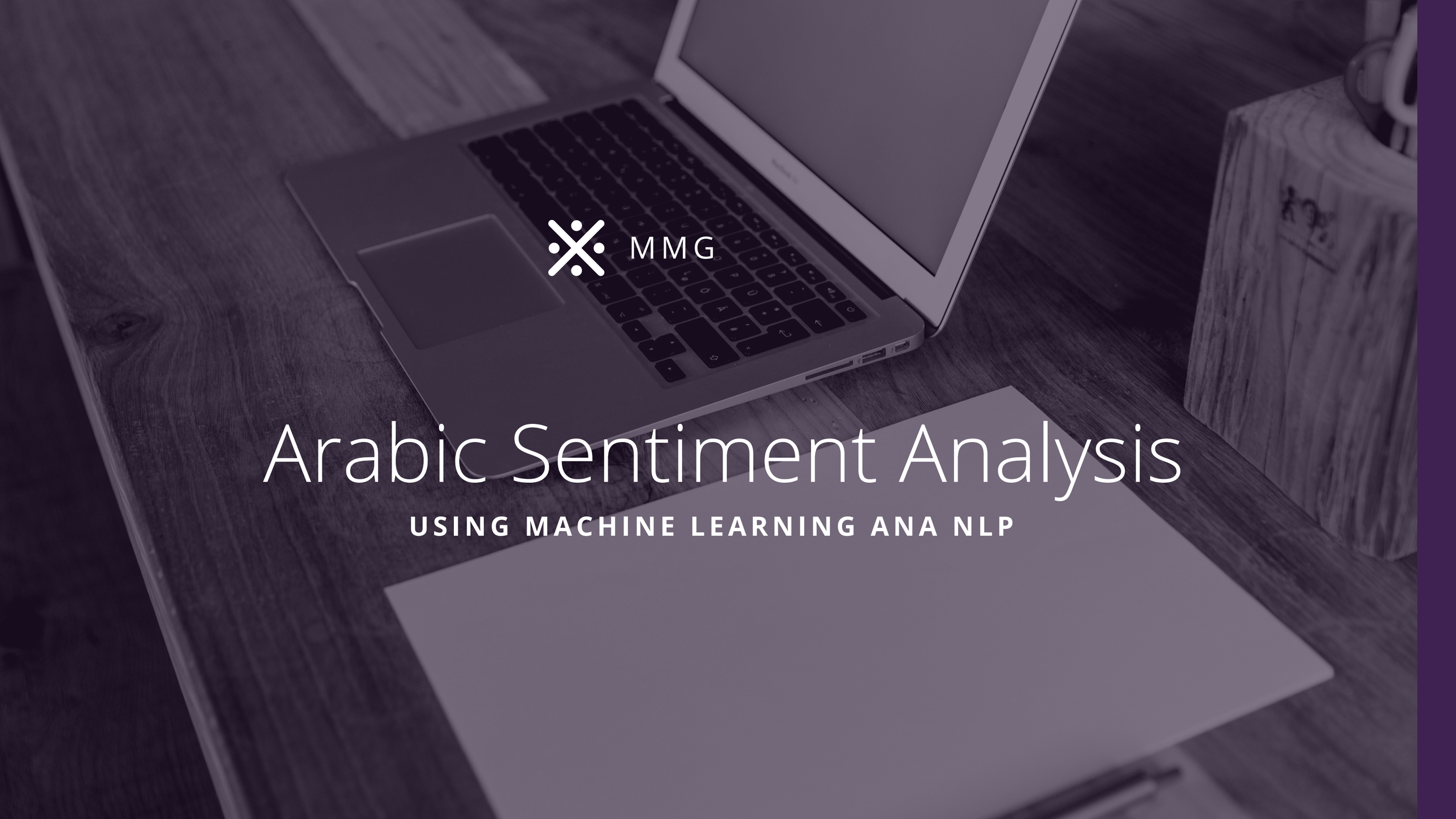


MMG

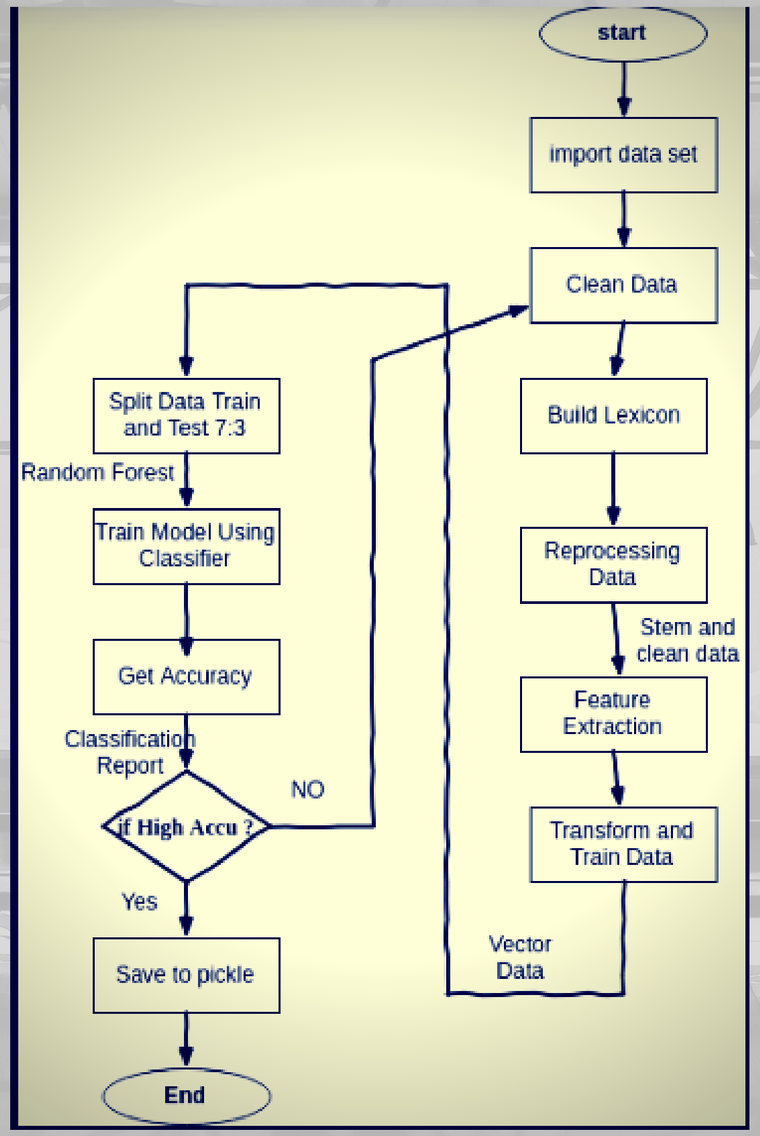
Arabic Sentiment Analysis

USING MACHINE LEARNING ANA NLP

CONTENTS

- Introduction
- Data and Lexicon
- Clean Data
- Reprocessing Data
- Feature Extraction
- ML Algorithm and Save

DESIGN FOR PROJECT



INTRODUCTION

Sentiment analysis is the process of determining a predefined sentiment from online texts written in a natural language with respect to a specific subject .

Arabic language has shown rapid growth in terms of its user on the internet, moving up to the 4th place in the world ranking of languages by users according to internet world stats .

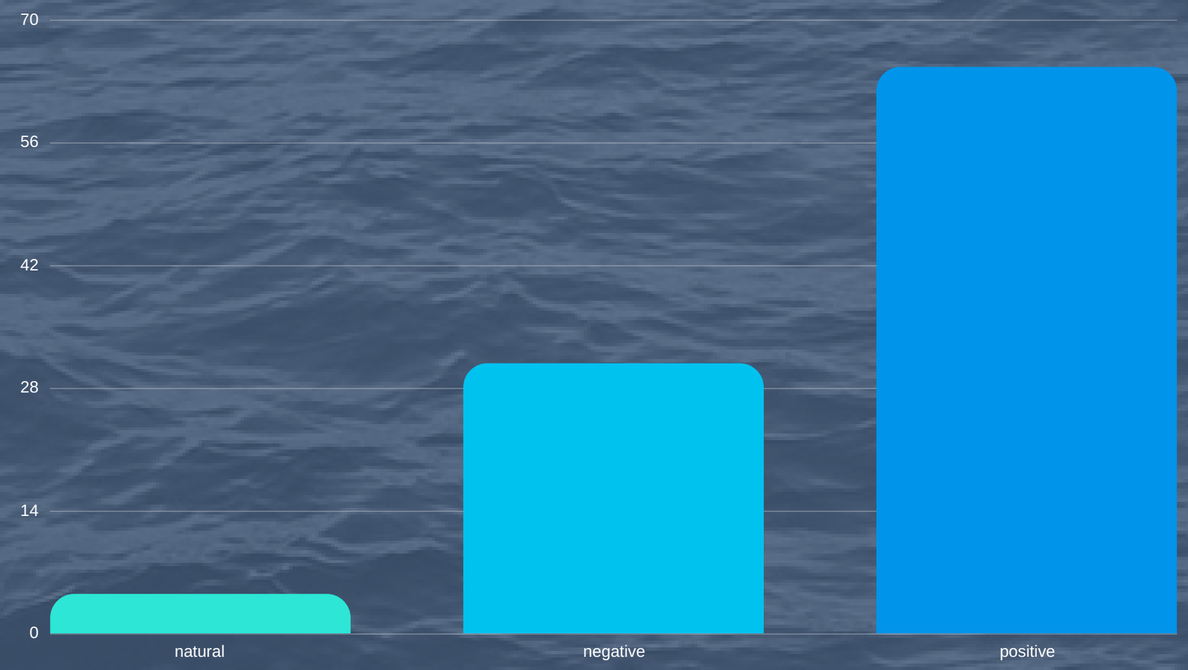
Because of that, there has been an increasing interest and research in the area of Arabic Sentiment Analysis. However, The Arabic Language remains under resourced with respect to the amount of the available dataset.

Aim of extracting a general opinion about one item or theme among the substantial amounts of unstructured data available on the Internet.

DATA AND LEXICON

- **Hotel Reviews (HTL)** For the hotels domain 15K Arabic reviews were scrapped from TripAdvisor Those were written for 8100 Hotels by 13 K users.
- **Restaurant Reviews (RES1)** scrapped from Qaym which 8.6 K Arabic reviews
- **Restaurant Reviews (RES2)** scrapped from TripAdvisor which 2.6 K Arabic reviews
- **Movie Reviews (MOV)** scrapping 1.5 K reviews from elcinema.com covering around 1 K movies.
- **Product Reviews (PROD)** For the Products domain, a data set of 15 K reviews was scraped from the Souq website
- **Random Arabic Tweet (data)** For the Arabic tweet collect randomly from twitter its size is 9.6 MB

REPRESENTATION OF DATS



SIZE

40.7 MB

64.6% data are positive

>DATA CLEANING

- *remove special character*
- *remove number*
- *remove URL*
- *remove name*
- *remove single character*
- *remove unneeded data such as places , names ,date info and othor*

UNNEEDED WORD

collect all places ,name ,days, month and weeks from data set then check if it is affect the accuracy of data we see that now change in accuracy.

>PREPROCESSING

- ***Tokenizing Words***
- ***Some of Stop words not all***
- ***stemming***

STOP WORDS

if we delete all of it the most affect word for prediction

such as

لكن, ليس, لا, غير, سوي, ... ,لن

those word it is make precision to project

STEMMING

we use ISRIStemmer it is the most common for Arabic stem also it is another level of clean data it is delete the prefixes of word

and the verb return to pure with the origin letter and Plural to Singular

TEST

```
st.stem('الحركات')
```
```
'حرك'
```
```
st.stem('الحركه')
```
```
'حرك'
```
```
st.stem('حركات')
```
```
'حرك'
```
```
st.stem('مراجعه')
```
```
'رجع'
```
```
st.stem('الذات')
```
```
'ذات'
```
```
st.stem('الموقع')
```
```
'وقع'
```

DISADVANTAGE

not deal with UN Arabic standard words such as places and Anonymous names

>FEATURE EXTRACTION

We use Term Frequency Inverse Document Frequency (TFIDF)

is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling.

MATHEMATICALLY BEHIND TFIDF

log (N/DFt)

DFt is the number of documents containing the term t.
N is the total number of documents in the corpus.

TFIDF

The higher the numerical weight value, the rarer the term The smaller the weight, the more common the term

SPLIT DATA

We use CROSS VALIDATION python library for splinting data

We divide data to train and test 70:30

We divide data randomly .

ML ALGORITHM

We us more than one algorithm

- **SVC**
- ***Decision Tree***
- ***Random Forest***
- **KNN**

The highest accuracy is Random Forest

83.4%

74%

SVC

75%

KNN

79%

Decision tree

83%

Random Forest

RANDOM FOREST

Why it is good for Sentiment Analysis ?

Check the Accuracy from over Fit

Classification report

```
Accuracy 0.834059254296
++++++++++++++++++++++++++++++++++++++++++++++++++
Classification_Report                precision    recall  f1-score   support

        -1.0       0.91      0.65      0.76      6456
         0.0       0.99      0.12      0.21       976
         1.0       0.81      0.97      0.88     13461

avg / total       0.85      0.83      0.81     20893
```

EXPLAIN CLASSFICATION REPORT

Confusion matrix

		Predicted/Classified	
		Negative	**Positive**
Actual	**Negative**	998	0
	Positive	1	1

precision

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

text

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

F1 SCORE

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

SAVE

Now time to save model we use ***pickle*** to save it

* used to convert the objects to byte

* as serialize

TEST

```
---------- عشوائي----------
الملف جيد جدا
[1]
الملف غير جيد
[-1]
الموبايل سيء
[-1]
الموبايل رخيم
[1]
الموبايل غالي
[-1]
الموبايل nice
[1]
اجمل  شي في الدنيا
[1]
النت غير بطي
[1]
التحميل مجاني
[1]
الاصدارات الاخيره كان زفت
[-1]
الفيلم very bad
[-1]
اعجبني طريقه وشكله رائع
[1]
اغبى منتج شوفته حياتي
[-1]
```

TEST

اقل من المتوسط
[-1]
ليس الجميل
[-1]
امكانيات جيدة بالنسبة للسعر, خصوصاً 2 جيجا رامات, وحجم الشاشة جيد, وكذلك سعة البطارية والمساحة التخزينية
[1]
تليفوون ممتاز وصوته نقي و البصمة رائعة
[1]
ممتاز من حيث الصوت البطارية والاداء
[1]
thank you
[1]
منتج جيد
[1]
جهاز معقول
[1]
شكرا امازون
[1]
بطيئ جدا
[-1]
موبايل محترم و سعر مناسب
[1]
المنتج بطيئ فعلا حتي في التصفح النت
[-1]
عمر البطارية طويل...حجم الشاشة كبير...امكانيات الجهاز عالية 2رام... اعلى جهاز فى فئته
[1]
عمر البطارية صغيره...حجم الشاشة كبير...امكانيات الجهاز ضعيفه 2رام..اقل جهاز فى فئته
[-1]
الموبايل رائع من حيث البطاريه والبصمه والكاميرا
[1]
سريع جدا والالعاب فيه ممتازة البطارية تدوم لوقت طويل من الاستخدام الشاق
[1]
ارخص سعر والتوصيل ممتاز
[1]

(لكن)

---------- بالنسبه ل لكن ----------

افكاره غريبة ولكنها رائعه
[1]

افكاره جيده ولكنها غريبه
[-1]

سريع ولكن سيئ جدا
[-1]

سريع ولكن غالي جدا
[-1]

(ادوات النفي)

لا يكون بطي
[1]

لا يكون سريع
[-1]

غير بطي
[-1]

ليس بطي
[-1]

غير سريع
[1]

ليس سريع
[-1]

ما بطي
[-1]

لا بطي
[1]

لا سريع
[-1]

غير بطي
[-1]

ليس بطي
[-1]


THANK
YOU