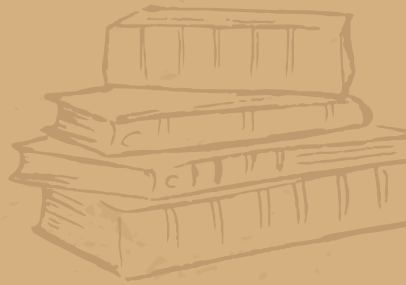
الشعر العربي

NLP project

Presented by:
Mada Abudahish
Mozah AlKhaldi

Workflow



01 introduction

02 Tools

03 EDA

04 Text prepressing

05 Topic Modeling

06 Classification

Introduction

Arabic poems are the oldest and the most prominent form of Arabic literature today.

Goal

In this project, we will develop an **Arabic poems** topic modelling to predict the correct phase (العصر) of poems that can correctly **classify** the topic.

Dataset

The dataset that will be used in this project has been downloaded from(الديوان)

- 1,831,770 **rows**
- 8 **columns**

Tools

- **Software Platform** :Jupyter Notebook
- **Programming Language**: Python
- **Python Libraries:**

  Statistics libraries:
  - Sklearn
  - Nltk

- **Data manipulation libraries:**
  - Pandas
  - Numpy
  - byArabic package
  - Camel tool

- **Visualization libraries**
  - Matplotlib
  - Seaborn
  - wordcloud

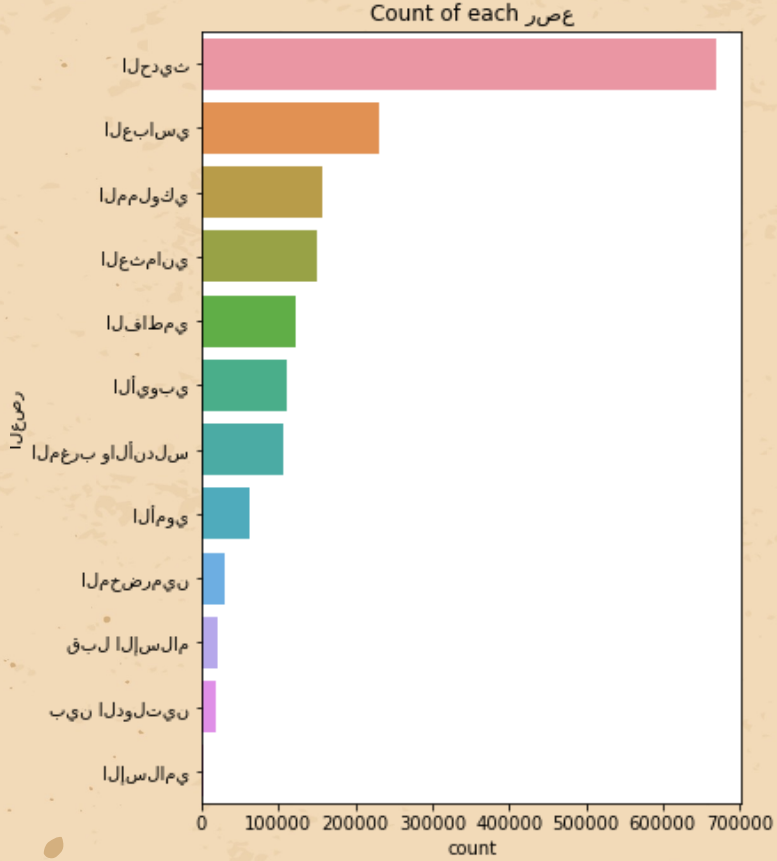
Exploratory data analysis (EDA)

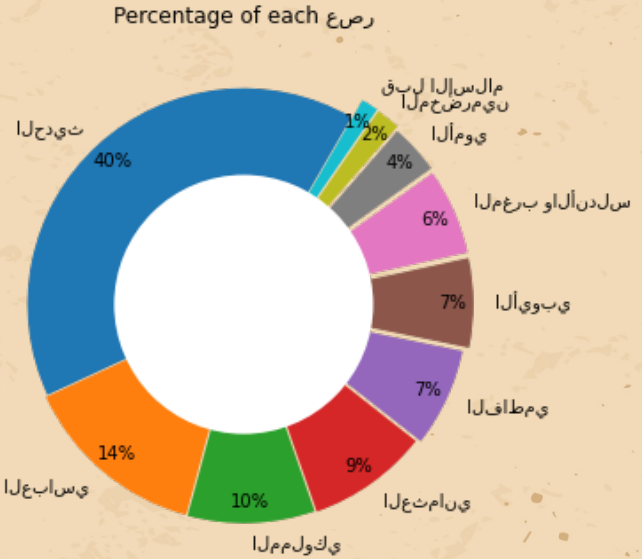
Remove Unnecessary columns

Remove NULLS and Duplicate values

Split Data by الديوان

Distribution of العصر





Word Clouds


يومي
يومه
يوما
يوسف
يومك
يولي
يودي
يهون
يهواه
يهوي
يهيج
يهمي
يود
يو
يهز
يهوى
يهتدي
ينقص
ينفع
ينفك
ينفي
يهتز
ينكر
يهاب
ينزل
ينهل
ينقض
ينظر
ينم

Text prepressing

01

Remove harakat and tashkeel , tatweel.

02

Remove extra spaces and punctuations

03

Tokenization
Stemming.

04

Replace أ إ آ with ا And ؤ with و
And repeated letter with single letter like .[ا],[ى], [وو]

05

Remove Stop Word.

Topic Modelling

Modelling

- Latent Dirichlet Allocation **(LDA)**
- Latent Semantic Analysis (**LSA**)
- Non-Negative Matrix Factorization (**NMF**)
- **CorEx**
- **Clustring**

Embedding

- **CountVectorizer**
- **TF-IDF**

NMF | Count Vectorizer

Topic 0
يوم , الدهر , الناس, الهوى, خير, الزمان, قلبي, الدنيا , الورى, الارض

Topic 1
الهوى, قلبي, الحب, القلب, حب, قلب, الفواد , النفس, الحياة , فوادي

Topic 2
الزمان , الايام , الناس, الدهر, يوم , الملك, الورى, العلى, الهدى, الدنيا

Topic 3
الارض, الاسى, دمي, الموت, الحزن, دم , المدى, الدماء , الدمع, الاولى

Topic 4
الحق, الناس, الاسلام, الاله , الحياة , خير, اليوم , العباد , محمد, الحر

Topic 0: الدهر والزمان
Topic 1: غزل
Topic 2: فخر
Topic 3: رثاء
Topic 4: اسلامي

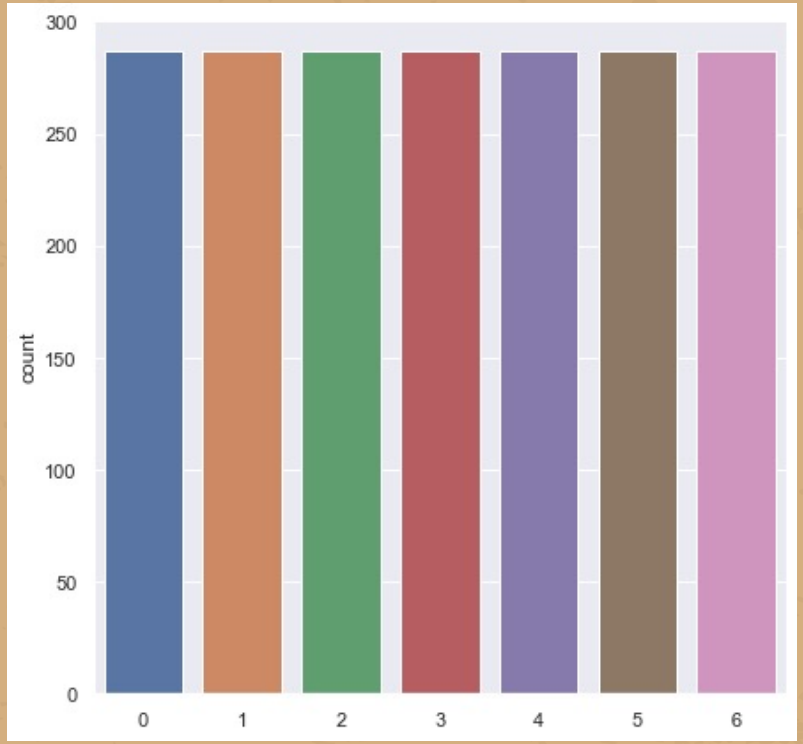
Classification

- XGBClassifier
- Gaussain Naive Bayes
- Support vector machines (SVMs)
- Random Forest
- Multinomial Naive Bayes
- Decision Tree
- Logistic Regression

Handling Data Imbalanced

SMOTE

The total observations before 733
The total observations after: 2009



Best Model

	F1	Accurecy
XGBClassifier	0.44	0.25

Future Work:

1.

Feed more features to the classifiers

2.

Recommendation system

Thanks!