Analysis Quality of Meals Restaurants By ML

★★★★★

Presented by: Annal Albeeshi

OUTLINE

01 INTRODUCTION

02 MOTIVATION

03 DATA DESCRIPTION

04 METHODOLOGY

05 CONCLUSIONS

INTRODUCTION

- we are going to analysis the Arabic language based restaurant reviews for the sentiment analysis. There are two types of sentiments which are 1 = positive and -1 = negative.

MOTIVATION

1- How to deal correctly with Arabic texts.
2- To build a system to classify the sentiments in arabic text based restaurant reviews using machine learning.

Datasets Description:

RES1

- The dataset is a collection of **Arabic texts**, which covers **Dialectal Arabic (DA)**.

- Dataset of **restaurant reviews** scrapped from qaym.com, 8364 reviews.

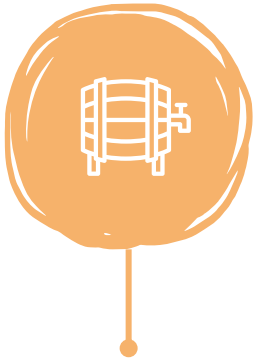
  - The dataset consists of **four columns** and **8364 rows**. It has **two classes** (**Negative and Positive**)

1. Polarity
2. Text
3. Restaurant_id
4. User_id

METHODOLOGY

NLP

Arabic text processing through NLP



Supervised learning

Classification

Unsupervised learning

Clustering

NLP

Arabic text processing

1. Remove the stop words.
2. Remove punctuations.

Arabic & English punctuations.

4. Normalize Arabic text.
5. Remove emojis.
6. Remove repeating char.
7. Stemming– ISRIStemmer

Information Science Research Institute

8. Remove extra whitespace.
9. Remove numbers.
10. Remove https.

"[ًٌٍَُِّْ]", ""

"ي" ,"ى"

"ء" ,"ؤ"

"ء" ,"ئ"

"ه" ,"ة"

"ك" ,"گ"

METHODOLOGY -- Pre-processing

Data Cleaning

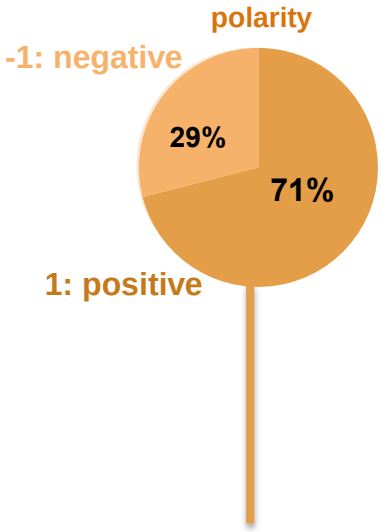
01 **Drop the duplicate values.**
The number of duplicates: 36.

02 **Missing values.**

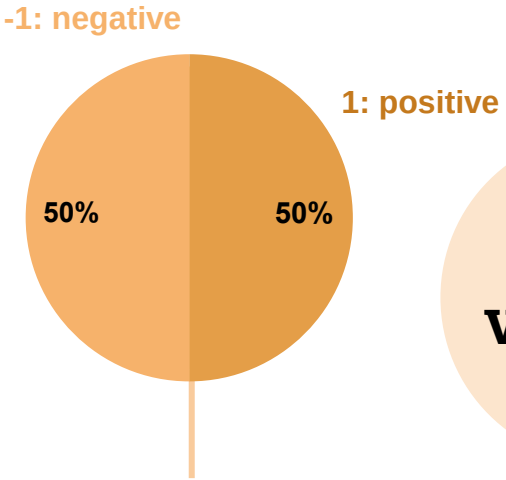
03 **Droping colunams**

Fig:1

Classification



Imbalanced



SMOTE

SMOTE for handle imbalanced

Count Vectorizer

TF-IDF Vectorizer

Classification

- Split the data into 80% training set and 20% testing set.

- Classification Algorithms:

1. LogisticRegression.
2. MultinomialNB.
3. BernoulliNB.
4. SGDClassifier.
5. SVC.
6. LinearSVC.

Classification

Classification Algorithms without SMOTE:

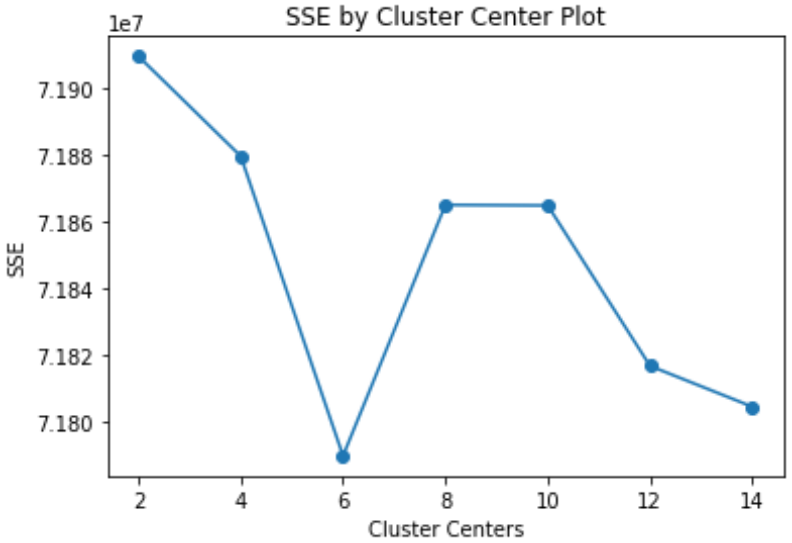
	Train_acc	Train_prec	Train_recall	Train_fbta	Test_acc	Test_prec	Test_recall	Test_fbta
LogisticRegression with CV	0.947734	0.943676	0.985350	0.964063	0.830816	0.857950	0.910883	0.883624
LogisticRegression with TFIDF	0.891541	0.870866	0.995117	0.928855	0.828399	0.821091	0.967438	0.888277
MultinomialNB with CV	0.895770	0.902322	0.957113	0.928910	0.830211	0.847179	0.926307	0.884977
MultinomialNB with TFIDf	0.753474	0.742668	1.000000	0.852334	0.714804	0.712279	0.999143	0.831669
BernoulliNB with CV	0.821148	0.827209	0.946285	0.882749	0.766767	0.784829	0.922022	0.847912
BernoulliNB with TFIDF	0.805438	0.788533	0.992781	0.878947	0.719637	0.718459	0.990574	0.832853
SGDClassifier with CV	0.973565	0.974869	0.988323	0.981550	0.809063	0.866494	0.862039	0.864261
SGDClassifier with TFIDF	0.992900	0.992813	0.997240	0.995022	0.847130	0.866774	0.925450	0.895151
SVC with CV	0.915106	0.901316	0.988960	0.943106	0.810876	0.808526	0.958869	0.877303
SVC with TFIDF	0.993505	0.992196	0.998726	0.995450	0.835045	0.831111	0.961440	0.891538
LinearSVC with CV	0.983535	0.982184	0.994904	0.988503	0.792145	0.853219	0.851757	0.852487
LinearSVC with TFIDF	0.997734	0.997879	0.998938	0.998408	0.849547	0.867200	0.928877	0.896980

Classification

Classification Algorithms with SMOTE:

	Train_acc	Train_prec	Train_recall	Train_fbta	Test_acc	Test_prec	Test_recall	Test_fbta
LogisticRegression with CV	0.975026	0.988771	0.960965	0.974670	0.765306	0.863939	0.785467	0.822836
LogisticRegression with TFIDF	0.984155	0.981830	0.986569	0.984193	0.827731	0.857613	0.901384	0.878954
MultinomialNB with CV	0.933578	0.900543	0.974816	0.936209	0.828932	0.850927	0.913495	0.881101
MultinomialNB with TFIDf	0.993494	0.998727	0.988248	0.993460	0.827131	0.901109	0.843426	0.871314
BernoulliNB with CV	0.766527	0.946240	0.565163	0.707660	0.533013	0.773913	0.461938	0.578548
BernoulliNB with TFIDF	0.881007	0.807973	0.999580	0.893621	0.723890	0.718045	0.991349	0.832849
SGDClassifier with CV	0.991605	0.993470	0.989717	0.991590	0.767107	0.853591	0.801903	0.826940
SGDClassifier with TFIDF	0.998321	0.998949	0.997692	0.998320	0.843337	0.860016	0.924740	0.891205
SVC with CV	0.932739	0.963579	0.899475	0.930424	0.720888	0.798102	0.800173	0.799136
SVC with TFIDF	0.998846	0.999160	0.998531	0.998845	0.778511	0.768968	0.973183	0.859107
LinearSVC with CV	0.996222	0.996848	0.995593	0.996220	0.746699	0.842991	0.780277	0.810422
LinearSVC with TFIDF	0.999685	1.000000	0.999370	0.999685	0.843938	0.863636	0.920415	0.891122

K-means clustering unsupervised



Ring was determined between 2 to 15, and the best Cluster centers 4.

K-means clustering unsupervised

```
Cluster 0
خدم,نظف,فضل,مره,طلب,جرب,لذ,سعر,اكل,طعم

Cluster 1
فطر,روع,ازن,يعب,ابو,حلب,فتش,بيخ,شاورم,عندعم

Cluster 2
رقرز,رقركنج,رقت,صرح,ريض,فتح,نهم,يهم,قهر,ابخ
```

Top Keywords:

Lastly, we'll cycle through the clusters and print out the top keywords based on their TFIDF score to see if we can spot any trends. I'll do this by computing an average value across all dimensions in Pandas, grouped by the cluster label. Finding the top words is simply sorting the average values for each row, and taking the top N.

CONCLUSIONS

- we've gone through all the steps required to design a text classification system for Arabic from NLP to Model Interpretation.

- Summary of NLP Results:

The texts in the dataset have been cleaned up, and the words are returned to their origins, but need improvement.

- Summary of Classic ML Results:

Best classfiers found for the dataset: **LinearSVC with TF-IDF**
Best classfiers found for the dataset**: MultinomialNB with Count Vectorizer**.
Performance across test datasets (numbers represent **precision score**).

- Summary of K-means clustering Results:

Ring was determined between 2 to 15, and the best Cluster centers 4.


THANK YOU
I HOPE YOU'VE ENJOYED