أكاديمية سدايا
SDAIA Academy

NLP

Presented by:
Amer Saleh
Naif Sulaihem
Khalid Alharbi

December 2021

I. Introduction

When bringing any product to the market, it is very important to know the customer review about the product, whether the review is negative or positive, from which the product is evaluated, analyzed, trying to develop the product, etc.

Customer reviews are very important in evaluating any product. We provide a service to analyze reviews, if they are negative or positive, for companies' products, to know the opinions of customers about any product.

II. Study Methodology

The methodology of this project is as follows, extracting data from Kaggle for a Arabic reviews and it was more than 100,000 rows and 2 columns.

In the preprocess step, we convert label as a float using label encoder because it was categorical , and there is no missing values in the dataset and outliers.

We explored the data, applied data cleaning and preprocessing with stopwords and normalize and lemmatizations methods to the data, and extracted important information from the data.

We will predict the review sentiment if that positive, negative, mixed, by analyze the words for each sentence.

III. Data Description

The data set is provided in .TSV format, contains text review and label of sentiment.

The data set was extracted from Kaggle

Variables	Description
Text	Text review.
Label	The sentiment of text.

IV. Tools and Libraries:

Technologies

- Python
- Jupyter Notebook
- PowerPoint for presentation
- web

Libraries

- ArabicLightStemmer
- libqutrub.conjugator
- naftawayh.wordtag
- tashaphyne.stemming
- plotly.graph_objs
- TruncatedSVD
- TfidfVectorizer
- CountVectorizer
- NMF
- strip_tatweel
- strip_shadda
- FarasaPOSTagger
- FarasaNamedEntityRecognizer
- FarasaDiacritizer
- FarasaSegmenter
- FarasaStemmer
- qalsadi.lemmatizer
- pandas
- numpy
- sklearn.linear_model

- sklearn.model_selection
- sklearn.preprocessing
- sklearn.metrics
- matplotlib.pyplot
- seaborn
- string
- nltk
- warnings

V. Classification Model

Model	Train Acc	Train Prec	Train Rec	Train Fbeta	Test Acc
Log Reg CV	0.9204	0.9205	0.9204	0.9203	0.6440
Log Reg TFIDF	0.8330	0.8327	0.8329	0.8326	0.6651
NB CV	0.8088	0.8104	0.8088	0.8084	0.6208
NB TFIDF	0.8165	0.8179	0.8165	0.8166	0.6435

VI. Topic Modeling

1- LSA

LSA	
Topic 0	رواية, كتاب, جدا, كاتب, كانت, احداث, قصة, اخر, نهاية, حياة, بشكل, قراءة, يكون, كثير, شخصية
Topic 1	رواية, احداث, شخصيات, نهاية, روايات, شخصية, قصة, بطل, لرواية, مراد, كاتبة, وصف, سرد, حبكة, تفاصيل
Topic 2	جدا, فندق, غرف, استقب, غرفة, موقع, جيد, يوجد, سيء, متاز, خدمة, مكان, نظافة, غرفه, فقط
Topic 3	فندق, كانت, انسان, حياة, اذا, حب, اخرى, يكون, عندما, وان, اخر, فقط, يوم, رجل, دين

2- LDA

LDA	
Topic 0	فندق, جدا, موقع, غرف, استقب, جيد, غرفة, يوجد, سيء, ضعيف, نظافة, مكان, فقط, خدمة
Topic 1	كتاب, رواية, كاتب, كانت, حياة, كثير, قراءة, حب, اخرى, اخر, عندما, يكون, وان, انسان
Topic 2	رواية, كتاب, جدا, كاتب, كانت, احداث, روايه, قصة, اسلوب, ده, نهاية, كتير, فكرة, مكن

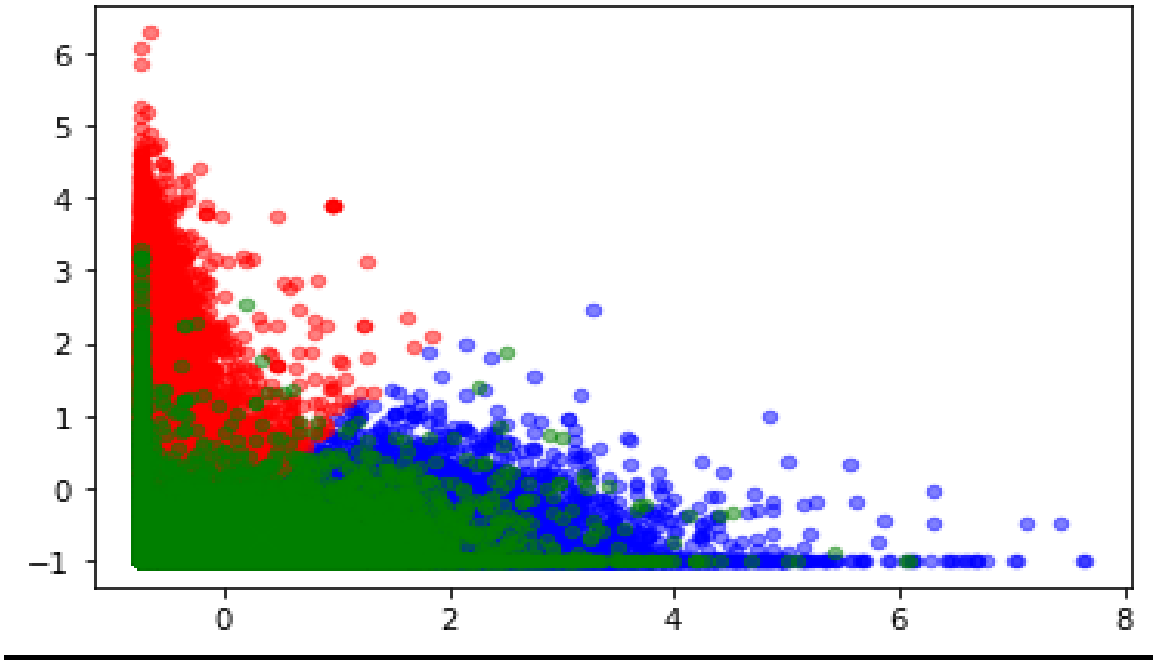
3-NMF

NMF	
Topic 0	رواية, كاتب, كانت, احداث, قصة, نهاية, شخصيات, شخصية, جدا, حياة, اخر, بشكل, حب, فكرة, روايات, اسلوب, عمل, رغم, تكون, قراءة
Topic 1	كتاب, كاتب, كتب, كانت, قراءة, اخر, يكون, كثير, انسان, حياة, دين, جدا, وان, ناس, اسلام, اخرى, فصل, فكرة, بشكل, فقط
Topic 2	جدا, فندق, غرف, استقب, غرفة, موقع, جيد, يوجد, كانت, سيء, فقط, متاز, مكان, خدمة, نظافة, راءع, غرفه, ضعيف, وجود, مواقف

Text	others	movies	books
ممتاز نوعا نظافة وموقع وتجهيز وشاطئ مطعم	0.022105	0.0000	0.00000
ارجع يه مرة اخرى قربه بحر مكان قديم توجد خدمات نجوم	0.018286	0.007591	0.003249
كتاب ضعيف جدا ولم استمتع قصه سرد لحه مشهد بدون فكره لقصه	0.00000	0.020182	0.040226

VII. Clustering

1-KMeans



VIII. Summary

1- Test Case

Test case	Results
فلم جميل و اختيار موفق كريستوفر نولان بتحفة فنية جميلة اخرى و تمتثيييلل ولا اروع من الاسطورة ليوناردو دي كابريو	Positive
كتاب سيئ و رواية مفرعة يوجد بها خروج عن النص بشكل فظيع قد يكون الكتاب الاخير الذي سأشتريه لهذا الكاتب	Negative
فندق مقبول المواقف قديمة الاثاث مستحدث انصح به اصحاب الميزانية المحدودة قد يكون خيار جيد	Mixed

2- Conclusion

We do the lemmatization and stemming with multi libraries, we decide qalsadi libraires is best one.

We found the best NLP practice for our data is Topic Modeling, with NMF, the cluster didn't give us a good result, and for classification model Logistic Regression with TFIDF.

Reference :

https://www.kaggle.com/abedkhooli/arabic-100k-reviews