Arabic Sentiment (عربي)

Big Data project

Dataset link

https://www.kaggle.com/code/ayaelneanaei/notebooka514080012/edit

Detailed Overview

Objective

Dataset

preprocessing

Model

Result

Conclusion

Objective

Our Arabic Compan Reviews (عربي)

Sentiment Analysis on Arabic Companies Reviews" presents an innovative approach to sentiment analysis in the context of the Arabic language. It utilizes a dataset of over 100,000 reviews related to Arabic companies. The main objective is to develop a reliable sentiment scoring system for businesses. The methodology includes reprocessing steps for refining the dataset, such as transforming emojis and removing Arabic diacritics, and using machine learning models like Logistic Regression, Naive Bayes, and Linear Regression. These models are adapted for sentiment analysis in Arabic text, considering the unique challenges of the language such as script variations and dialects.


Canva

Dataset:

Our Dataset

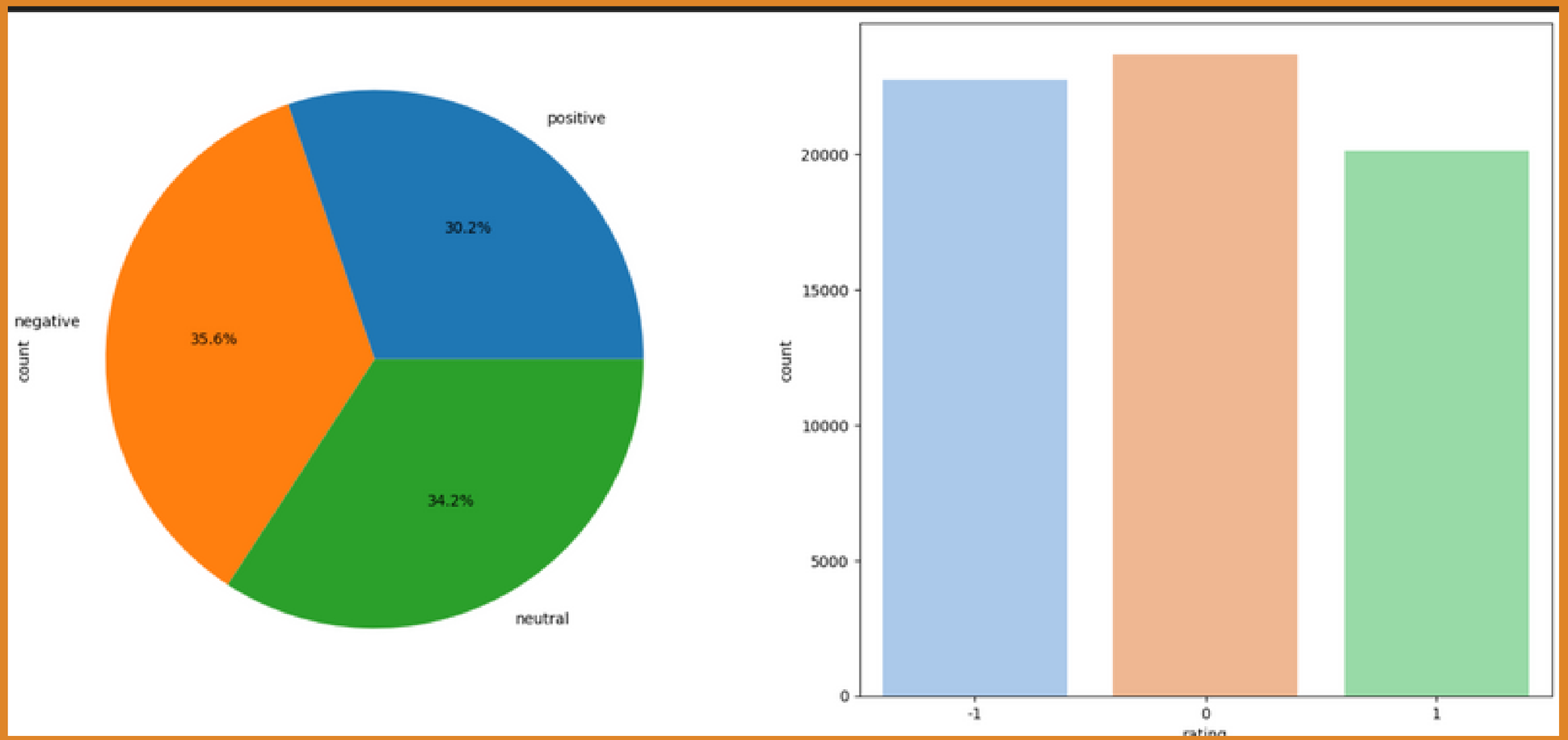
A collection of over 100,000 Arabic-language reviews."
Each review is associated with specific companies and labeled with sentiment ratings.

Review Distribution Highlights:

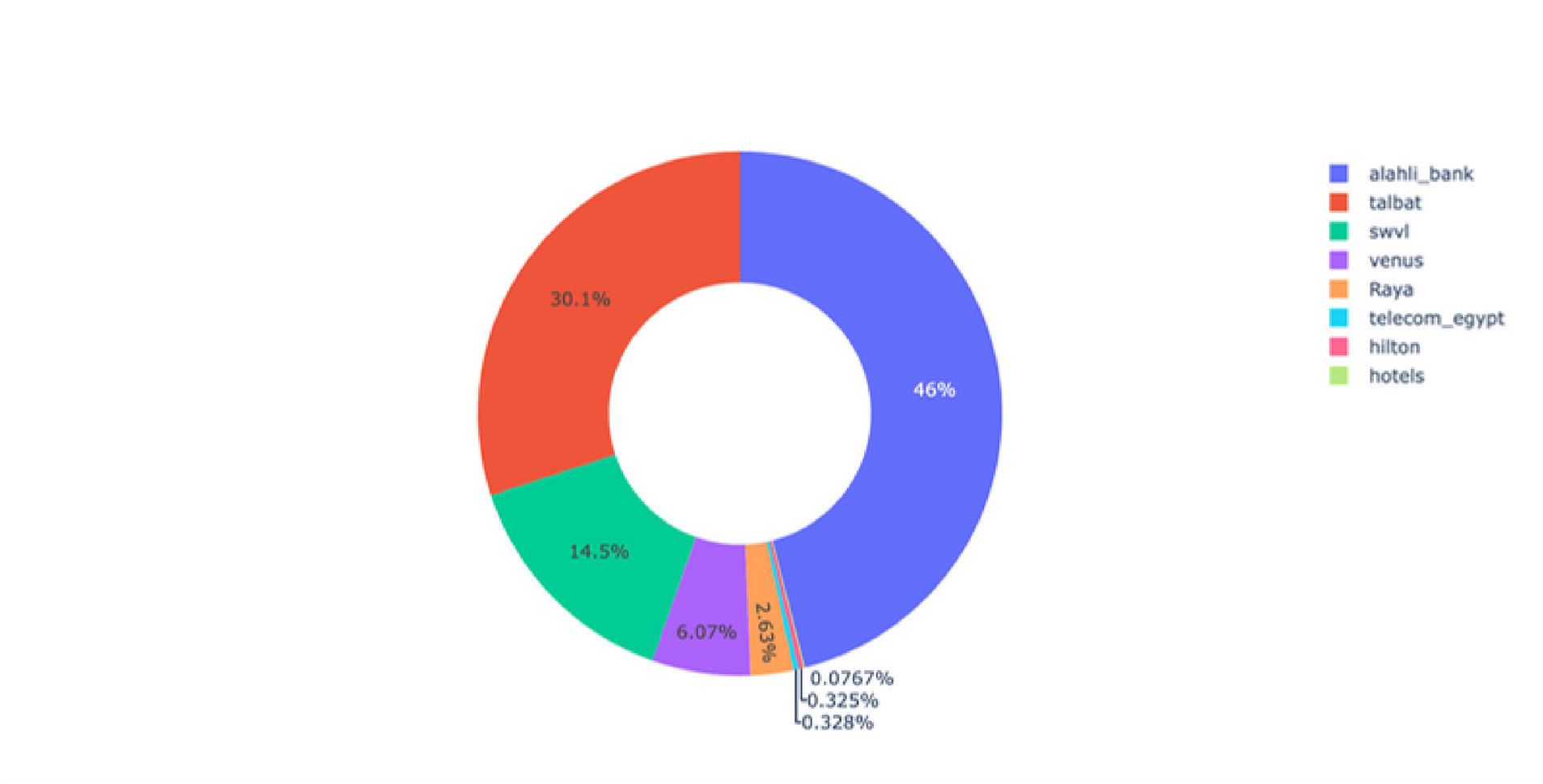
Alahl Bank leads with 47.7% of reviews.
Talbat and Swvl follow with 27.6% and 13.3%, respectively.
Other notable mentions include Venus, Raya, and smaller companies like Capiter, TMG, Ezz Steel, Domty.

Methodology

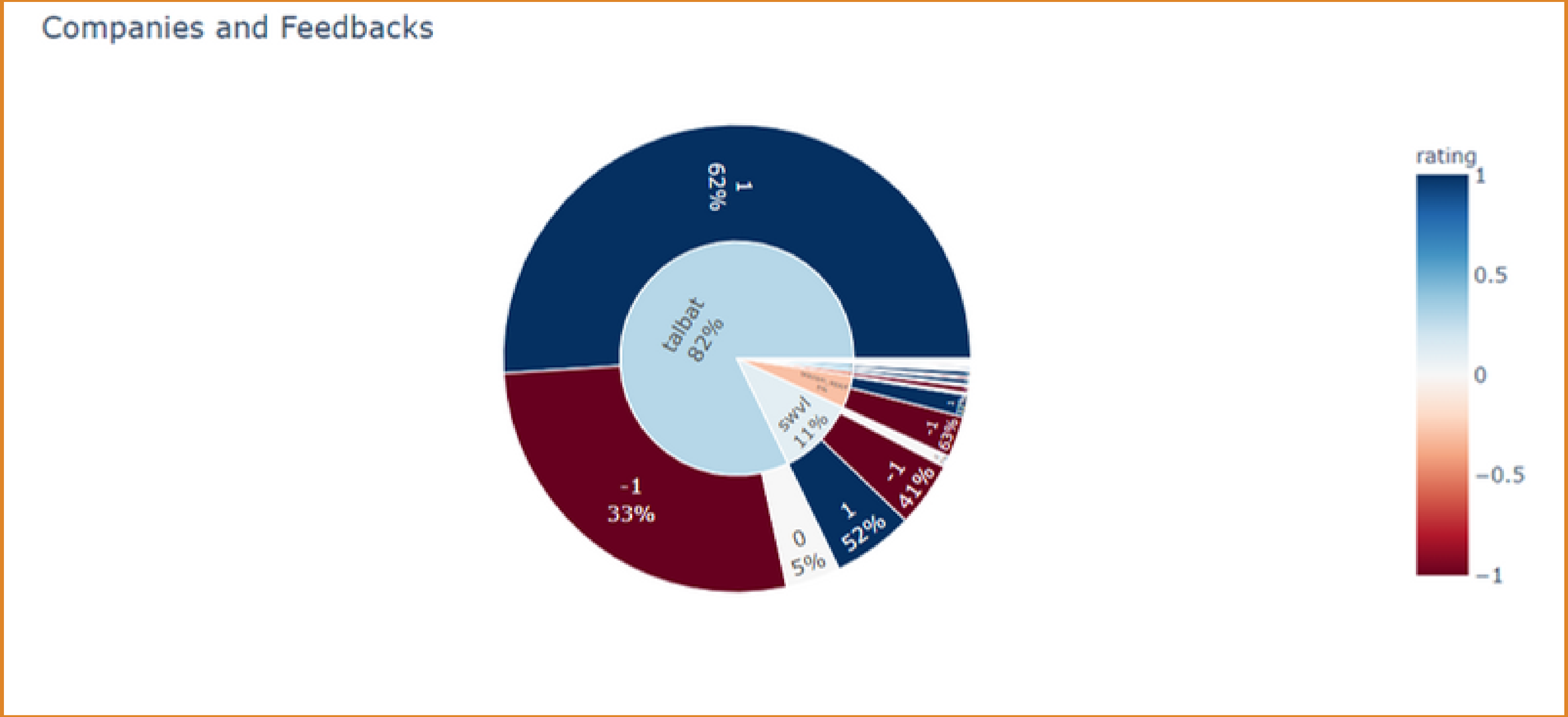
presents a pie chart illustrating that the positive ratings (1) outnumber both the neutral (0) and negative (-1) ratings



illustrates that 'Talabat' has received the most reviews in the dataset, accounting for 82.2%, which is a higher percentage than any other company.



illustrates the feedback for various companies, highlighting, for example, that 'Talabat' received 62% positive, 33% negative, and 5% neutral reviews in their overall feedback distribution.



Pre-processing -Clean text

1 Check Nulls Values

```
+------------------+------+-------+
|review_description|rating|company|
+------------------+------+-------+
|               407|   891|   1376|
+------------------+------+-------+
```

2 check duplicated

```
Total number of duplicate entries:  138
```

4 Removing mentions

5 Removing tages

6 Removing number

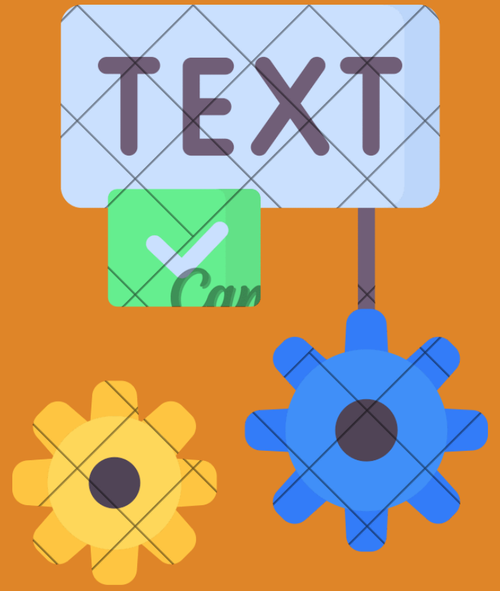
7 Removing all Viaa

8 Remove tashkeel

9 Remove tatweel

10 removeExtraChar

11 Remove links



Remove English

Stopword Removal

duplicated words

replace emoji with text

RemoveSpecialChars

Tokenization

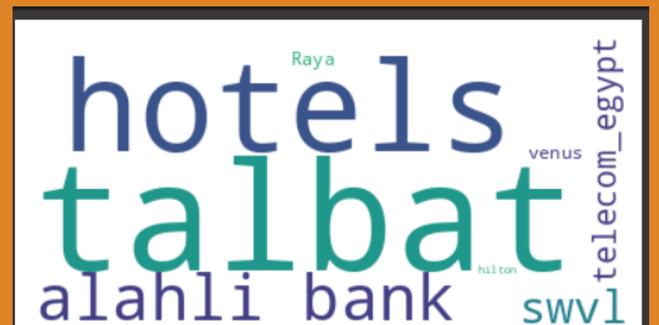
Normalization

Handling Arabic-specific Challenges

Stemming & Lemmatization

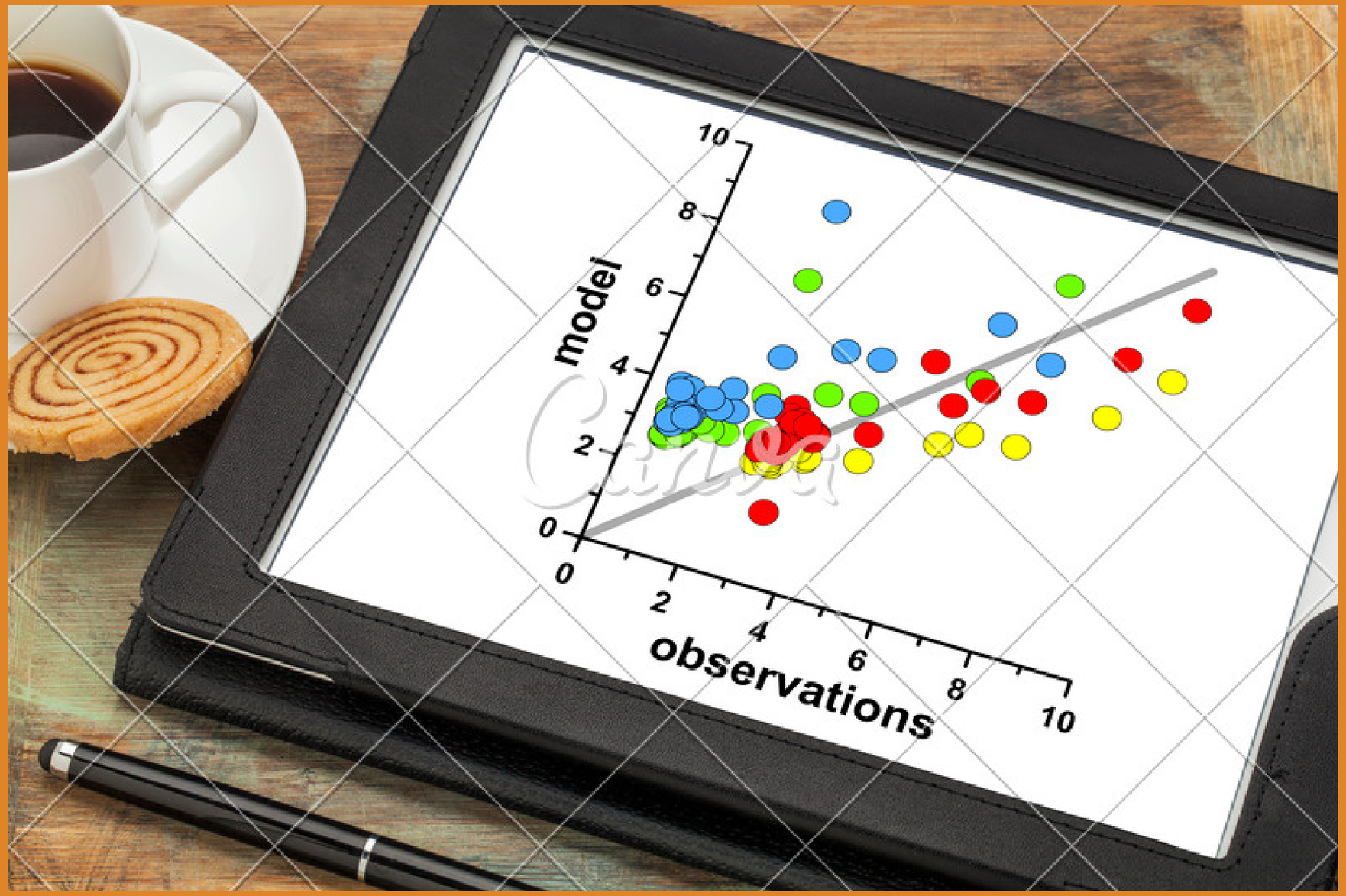
Spell Checking and Correction

Word Cloud



Franco translate

Models



TF-IDF

Term Frequency - Inverse Document Frequencyt

Nurmerical Statistic: reflect how important a word is a document in a collection

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

After TF-IDF

check the number of rating values in the train data.

```
label_counts =
train_df.groupBy('labeli').count()
```

labeli	count
1.0	18045
0.0	18946
2.0	15795

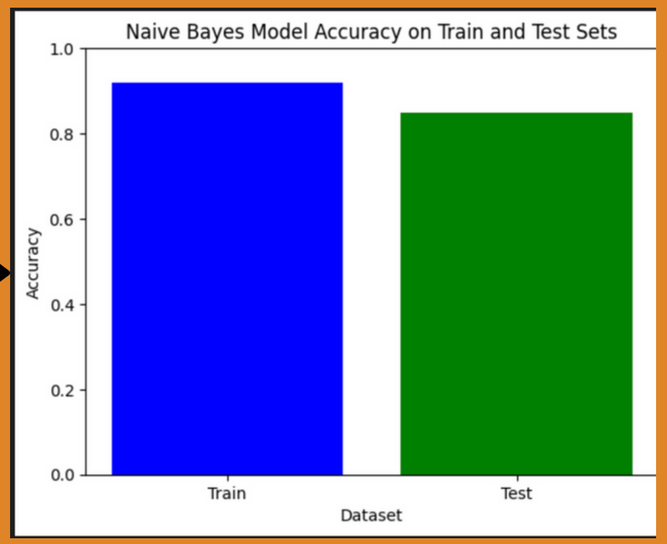
The numbers are close to each other, I do not need them to be equal.

NaiveBayes Model

Naive Bayes is effective for predicting review ratings due to its ability to efficiently handle text classification tasks by modeling word relationships with ratings, making it well-suited for sentiment analysis.

The output for the Naive Bayes model is as follows:

- Accuracy on the training set: 91.89%
- Accuracy on the test set: 84.81%

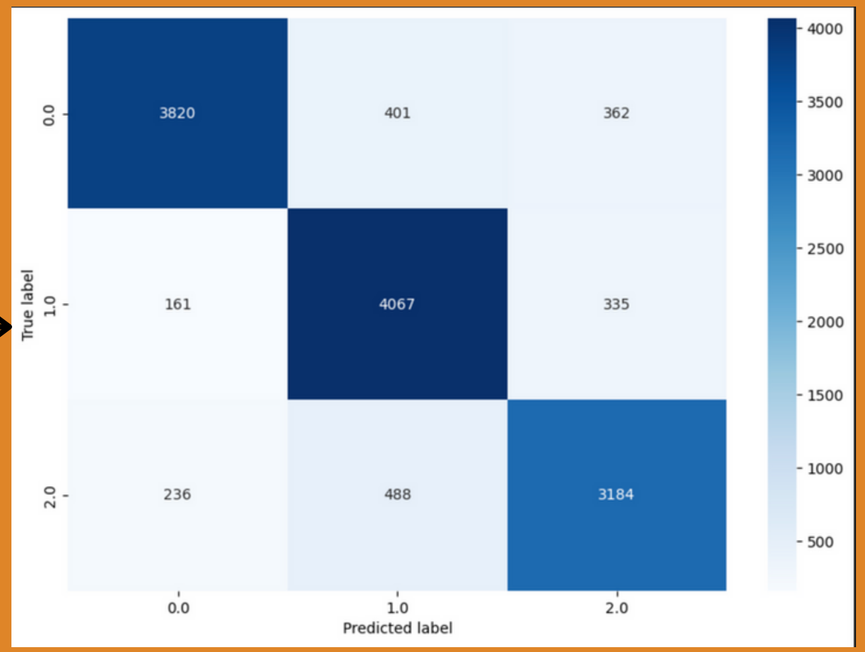


Test Model

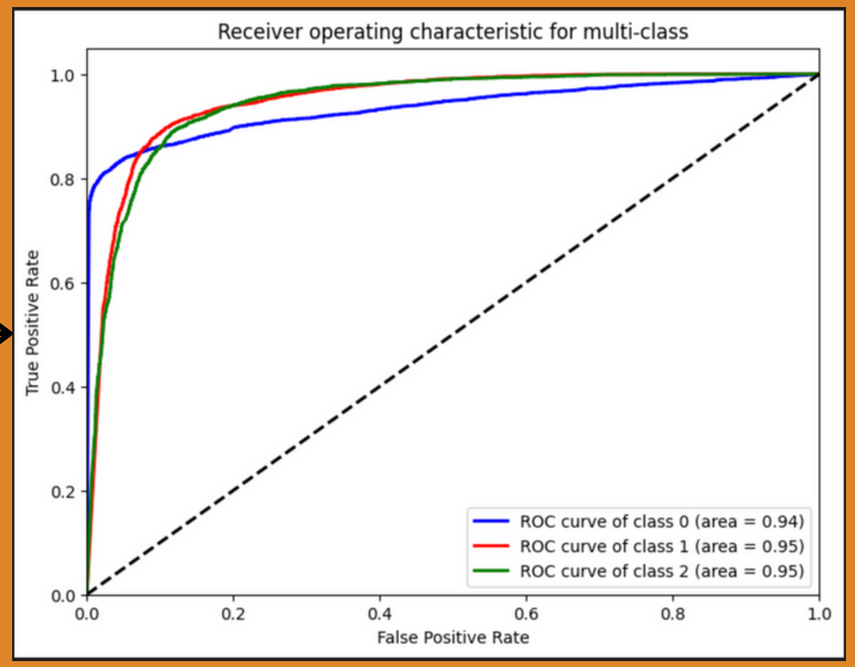
```
new_text = "إنـه جميـل "
```

```
'إنـه جميـل ' is: Positive
```

Predicted Label For Naive Bayes Model



Receiver operating characteristic for multi-class



Second Model Linear Regresstion

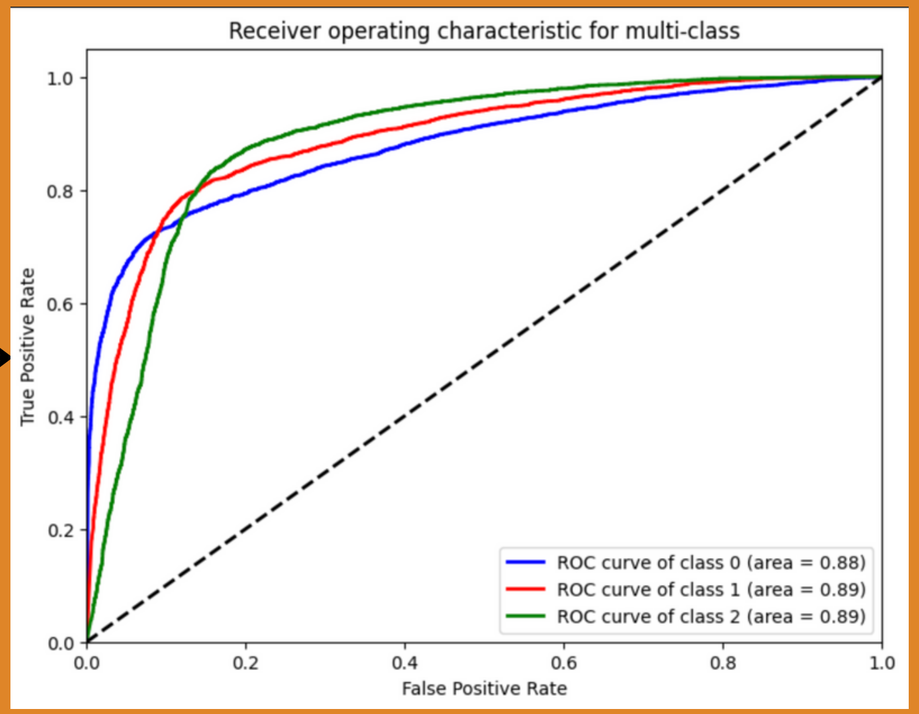
The Root Mean Squared Error (RMSE) for the model is 1.08.

```
Root Mean Squared Error (RMSE): 1.0756346689076184
```

Third model LogisticRegression Model

```
Accuracy: 0.7685766814769419
```

Receiver operating characteristic for multi-class



When need to improve accuracy we make parameter grid

```
Logistic Regression (CV) Accuracy: 0.8463306266278535
Best Max Iter:  5
Best Reg Param:  0.01
Best Elastic Net Param:  0.0
```


CONCLUSION
00:01:16:22
94.05%

CONCLUSION

1. **Data Preparation for Advanced Machine Learning**
2. **Challenges and Feature Engineering**
3. **Model Application and Visualization**

Thank You

For Your Attention

The end