Movie Genres

Hamza Al-Habash

ID:19110034

HTU: School of Computing and Informatics

Date: 2021/11/10

Problem statement

- My head of department has placed me in the development team to help correctly classify a movie dataset which includes for each movie a small description and the genre the movie belongs to. The dataset consists of 85,855 movies with 1257 movie genres.

- I have been tasked to develop a top-down AI application using NLP and machine learning to build a model which can classify movies to their correct genre using the movie description only.

- For my task, I’ll pre-process and build classifiers for classifying “Horror” and "Comedy, Romance” , and another classifier(s) for classifying "Drama" and "Comedy" .

Goals/Aims Of The Project

- Assist to correctly classify a movie dataset which includes for each movie a small description and the genre the movie belongs to.

- Developing a top-down AI application using NLP and machine learning to build a model which can classify movies to their correct genre using the movie description only.

- Pre-processing & building classifiers for classifying “Horror” and "Comedy, Romance” movie genre, and another classifier(s) for classifying "Drama" and "Comedy“ movie genre .

The Methods Used To Develop The Project

- To pre-process the data, I’ve used stop_words to remove br tags, punctuation, numbers, and stopwords.

- To analyze the dataset I’ve used two word representation models: Bag of Words and TF-IDF.

- Also, I’ve used two classification algorithms: Naïve Bayes and Logistic regression to classify the dataset.

- **For each word representation and classification model I’ve included the following:**

  - All text pre-processing steps.
  - The classification accuracy.
  - The confusion matrix.
  - Using the LIME package, I’ve shown Visualization of top features. Including two sentences for demonstration purposes.

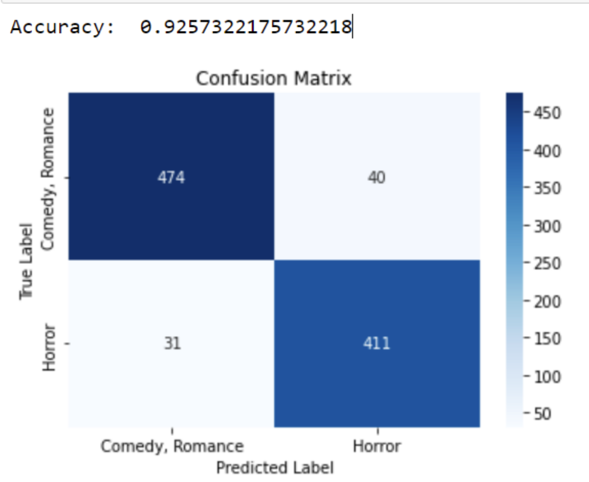
The Performance Measure Of My Models

Comparing confusion matrices and classification accuracies

[“Horror” and "Comedy, Romance”]

❖ Bag of Words

Naïve Bayes

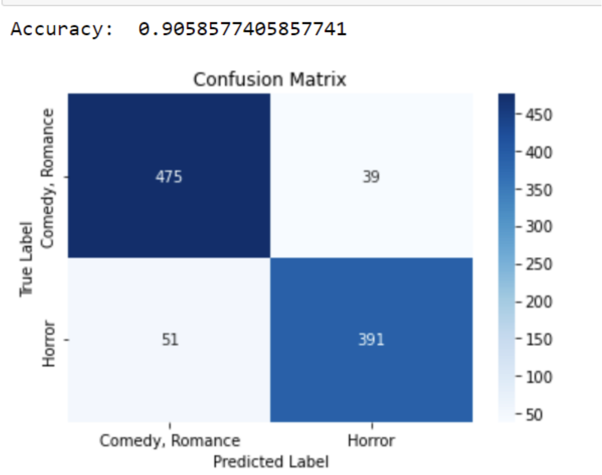


```
Accuracy:  0.9257322175732218
```

I found that in both the Bag of Words & TF-IDF, the Naïve classifies the data with a higher classification accuracy, and better confusion matrix than a Logistic does

I found that the TF-IDF Naïve classifies the data with a higher classification accuracy, and better confusion matrix than the Bag of Words Naïve.

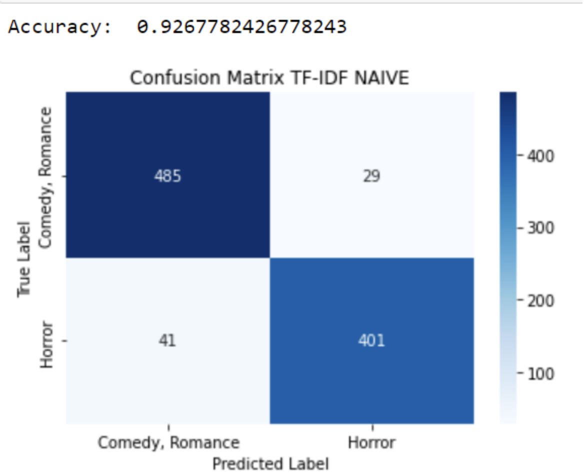
Logistic regression



```
Accuracy:  0.9058577405857741
```

❖ TF-IDF

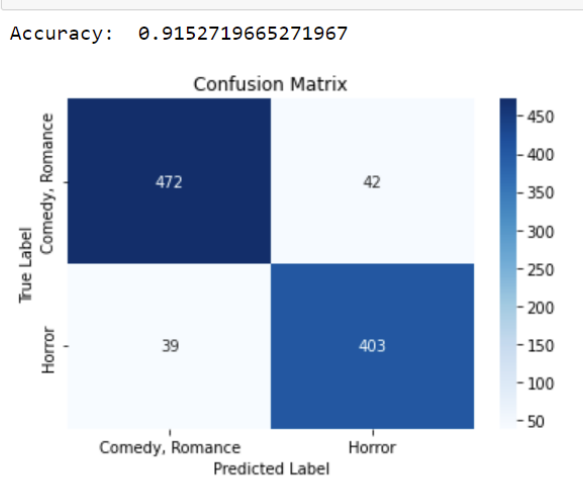
Naïve Bayes



```
Accuracy:  0.9267782426778243
```

I found that the TF-IDF Logistic classifies the data with a higher classification accuracy, and better confusion matrix than the Bag of Words Logistic.

Logistic regression



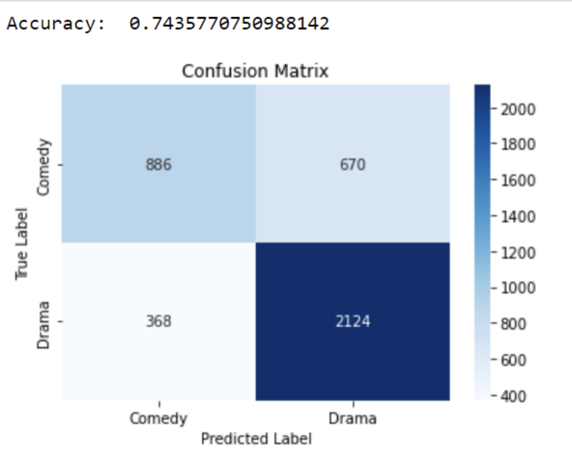
```
Accuracy:  0.9152719665271967
```

The Performance Measure Of My Models

[“Drama” and "Comedy”]

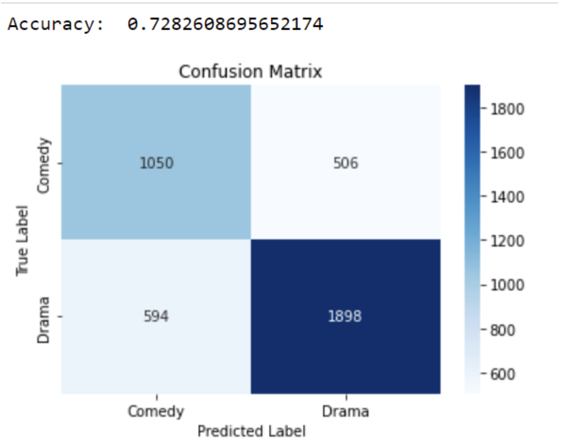
❖ Bag of Words

Naïve Bayes



```
Accuracy:  0.7435770750988142
```

Logistic regression



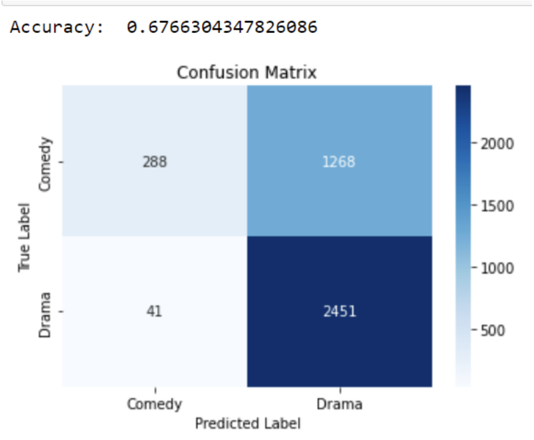
```
Accuracy:  0.7282608695652174
```

I found that in the Bag of Words, the Naïve classifies the data with a higher classification accuracy, and better confusion matrix than a Logistic does

I found that in the TF-IDF, the Logistic classifies the data with a higher classification accuracy, and better confusion matrix than a Naïve does

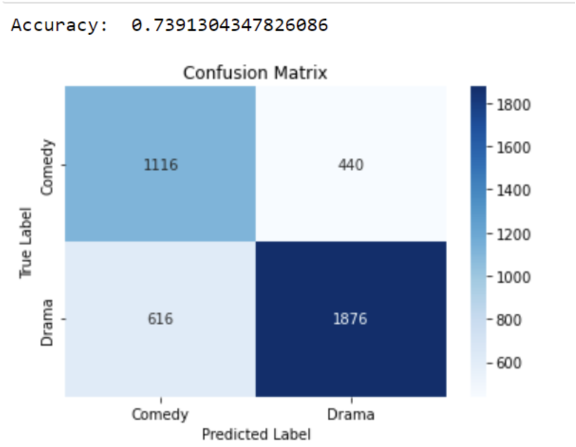
❖ TF-IDF

Naïve Bayes



```
Accuracy:  0.6766304347826086
```

Logistic regression



```
Accuracy:  0.7391304347826086
```

I found that the Bag of Words Naïve classifies the data with a higher classification accuracy, and better confusion matrix than the TF-IDF Naïve.

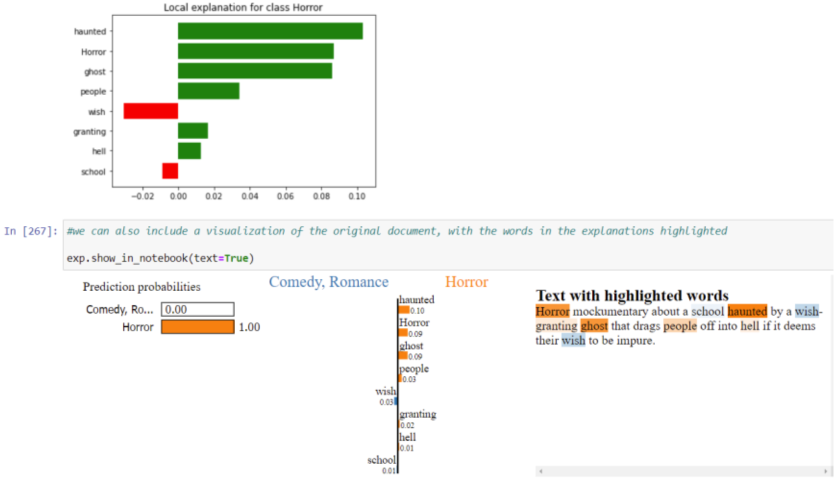
I found that the TF-IDF Logistic classifies the data with a higher classification accuracy, and better confusion matrix than the Bag of Words Logistic.

Example of LIME visualization

[“Horror” and "Comedy, Romance”]

❖ Bag of Words

Naïve Bayes



```
In [267]: #we can also include a visualization of the original document, with the words in the explanations highlighted

          exp.show_in_notebook(text=True)
```

Prediction probabilities

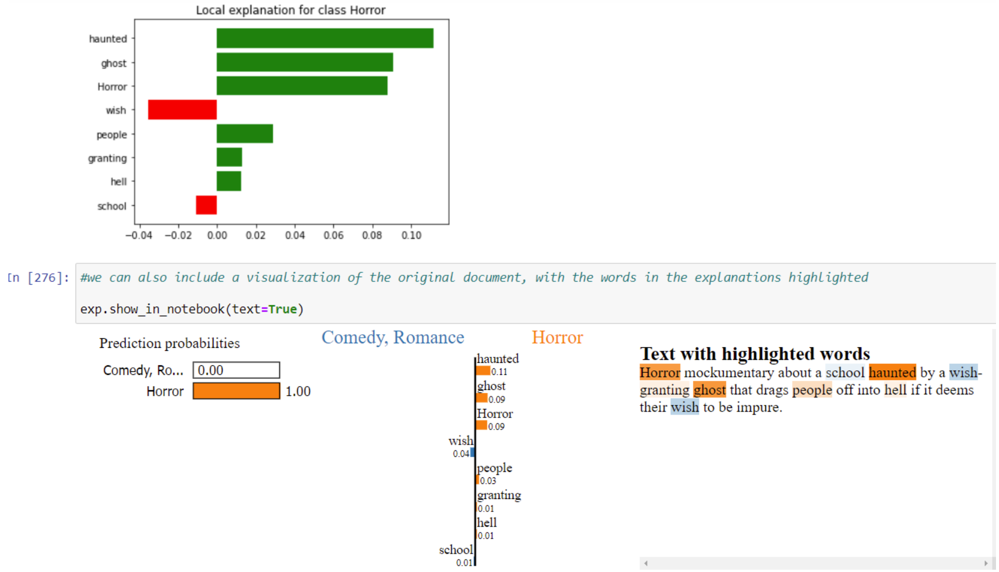
Comedy, Ro... 0.00

Horror 1.00

Text with highlighted words

Horror mockumentary about a school haunted by a wish-granting ghost that drags people off into hell if it deems their wish to be impure.

Logistic regression



```
[n [276]: #we can also include a visualization of the original document, with the words in the explanations highlighted

          exp.show_in_notebook(text=True)
```

Prediction probabilities

Comedy, Ro... 0.00

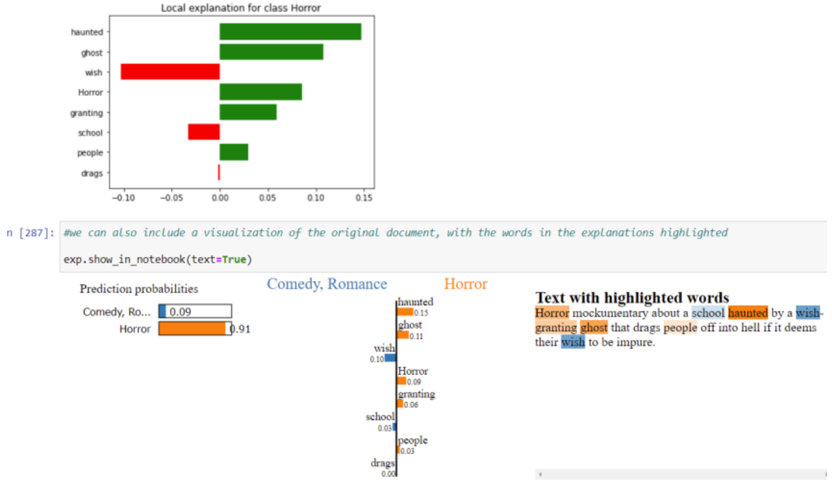
Horror 1.00

Text with highlighted words

Horror mockumentary about a school haunted by a wish-granting ghost that drags people off into hell if it deems their wish to be impure.

❖ TF-IDF

Naïve Bayes



```
n [287]: #we can also include a visualization of the original document, with the words in the explanations highlighted

         exp.show_in_notebook(text=True)
```

Prediction probabilities

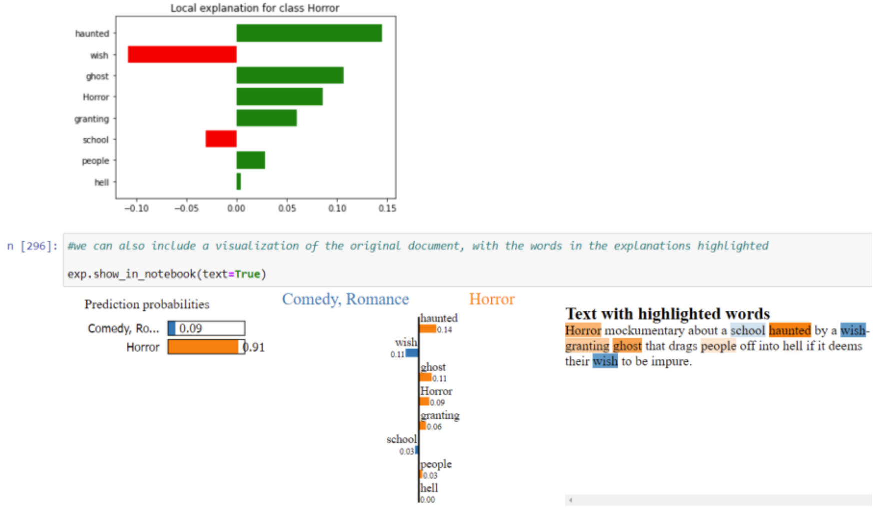
Comedy, Ro... 0.09

Horror 0.91

Text with highlighted words

Horror mockumentary about a school haunted by a wish-granting ghost that drags people off into hell if it deems their wish to be impure.

Logistic regression



```
n [296]: #we can also include a visualization of the original document, with the words in the explanations highlighted

         exp.show_in_notebook(text=True)
```

Prediction probabilities

Comedy, Ro... 0.09

Horror 0.91

Text with highlighted words

Horror mockumentary about a school haunted by a wish-granting ghost that drags people off into hell if it deems their wish to be impure.

Discussing The Differences In Classification Accuracies

[“Horror” and "Comedy, Romance”]

❑ I found that in both the Bag of Words & TF-IDF, the Naïve classifies the data with a higher classification accuracy, and better confusion matrix than a Logistic does.

❑ I found that the TF-IDF Naïve classifies the data with a higher classification accuracy, and better confusion matrix than the Bag of Words Naïve.

❑ I found that the TF-IDF Logistic classifies the data with a higher classification accuracy, and better confusion matrix than the Bag of Words Logistic.

[“Drama” and "Comedy”]

❑ I found that in the Bag of Words, the Naïve classifies the data with a higher classification accuracy, and better confusion matrix than a Logistic does.

❑ I found that in the TF-IDF, the Logistic classifies the data with a higher classification accuracy, and better confusion matrix than a Naïve does

❑ I found that the Bag of Words Naïve classifies the data with a higher classification accuracy, and better confusion matrix than the TF-IDF Naïve.

❑ I found that the TF-IDF Logistic classifies the data with a higher classification accuracy, and better confusion matrix than the Bag of Words Logistic.

❑ [“Horror” and "Comedy, Romance”] vs [“Drama” and "Comedy”]:

❑ According to my results the Horror & Comedy, Romance has a much better & higher classification accuracies, and better confusion matrices.

❑ Why this happened? The data among these two is far different which includes the number of words, size of the data, quality of the datas and many other factors that contributes to a different accuracies

Main Conclusions

- ❑ In an overall view, I think that the [“Horror” and "Comedy, Romance”] model performed better as we saw before because the quality of its data must & many other factors be more accurate than the [“Drama” and "Comedy”] model.

- ❑ In my perspective, the model I’ve built for sure it’s ready to go to the next stage in the model development pipeline, because the data was processed and cleaned before starting building the project.

Future Work

- For me, the data pre-processing, data quality, data size was cleaned and very good, there was no N/A values and data size is manageable I didn’t have any obstacles when analyzing it.

- As a future work we may try to use and try many other types of word representation models & classification algorithms, so we can have a different results and decide which is better for our projects.