Report

1	[{'id': 35, 'name': 'Comedy'}]	When Lou, who has become the "father of the Internet," is shot by an unknown assailant, Jacob and Nick fire up the time machine again to save their friend.
2	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10751, 'name': 'Family'}, {'id': 10749, 'name': 'Romance'}]	Mia Thermopolis is now a college graduate and on her way to Genovia to take up her duties as princess. Her best friend Lilly also joins her for the summer. Mia continues her 'princess lessons'- riding horses side-saddle, archery, and other royal. But her complicated life is turned upside down once again when she not only learns that she is to take the crown as queen earlierthan expected...
3	[{'id': 18, 'name': 'Drama'}]	Under the direction of a ruthless instructor, a talented young drummer begins to pursue perfection at any cost, even his humanity.
4	[{'id': 53, 'name': 'Thriller'}, {'id': 18, 'name': 'Drama'}]	Vidya Bagchi (Vidya Balan) arrives in Kolkata from London to find her missing husband Arnab Bagchi. Seven months pregnant and alone in a festive city, she begins a relentless search for her husband. With nothing to rely on except fragments from her memories about him, all clues seem to reach a dead end when everyone tries to convince Vidya that her husband does not exist. She slowly realises that nothing is what it seems. In a city soaked in lies, Vidya is determined to unravel the truth about her husband - for herself and her unborn child - even at the cost of her own life.

An example of a dataset movies

1- Preprocessing

To begin with, we first did a preprocessing of all the data (including test data and the most). Of the 3, the removal of stop words and other items that were less important in this database was continued. We then applied to stem and lemmatizing to get cleaner and more appropriate texts.

2- Embedding

In this section, with the help of word2vec, we implement the Embedding layer to get meaningful vectors of word properties from raw words and data.

3- Representation of corpus:

In this step, you have represented the paragraphs in three ways, which are as follows.

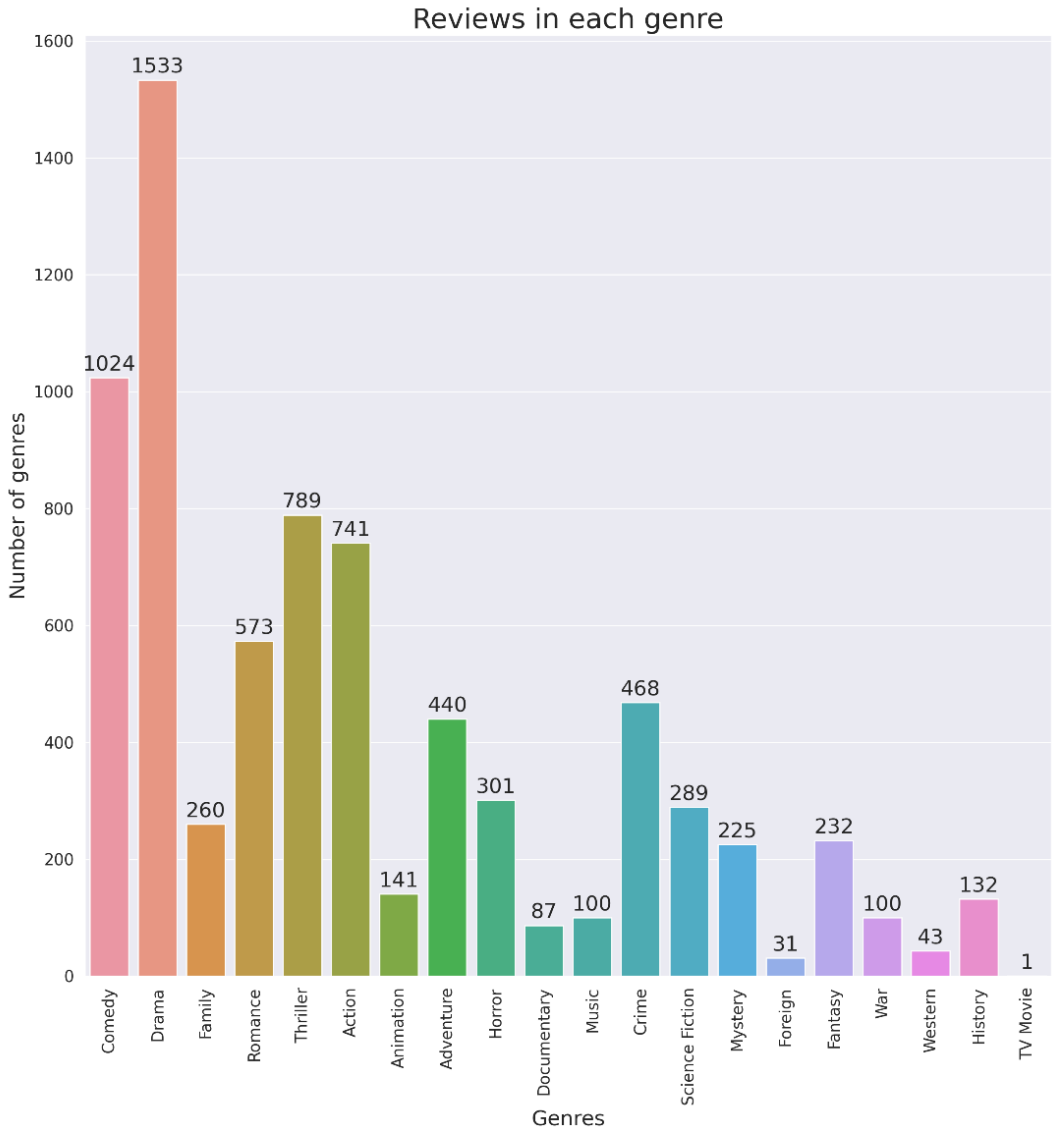
Average words of the paragraph: In the first method, we took a simple average of the words in the text according to the vector obtained from the words in the previous step. I also normalized the resulting vector. After I displayed the words in word2vec with 400 attributes in this section, the reviews do not have more than 400 attributes. Due to the length and size of the text, this number is small, but we must continue to check to be sure of the results.

Bag of Words: In the second method, by calculating tf-idf and examining the words in the text, a very large space to represent the decimal point, whose dimensions were equal to the number of words in the entire data.

Weighted average: In the third method, using the tf-idf obtained from the second method and the word vectors, I decide to take a weighted average of the words inside a paragraph instead of a simple mean. There is still the problem of lack of features, but the way of averaging logic is more than the first case.

4- Classification

First, data analysis was performed to provide a statistical analysis of the data and how the tags were distributed. The results were as follows:



Genre charts based on the number of reviews that have that genre

The classifiers we used in this step include MLP, svm_classifier, and kn_classifier.

Also, the evaluation was done by Jaccard. In this method, if, for example, one of the three labels is identified correctly, it receives one third of the points.

MLP

After testing several MLPs with different architectures, we chose the following architecture. The layers include 200, 100, 50 perceptrons and the relu activation function, respectively, and in the last layer, 20 perceptrons with the sigmoid activation function. By testing different values for epoch and batch_size, we reached batch_size = 30, epochs = 30 and determined the output

using a threshold. In this way, if it was more than the threshold, we would consider that tag as 1, and if it was less than that, we would set it to 0.

The max results obtained in this section in the 3 phases mentioned (3 methods of representation of corpus) are as follows:

Phase1:

Threshold = 2.2 , test jaccard = 0.28

Phase2:

Threshold = 0.01, test jaccard = 0.37

Phase3:

Threshold = 2.2 , test jaccard = 0.28

SVM with linear kernel

The results obtained in this section in the 3 phases mentioned (3 methods of representation of corpus) are as follows:

Phase1:

Train Jaccard = 0.14
Test Jaccard = 0.14

Phase2:

Train Jaccard = 0.34
Test Jaccard = 0.23

Phase3:

Train Jaccard = 0.14
Test Jaccard = 0.15

KNeighborsClassifier

The results obtained in this section in the 3 phases mentioned (3 methods of representation of corpus) are as follows:

Phase1:

Train Jaccard = 1.0
Test Jaccard = 0.14

Phase2:

Train Jaccard = 1.0
Test Jaccard = 0.21

Phase3:

Train Jaccard = 0.24
Test Jaccard = 0.09

5- Conclusion

As you can see, of the various methods mentioned during the report, the best method that had the closest result to our goal and the best training we saw in it was the second method or Bag of Words. In the other two methods, the biggest problem we encountered was the unequal distribution of tags in the data, which caused our Agent to learn to label all the data in one way, in addition to the limitations of the model of those two methods. Reach well. We also saw in MLP that from a threshold onwards, exactly the same thing happened for the first and third methods, and like other models like SVM with a linear kernel, it returns a constant value for accuracy and evaluation, which my impression of this The subject was the same learning error that I explained. This problem is possible from two issues. The first is the very small amount of data in the data and its small size, which showed itself in all parts of the work. For example, for MLP, the small amount of data meant that we did not have complete and good training, or for SVM, it did not show enough state so that SVM could find the right space and have good learning.

Another issue was how to model and embed raw data. As we saw, the averaging method was not suitable at all, and the only place we saw a little learning was in the second method, or Bag of Words. With this in mind, Bag of Concept is expected to give the best results on this database. A very important point in implementing this method is proper clustering of the words we got from word2vec. In the incomplete implementation I had of this part, by running mean-shift on all words, 14 clusters were automatically identified, which is not appropriate at all due to a large number of words in the database (mean-shift was chosen because the number of clusters We do not know (so by solving this problem and doing new modeling of critiques, there is a very high probability that we will achieve a better result.

In the end, the best score we got from the Jaccard was about 37%, which was achieved in a very low threshold in the second method with the help of MLP. It may then use Bag of Concept or other methods to embed data other than word2vec (such as glove) to get better accuracy and get better results from the same low data.