Ravi Patel and Dylan Toothaker
Info 490C Final Report

Analysis on Movie Genres from Different Reviews

Introduction:

In recent times, movies have become a staple of our culture, and can be considered as one of the top pastimes all around the world. Although, when we choose a movie, one of the main things that we classify by is what genre the viewer is planning to watch. We watch movies to pass time, and we sit down for a couple hours to enjoy the screen. We continuously grow curious as to how much more information is available to the public behind the scenes. Leading us to our core research question, given a selection of reviews and genres for a movie, how do people describe the individual genres through the movie reviews. In addition to wondering how each individual genre overlaps/compares with other genres? Can the extracted features from movie reviews predict the genre of the movie? We plan to use much of what we have learned in class to assist us in our process to find out what makes the reviews unique to each of the respective genres. We hope to use Comparison and Clustering to determine which genres are most closely related, as well as the least related. We can use Classification and Feature Analysis to specifically look at the similarities and differences between related, or not so related, genres.

Related Work:

[2] A study has been done by Quan Hoang, also from the University of Massachusetts Amherst, where he attempts to predict 20 movie genres based off of the plot summaries using natural language processing (Word2Vec). At the same time, he also compares genres to one another using a Naive Bayes classifier and the Jaccard similarity. The focus of our project is to understand the relation between movie reviews and genres by using feature analysis to investigate the relationship between two genres. Moreover, we also plan to use the multinomial Naive Bayes classifier to investigate the similarities between genres through their reviews.

[3] Sandipan Dey has also conducted a similar analysis to ours, where she focuses on exploratory analysis on movie-ratings along with Fraud-Detection with Credit-Card Transactions. Her analysis focuses on movie ratings across different genres where she uses seaborn to plot multiple different graphs for each genre, and this is similar to ours because one of the additional packages that we used to graph our results was seaborn as well. Additionally, when analyzing the fraud-detection, similar to us, she uses the

sklearn package models to predict the test dataset and test the accuracy as well. Although we aren't using the packages in the exact same way, they provide us with some guidance and some insight on how we could analyze our dataset differently than we had been.

Dataset:

Our dataset is from IEEE DataPort and is titled IMDB Movie Reviews Dataset. There are actually two datasets given from this page, a folder called 1_movie_per_genre and a folder called 2_reviews_per_movie_raw. For the purpose of this analysis we will be using the first folder. This folder contains seventeen different CSVs, each representing a genre. The genres included are action, adventure, animation, biography, comedy, crime, drama, fantasy, history, horror, music, mystery, romance, sci-fi, sport, thriller, and war. Each genre CSV includes one hundred movies pertaining to the genre, each movie has a star rating, age rating, and a link to an HTML page with a list of the reviews the movie has. In total, there are 1700 different genre/movie pairs, for 1150 unique movies. During our analysis we ran into issues with multiple genre CSVs. After working around these issues we are now working with only 12 genres, 1200 different genre/movie pairs, and 960 unique movies. The genres we are working with include action, adventure, animation, biography, comedy, drama, fantasy, horror, music, mystery, thriller and war. The limits of this dataset include our removal of five genres, and the amount of reviews for each movie is limited. Each movie includes somewhere between seven and twenty five reviews, with the majority having over twenty. If we had more time, we would have liked to have been able to work with both datasets given from the IEEE web page, as there would have been much more reviews to process, as well as metadata like reviews per movie, run time, and number of ratings.

ID	Movie	Genre	Age Rating	Star Rating	Review URL	Reviews
1	"The Dark Knight	Action	PG-13	9	https://www.imdb.com/title/tt0468569/reviews/_ajax?ref_= ...	['Confidently directed, dark, brooding, and packed with ...
2	"Inception	Action	PG-13	8.8	https://www.imdb.com/title/tt1375666/reviews/_ajax?ref_= ...	['My 3rd time watching this movie! Yet, it still stunned ...
3	"The Matrix	Action	R	8.7	https://www.imdb.com/title/tt0133093/reviews/_ajax?ref_= ...	['When this came out, I was living with a roommate. He w ...
4	"The Lord of the Rings: The Fellowship of the Ring	Action	PG-13	8.8	https://www.imdb.com/title/tt0120737/reviews/_ajax?ref_= ...	['Director Peter Jackson and wife Fran Walsh have succes ...
5	"The Dark Knight Rises	Action	PG-13	8.4	https://www.imdb.com/title/tt1345836/reviews/_ajax?ref_= ...	["Christopher Nolan's epic trilogy concludes in glorious ...

After tedious data manipulation, the image above shows how our data structure is set up to begin our analysis. To structure and display the data this way, we used the Table() method from the datascience package.

Methodology:

Now that our data is structured in a way we can perform analysis on it, we begin our methodology. Our first step was to tokenize the entirety of all the reviews in our dataset. After we tokenize, we create counters that are genre specific, and movie specific, as well as a total counter. Now it is time for us to begin our feature analysis. We chose to start with feature analysis because it was the most recent topic in class. Our first step was to look at the most occurring and the least occurring words for a genre. We decided to use Horror and Comedy as our genres of choice because they are two genres we know are very different. The least and most occurring words don't really tell us much, so we move on to find features that are distinct and characteristic. First, we import the contingency table and the dunning gscore functions we've seen in class. We also bring over the most distinctive function from class. The most distinctive function will be able to tell us the top words in the horror and comedy genres that have a high dunning gscore. What this dunning gscore tells us is what proportion of that word is found in the horror genre and the comedy genre. Words with a higher score are strongly associated with one genre, and not the other. Words with a low score are likely to be found in either genre's reviews. We want to see whether or not the features extracted from the reviews for the comedy genre and the horror genre actually make sense to their corresponding genre. This ties back to our core question because if the extracted features for their respective genres make sense, then we can use those features as a predictor for the genre of the movie. Our results from this were a strong support for our core argument. The features extracted from comedy strongly represent that genre, and the features extracted from horror strongly represent that genre. (i.e 'horror', 'scary' are associated with horror and 'funny', 'comedy' are associated with comedy).

Our next step was to try classification. We took many functions from class notebooks, including build_vectors and classify_genre. We built the vectors for each genre and then performed multinomial and binary classification on these vectors. This would help us answer our core question by classifying all the movies by their correct genre. Our vectors represent the counts of words for each genre, and we did classification based off of these results. Our multinomial classification results were inconclusive, we could not get accurate results to display. So instead we used binary classification which might be a better choice for our case anyways. The results we got from the binary classification seem much more accurate. We know that we are able to predict the genre of the movie a review comes from with an average of ~75%. This is a relatively good accuracy for just predicting based on counts of words for each review.
We decided to try another small classification technique we saw in class to see if we could show anything else. We found the word probabilities for all the genre names (i.e horror, comedy, animation) occurring in said genre, and all the genre names occurring in the other genres. The resulting table didn't give us much information but you can clearly

see the name of the genre, which occurs much higher in that genre than any other. For example the word comedy appears in reviews for comedy movies more than horror movies. It is fair to assume these results would be this way anyways, but it gave us clarity that we are working in the right direction.

The next step we took was to try and cluster our movies by genre, based on the content in the reviews. We once again tokenized our entire collection. We shortened our vocabulary down to words that occur in more than one genre. Then we build vectors from our counts of each word as we did in the previous method. We use the print_clusters method from our class notebook to display each movie, its genre, and the cluster it was placed in. To do this we also imported AgglomerativeClustering from the python package sklearn. The idea here was to see multiple movies of the same genre in the same cluster. We can then display this visually using dendrogram from the scipy.cluster.hierarchy package. Being able to view the clusters visually can help us determine if our clustering method is working correctly. If it is working correctly, it will be a critical part in answering our core question.
Our results here are insufficient. Multiple problems were run into trying to translate the given code into something that would work for our dataset. Our clustering results showed the first 1189 movies in the first cluster, and one movie per cluster for the last eleven clusters. We were unable to solve this problem in time. If we had more time we would have gone back to try and step by step recreate the code from the notebook so that it worked for our data.

Last but not least, comparison. We chose to do comparison last because we thought the results would be the least significant for our analysis. Our related works used Jaccard similarity to compare two genres together. We plan on using cosine similarity to compare the reviews for each genre. If the reviews of two genres are similar, we might conclude the two genres are also related. If the reviews of two genres are dissimilar, we might be able to conclude the genres are not so related. For example, as we've previously discussed the horror genre and the comedy genre have very different features. By looking at the probability table from the classification method, we can assume action and thriller, and mystery and horror are two pairs of closely related genres. They have the largest probability for two different genres in that probability table. Finding the jaccard similarity between like and dislike genres can help us work towards our core question by establishing which genres are closely related.

Analysis:

Comparison: As we were going about our plan to use the cosine similarity for analysis, there was an error in the code that would not let us execute our code to the fullest

extent. If there was no error, we attempted to find the similarities between the different genres.

Clustering: In order to cluster our movies together we attempted to use a dendogram to visualize all the movies, along with the clusters they belong to. Our preliminary results displayed our clusters in a very unrealistic way. The first cluster contained all but 11 movies, whereas the remaining movies were split into individual clusters of their own. This result did not help us in supporting our dataset due to the incomplete results that were provided to us.

Classification: As we were classifying our dataset, we decided to look at the probabilities of how often all the genre names appeared within all the other genres. One thing to realize when looking at this table is that the column names are the genres that we are investigating, and the row titles/names are simply words that also match the name of the genre.

GENRE	action	adventure	animation	biography	comedy	drama	fantasy	horror	music	mystery	thriller	war
action	0.00303204	0.000258098	0.000101309	1.20606e-06	0.000253273	0.000209855	0.00011337	9.16608e-05	0.000334079	7.47759e-05	0.000118194	0.000560819
adventure	0.00200134	0.000365966	0.000436019	0	0.000227068	0.000189626	0.000170301	9.5417e-05	0.000444474	7.0053e-05	4.71046e-05	0.00045776
animation	0.000765857	0.000421085	0.00276908	0	0.000470144	0.000136274	0.000166254	6.26858e-05	0.00065275	5.85976e-05	1.36274e-06	0.000208499
biography	0.000454427	5.62055e-05	1.19586e-05	0.000114803	0.000259502	0.00047117	2.98965e-05	0.000101648	0.000756981	3.22883e-05	9.80607e-05	0.000758176
comedy	0.000672284	0.000192081	0.000597122	2.78379e-06	0.00180111	0.000231054	0.000128054	0.000121095	0.000432879	7.09865e-05	3.75811e-05	0.000143365
drama	0.000710796	8.04015e-05	4.54443e-05	1.28176e-05	0.000212074	0.000341415	0.000174786	0.000221395	0.000510375	0.000100211	0.000185273	0.000720118
fantasy	0.00128621	0.000389831	0.000182077	2.33432e-06	0.000463362	0.000176241	0.000452857	0.000285954	0.000407338	7.93667e-05	4.4352e-05	0.000260276
horror	0.000958811	5.61851e-05	1.22142e-05	2.44283e-06	0.00037131	0.000149013	9.03848e-05	0.00393907	0.000398182	0.000211305	0.000328561	0.000136799
music	0.000142606	4.70868e-05	5.38135e-06	6.05402e-05	0.000731864	0.000313464	6.99575e-05	8.07202e-05	0.00322343	6.18855e-05	4.70868e-05	0.000180275
mystery	0.000964201	8.22096e-05	5.87211e-06	1.17442e-06	0.000139756	0.000218443	7.51631e-05	0.00139756	0.000347629	0.000571944	0.000634188	0.000250152
thriller	0.00313296	0.00010114	8.32917e-06	4.75952e-06	0.000142786	0.000267723	9.28107e-05	0.000424788	0.000312939	0.000186811	0.000553295	0.000348635
war	0.000808977	0.000123798	3.67717e-05	4.90289e-06	0.00039836	0.000417972	0.000100509	0.000199793	0.000357911	4.78032e-05	5.7609e-05	0.00479871

By looking at this table, we noticed that the relationship between the word thriller, and the action genre are the most integrated with one another.

In addition to the table of probabilities we also ran a classify_genre method to run a binary classification of a genre in comparison to all the other genres.

```
action - Accuracy: 50.0%
adventure - Accuracy: 58.3%
animation - Accuracy: 91.7%
biography - Accuracy: 75.0%
comedy - Accuracy: 75.0%
drama - Accuracy: 75.0%
fantasy - Accuracy: 58.3%
horror - Accuracy: 83.3%
music - Accuracy: 83.3%
mystery - Accuracy: 75.0%
sport - Accuracy: 100.0%
thriller - Accuracy: 83.3%
war - Accuracy: 91.7%
```

Our result does not demonstrate the result we expected due to the unnatural accuracy probabilities that are portrayed in the result. The fact that there are many probabilities that are either ½ or ¾ makes it seem like this result could be inaccurate, but when we noticed that the accuracy of the sport genre is 100%, this solidified our belief that the result shown is most likely inaccurate.

Feature Analysis: While looking through our dataset we focused specifically analyzing the comedy and horror genres. Comedy and horror are two very different genres that still overlap through very different movie concepts. First we looked at the most and least occurring words from each review for both genres. We look at this first because we want to obtain a small understanding of how each genre is described by the viewers throughout the different genres.

```
Most occuring words for the Comedy genre:
1. the: 41996
2. and: 23614
3. of: 18137
4. to: 17250
5. is: 14681
_____________________________________________
Least occuring words for the Comedy genre:
1. leon: 1
2. marsan: 1
3. leith: 1
4. dequel: 1
5. womans: 1
```

```
Most occuring words for the Horror genre:
1. the: 52292
2. and: 25223
3. of: 21237
4. to: 19123
5. br: 16598
_____________________________________________
Least occuring words for the Horror genre:
1. naiive: 1
2. reflexion: 1
3. denounces: 1
4. insensible: 1
5. woes: 1
```

By looking at the most occurring words for each genre exclusively we can see that most of the top occurring words for both genres are connecting words such as prepositions, and conjunctions, with the exception of the word “the” which is an article. As for the least occurring words between the two genres exclusively, we can see that the words

seem to occasionally be misspelled words, such as “dequel” or “naiive” from the Comedy and Horror genres respectively. Other than that we can see that there is a foreign word, “reflexion” that is seen in the least occurring words for the Horror genre, which could indicate that there are more English reviews written for these genres in relation to the amount of reviews written in a foreign language.

Case Study:

As we began our research focusing on the comedy and horror genres as our test genres, we found that our results seemed to produce interesting results. Due to this, we did a further analysis of the Comedy and Horror genres inclusively by looking at the most/least distinctive words of both genres combined, and then we also looked at the ranking of the most significant words between the genres and the count of the words within each genre.

```
Most distinctive words in reviews of the horror genre and comedy genre:
1. horror: 3396.178495941959
2. funny: 816.828283106615
3. pixar: 804.9292057879219
4. comedy: 797.1737990663137
5. scary: 697.1194736250409

_____________________________________________
Least distinctive words in reviews of the horror genre and comedy genre:
1. never: 2.525766973349164e-08
2. took: 2.6295282111732377e-06
3. meticulous: 3.0911291915458605e-05
4. chronicles: 3.0911291915458605e-05
5. understatement: 3.0911291915458605e-05
```

When investigating the most distinctive words between the two genres, one thing we can see is that the name of the genres themselves are within the top five most distinctive words. Additionally we can see that there is an animation studio, “pixar” within the distinctive words, and we can infer that this studio is there because they are probably very related to creating one of the two genres, but from our personal knowledge of movies we can assume that “pixar” is more likely to be related to the comedy movies. We can also see that the top adjectives used to describe the Comedy and Horror genres when looking at them inclusively are funny and scary. When looking at the least distinctive words between the two genres most of what we can see is not too surprising, we can see that the word “never” is the least distinctive between both genres. Also, one factor that was surprising is that of all the words from both most/least distinctive words, understatement is the only word that is a literary device.

```
Rank Score| Term: Horror Comedy
_____________________________
1. 3396.2 | horror: 3225 87
2. 816.8 | funny: 282 1270
3. 797.2 | comedy: 304 1294
4. 697.1 | scary: 783 47
5. 587.0 | vampire: 484 2
6. 574.0 | disney: 11 436
7. 569.9 | animation: 10 429
8. 541.8 | evil: 856 117
9. 520.9 | alien: 749 86
10. 487.7 | animated: 4 346
11. 482.1 | toy: 19 403
12. 478.1 | gore: 582 45
13. 417.4 | carrey: 4 299
14. 384.5 | blood: 484 41
15. 383.3 | scares: 370 11
16. 373.1 | woody: 1 253
17. 360.6 | marty: 4 261
18. 360.0 | hilarious: 112 536
19. 353.9 | creepy: 416 29
20. 308.1 | jokes: 65 395
21. 299.7 | terror: 254 2
22. 293.9 | dead: 843 232
23. 293.4 | murray: 2 206
24. 286.2 | his: 3180 4145
25. 280.9 | comedies: 21 266
```

When looking at the top 25 most significant words between the two genres, we can see that “horror” is the name of a genre that we are investigating, along with being the most significant word between the two genres Comedy and Horror. We know that four of the top five most significant words were also in the analysis of the most distinct words between the two genres, but surprisingly the outlying word from the most distinct words that was not significant is “pixar”. This is surprising to us because we believed that “pixar" would at least be within the top 25 significant words, especially considering they are an animation studio, and the words “animation” and “animated” are within the top 10 significant words. An interesting find in regards to the absence of the word “pixar”, we see that “disney” shows up and replaces “pixar” as the only animation studio on the list of significant words. When we change our perspective from looking at the words themselves and transition to looking at the frequency of the words in each genre, something we found that was interesting is that the word “vampire” only appears two times in comedies, and we found this surprising due to all the comedic animated movies there are about mythical monsters, such as Hotel Transylvania, or Scooby-Doo. One other thing that caught our attention is how “horror” holds the place as the most significant word between the two genres yet it only shows up in Comedies 87 times. We can also see that the Comedy genre is more of an open and free genre, rather than how

the Horror genre will mainly be restricted to the genres that fall in a similar category to itself due to how “comedy” appears in horror about 304 times, in comparison to how many times “horror” appears in Comedy.

Conclusion:

We have concluded that there definitely exists a correlation between genres along with the reviews for those genres. One thing to note is that we believe there is a correlation between all the genres but there exists a higher correlation between genres of a similar type. It is similar to pairing, horror and thriller genres together, or matching comedy and animation and romance together. A direction that we can aim for in the future is to be able to predict the genre using the movie reviews, although this would incorporate concepts that we have not fully covered in class. Another thing we are curious about is the most/least distinctive, and the most/least occuring words when looking at an inclusive dataset of all the possible genres, instead of confining it to two. We learnt a great deal of material throughout the process of this class and this project, but one skill we feel accomplished in is the ability to work with text and analyze text, especially with the usage of a Counter, and tokens. Additionally, we wanted to figure out how to manipulate text to make predictions about other features. We saw this in articles we read, so we were curious as to how we would go about doing this. If we had unlimited time and resources we would definitely attempt to gain another collaborator to this project, but aside from that we would look more into NLP(Natural Language Processing) as it seems to be a big part of creating predictions off of given features, along with attempting to better our understanding of text analysis.

References:

[1] : https://ieee-dataport.org/open-access/imdb-movie-reviews-dataset

[2] : Hoang, Quan. (2018). Predicting Movie Genres Based on Plot Summaries. https://www.researchgate.net/publication/322517980_Predicting_Movie_Genres_Based_on_Plot_Summaries

[3] : Dey, / Sandipan. “Data Science with Python: Exploratory Analysis with Movie-Ratings and Fraud Detection with Credit-Card Transactions.” *Sandipanweb*, 2 July 2018, https://sandipanweb.wordpress.com/2017/12/16/data-science-with-python-exploratory-analysis-with-movie-ratings-and-fraud-detection-with-credit-card-transactions/.