– Independent Work Report Fall 2019 –

Analyzing Objectivity in Film Maturity Ratings

Natalie O'Leary
Advisor: Christiane Fellbaum

January 7th, 2019

Abstract

The maturity ratings of films given by the Motion Picture Association of America have long remained mysterious to the public. This paper seeks to demystify these ratings somewhat by analyzing the ability of various classification models to predict the ratings given to a test set of films. Because the ratings are determined by a group of individuals with a vested interest in the outcome, the prediction is that models will not be particularly successful at determining ratings from the scripts alone. This prediction is supported by the results of this project.

1. Introduction

As any modern movie goer knows, maturity ratings are attached to any and all content that is screened in theaters, including previews, short films, and feature films. These ratings, ranging from PG to R, determine who is and is not allowed into a given screening and also advise parents on what movies to bring their kids to. However, these ratings are determined by a committee of people known as the Motion Picture Association of America, the MPAA for short, who – as human beings – are incapable of being completely objective. Over the years, the MPAA's ratings and standards have been persistently unpredictable and largely inconsistent, with one notorious example of unpredictability being the violent, shark thriller Jaws, which was rated PG at the time of its release. The MPAA is made up of representatives from Paramount Pictures, Sony Pictures, 20th Century Fox, Universal Pictures, Walt Disney Studios, Warner Bros and Netflix, which are the 7 largest studios in America. As a result, these studios have a distinct advantage over smaller, independent studios in attaining desired maturity ratings, as they are among those who get to make those decisions. Furthermore, the Disney Company has an even bigger advantage, in that

they own both Walt Disney Studios and 20th Century Fox. The MPAA has also never published any guidelines regarding how they determine their ratings (e.g. swear quotas or length of nudity scenes). This lack of transparency allows for potential corruption, and allows these larger studios to use their resources to push the MPAA toward one rating or another, depending on a movie's desired demographic.

Additionally, several studies [8][1] have shown that ratings have tended to "creep" over time, meaning that today's films tend to contain more sex and violence than those that received the same ratings in previous years and decades. This is cause for great concern for parents and children, as younger and younger children are being exposed to mature material. Studies such as one done at Çukurova University in 2017 [4] point to evidence that violent films and media content can negatively affect child and adolescent behavior through observational learning. This means that the corrupt and ineffective nature of the MPAA has a direct effect on the well-being of children, as well the general disposition and affinity for violence of the general population. Thus, the maturity ratings granted to films have a direct and dire impact on the public, and should accordingly be determined in an objective manner that accurately informs viewers and prevents impressionable children from seeing inappropriate material.

So, if one were to pursue a completely objective rating system, ideally only the source materials and contents of the film should be considered, without any human emotion or influence involved in the decision. Thus, it follows that a machine, if given the correct information, should be able to determine an accurate maturity rating more objectively than a committee of financially invested humans. The complete scripts for feature films are a natural data set to turn to for this problem, as they contain all the dialog in the films as well as descriptions of the non-verbal scenes. From this information the frequency and weight of different words and features can be determined in relation to other scripts, including words that are spoken and words that pertain to visual information, such as nudity or violence.

This project seeks to quantitatively analyze the objectivity of the MPAA's ratings of around 300 movies over the past 4 decades, by training different classifier models on their scripts and

comparing the models' predicted maturity ratings on the test set of scripts to the actual ratings for those films. I have chosen to make these predictions using several different models, as my goal is to show that none of the models are able to accurately predict the ratings, giving evidence that factors outside of the contents of the films play into the rating process. The use of multiple models minimizes the chance that any one classifier's poor performance could be due to a poorly chosen model, as well as showing that even the model with the best results will still not be able to accurately produce the correct ratings the majority of the time.

2. Problem Background and Related Work

This section aims to delve into some related research projects in this field of study, however, the range of related projects is rather broad, as the field of film classification is surprisingly unexplored. I was unable to find any studies that directly deal with classification based on maturity ratings, however there are some related movie classifiers which deal with genre and preference.

Although it is rooted in genre classification, one project, entitled, "Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features" [2] by Alex Blackstock and Max Spitz was particularly useful to me in my research. The goal of their work was to train both a Maximum Entropy Markov Model and a Naive Bayes classifier on movie scripts, and to use them to predict movie genres and compare the results. To make up for the lack of visual information in scripts, they separated the words in the scripts into different types of "frames", either stage directions or dialog. The stage direction frames were identified by their setting and the dialog frames by their speaker. They used these frames to quantify the ratio of dialog driven scenes vs visually driven scenes, as some genres may be more visual, while others rely heavily on dialog. Because the end goal was to determine genre, they used many different features to infer nuances of the film, including ratios of descriptive language, personal pronouns, wh questions, locations, – historical or otherwise – and exclamations within the text. They ultimately found that the MEMM model had a marginally better success rate, correctly generating genre about 55% of the

time. Many of these factors, like location and number of speakers, while they may be pertinent in differentiating between a historical drama and a documentary with many speakers, are not extremely relevant to determining maturity rating, which is less dependent on motifs and more driven by the language itself. Nevertheless, their use of, and processing of scripts was very helpful in my research. The drawback to working with genre classification is that it is not so cleanly black and white like maturity ratings are. Many films may fall into more than one genre, like crime thrillers or romantic dramas. In fact as many as half of the scripts they used were labeled with at least one of Drama, Thriller, Action, Comedy or Crime, meaning even a completely random generator would be right about 50% of the time if it drew from only these labels [2]. This constant overlap and blending makes genre classification a slightly less feasible problem to tackle using NLP, however their work was instrumental in informing my efforts.

Another study, "Smart Trailer: Automatic Generation of Movie Trailer Using Only Subtitles," [4] conducted at Misr International University goes a bit further than classification, as they first classify the films by genre and then use that information to select the most important sections of the subtitles, which can be used to generate a trailer for the film. This idea came to be as a result of the massive increase in video data as users began to be able to create, publish and post their own videos. The researchers built their model by parsing through the subtitles and separating them into distinct n-grams with linguistic significance. These key phrases are then used to determine genre using a K-nearest neighbor classifier. Once a genre has been determined and the various key phrases separated out, the phrases are ranked using the PageRank algorithm [3] and the top ranked key phrases that fit within the specified time frame for the trailer are used to select the scenes for the trailer, which they hoped would be an adequate summary of the film. The classification portion of this study is the most relevant to my project, so I will be focusing on those results. Their genre classifier, in contrast to the Blackstock and Spitz model [2] returned as high as 89% accuracy predicting the genre of Action films, while it had only 55% accuracsy determining the genre of Fantasy films, with the rest of the genres falling somewhere between these values. This suggests that some genres have a more clear cut dictionary of words

and phrases, while others are more varied. Logically, this should be true of maturity ratings as well, as R rated movies tend to contain a lot of violence, profanity and sex, while G rated movies can have many different subjects, from children's films to documentaries.

The final important and relevant work to this project is one done at USC entitled "Recommendations Without User Preferences: A Natural Language Processing Approach" [6] which uses NLP to generate similar movies given a singular film. This project is a little different from the others as it uses plot summaries of the films for a corpus, rather than the actual scripts or dialog from the movies themselves. However, the object of this project was to remove humans from the movie recommending process, much like the aim of my project is to remove humans from the rating process. They used two algorithms to achieve this result, the first being a word-space similarity metric that transforms each summary into a vector of binary features similar to the one I implemented in this project [6]. The second algorithm they implemented attempts to give each movie a score within each genre, reflecting how closely it matches that genre. Films with similar genre scores are determined to be similar in subject matter and theme. They had each of their algorithms generate 5 films determined to be most similar to the given film, and gave these recommendations to test subjects to compare to human generated recommendations to determine which are closest to the original film. They ultimately found that the genre based algorithm performed comparably to the recommendation software implemented by the popular movie site, IMDB. The goal of this project approaches but does not quite touch on the goal of my project, which is to assess human bias in the movie rating and recommending industry. Personal preferences are not universal across all human beings, just as opinions about what is and is not appropriate for children and adolescents to see may vary among parents. Both this study and my own seek to quantify these somewhat ambiguous qualities, so that they may better serve the public. Whether or not you agree with the classifications, the public seeks them out, both ratings and recommendations, and so it is to our benefit that they be universally, objectively determined.

3. Approach

3.1. Data Set

3.1.1. Ratings There are a variety of different ratings that have been used by the MPAA throughout its years of operation, however, today some are used far more frequently than others. The rating of PG-13 was only introduced in 1984, but it is now one of the most commonly used ratings, while the G rating has become largely obsolete. As it currently stands, the MPAA ratings are defined as in Table 1 below.

Rating	Explanation	Who's Allowed In
G	Nothing that would offend parents for viewing by children	All Ages Admitted (not restrictive)
PG	Parents urged to give "parental guidance." May contain some material parents might not like for their young children	Some material may not be suitable for children (not restrictive)
PG-13	Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers	Some Material may be inappropriate for children under 13 (not restrictive)
R	Contains some adult material. Parents are urged to learn more about the film before taking their young children with them	Under 17 requires accompanying parent or guardian (restrictive)
NC-17	Clearly adult. Children are not admitted	No One Under 17 Admitted (restrictive)

Table 1: MPAA Ratings in 2019. [9]

As I have already established, the MPAA is made up of representatives from giant movie corporations with a vested interest in how the ratings affect movie sales. This has caused the number of movies rated G and NC-17 to be drastically lower than the other three ratings. This is because the rating PG has become synonymous with G in terms of content, however it saves the movie company from being held accountable should any parents complain that a PG film contains lewd content since it is left to the parents discretion. Additionally, an R rating is a far preferable rating to NC-17 for studios, as kids under 17 make up 24% of movie goers,[7] and R allows them

into the theater with parental supervision while NC-17 completely cuts off potential revenue from that demographic. I therefore chose to remove those two ratings from my study, and the only labels used in my classifier are PG, PG-13 and R. I attained the ratings data from an online database on data.world called "IMDB Top 250 Lists and 5000 Plus IMDB Records," [11] which contains the IMDB stats on over 5000 movies including ratings and titles.

3.1.2. Corpus There are many different ways to process movie data, as films are an audio and visual experience generated from textual scripts full of lines and directions. While music and lighting may be important in genre classification, they are not quite as pertinent to maturity ratings, so I chose to use only raw script data for my corpus. The most prevalent online movie database for textual data is the Cornell Movie Dialog Corpus, [5] which is a great resource for film dialog and conversation. However this was not sufficient for my work, as this corpus does not include stage directions, which are just as important in determining the ratings as the dialog itself. While profanity plays a big part in what we know of the rating system, things like nudity and violence are mostly portrayed visually, rather than aurally, and can therefore only be described in the stage directions of a text document. I was able to attain about 300 viable individual scripts with an equal number of each of the three ratings from Film Corpus 2.0, compiled at UC Santa Cruz [12]. I found it unnecessary to pre-process the scripts as thoroughly as Blackstock and Spitz [2] who split the scripts into dialog and stage direction or grouping the dialog into discinct scenes like the Smart Trailer project, [4]because ratings are more dependent on overall volume of inappropriate material, rather than the ratio of scene to dialog or importance of the racy scene.

3.2. Classification

3.2.1. Feature Extraction and Weighting I used tools from scikit-learn [10]primarily for this segment of the process. Since it is unclear exactly what kinds of metrics the MPAA uses to determine rating, I have to assume that all words in the script are important, thus I tokenized all of the scripts using scikit-learn's CountVectorizer, which transforms the whole text into an array

of tokens and then creates a vector that represents the frequency of each token in a given text in relation to all of the other text samples. I also utilized scikit-learn's TfidfTransformer tool, which uses a count vector to produce a tf-idf vector, a kind of vector that assesses the importance of each token (word) in a document within the context of the entire corpus. These tools would be the key to identifying the most significant differences between the differently rated films in training, and thus the basis on which all of my models produced classifications.

3.2.2. Evaluation Metrics There are two main methods through which I determined the success of my models: Precision-Recall Curves and ROC-AUC curves, both of which depend on the idea of false positives, false negatives, true positives and true negatives. For each label a positive value is any instance in which the classifier identifies a film as that label, and negative is every time the classifier outputs any other label. True means that the label was correct and false means it was incorrect. Thus, for any given label, a true positive is when the classifier correctly outputs that label, while a false negative is when it incorrectly chooses a label other than the one in question as shown in Figure 1. Mathematically, precision is defined in equation 1 and recall in

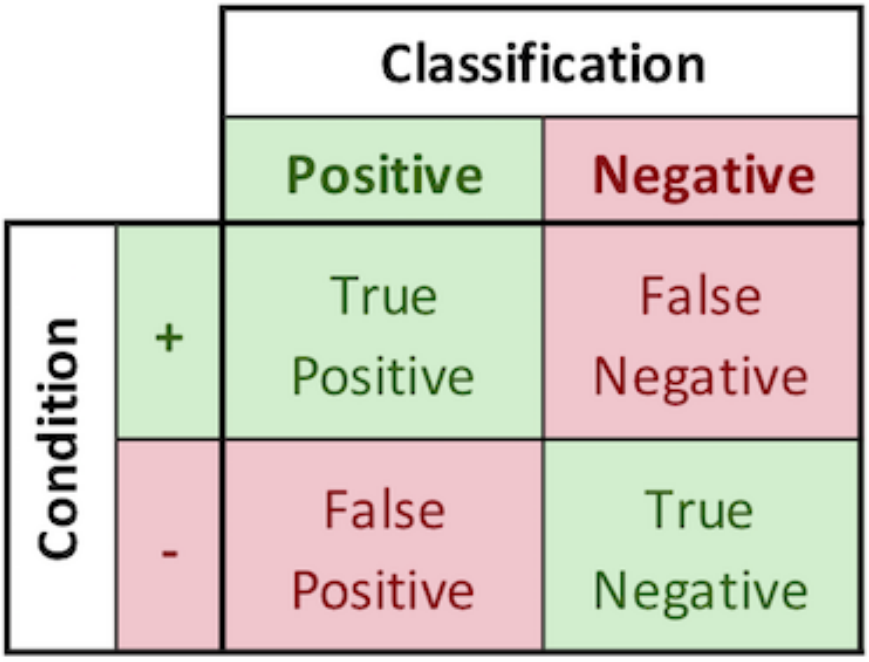

Figure 1: Confusion Matrix

equation 2. Essentially precision measures correctly identified labels taking into account how often this label was incorrectly given, while recall measures correctly identified labels taking into account how often other labels were applied to something of this type. These are significantly more enlightening statistics than something like a percent accuracy, especially when using a multi-class classifier, as it could appear that the classifier has 100% accuracy in identifying R rated films, however if it labels all of the films as R, it is clearly not doing its job correctly as it has no precision. On the other hand recall is a measure of the percent of true positives out of the total number of films in the data set that had that rating. The F1 score is the harmonic mean of precision and recall calculated as in equation 3. F1 curves, or precision-recall curves plot the F1 scores for different threshold values.

$$Precision = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{1}$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{2}$$

$$F1 = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \tag{3}$$

ROC-AUC curves on the other hand, use true positive rate (which is the same as Recall) and false positive rate, which can be calculated as in equation 4, or by simply subtracting the true positive rate from 1, as they are inverses. The ROC-AUC curve plots true positive rate against false positive rate, thus comparing how often a model correctly gets a label compared to how often it incorrectly selects that label. A sample multi-class ROC-AUC curve from scikit-learn [10] is shown in figure 2. The greater the area under the curve, the more successful the model.

$$FalsePositiveRate = \frac{\text{False Positives}}{\text{False Positives + True Negatives}} \tag{4}$$

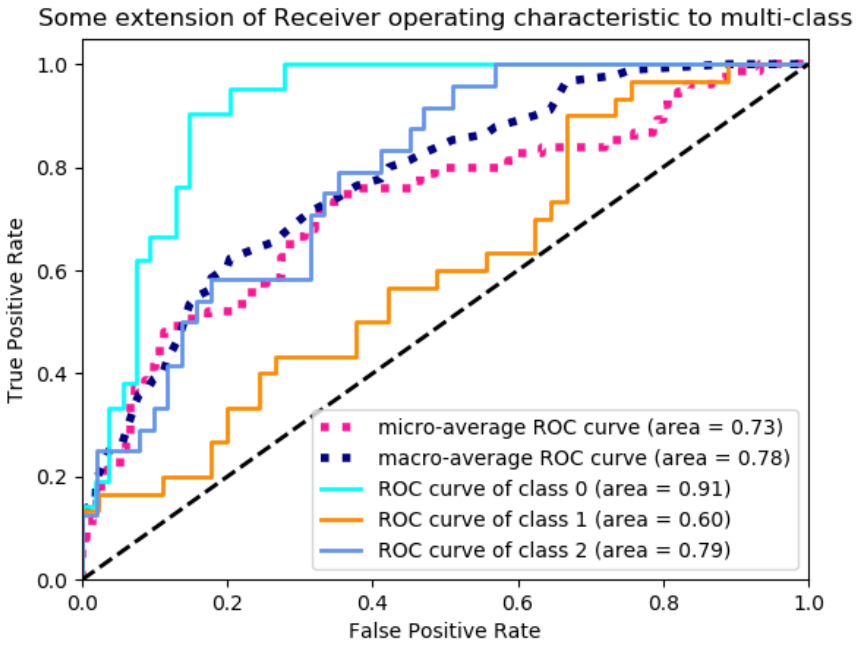

Figure 2: Sample Multi-Class ROC-AUC Curve [10]

4. Implementation

4.1. Gathering Data

A good portion of my time was spent collecting and preparing the necessary data for testing. The trouble with collecting scripts is that there is no universally adapted format for them, so while some scripts contained the title and MPAA rating at the top of the document, others placed that data elsewhere or didn't include it at all. As I mentioned in section 3.1, I was able to find the majority of my scripts from Film Corpus 2.0 [12], which stores 1070 different scripts in text files where the name of each file is the title of the film. I then had to write a program that took these titles and searched for them in the ratings database [9] to actually match the scripts to their correct MPAA rating. After this process was complete, I was left with 648 scripts with ratings. However, the ratio of PG to PG-13 to R was by no means equal in this data set, which was about 60% rated R 32% rated PG-13 and only 8% rated PG. This was problematic as I only had about 66 films rated PG, with the majority being rated R, which could skew the classifiers toward R ratings disproportionately often. So, I had to supplement my data by manually searching for PG rated

scripts online and either locating their ratings in my database or finding their ratings online as well. I ended up manually inputting 25 scripts, for a total of 81 PG films. I then randomly chose 101 of the 441 R rated scripts and 100 of the 207 PG-13 rated scripts for a total of 283 feasible, rated scripts that were ready to be pre-processed.

4.2. Prepossessing

Although I used a few different models, each of them requires the data to be prepared via the same key steps which I outline below.

4.2.1. Count Vectors In order for text data to be properly interpreted by each of the models, the words need to be transformed first into an array of singular tokens rather than strings of text. This was done using the Count_Vectorizer tool in scikit-learn, [10] which first tokenizes the text, and then uses these tokens to create count vectors for the whole data set. A count vector contains a value for all words used in the vocabulary of the entire corpus depending on the frequency of that word in a given script. So each script is now represented as a singular count vector containing its token frequencies.

4.2.2. TF-IDF Vectors Count vectors are not the final transformation of the corpus however, as we must now find a way to relate each of these count vectors to each other, so they actually mean something relationally, and can be analyzed by the classifiers. For this, we will calculate tf-idf vectors for the corpus, which are a way of weighting certain terms in terms of importance within the entirety of the corpus, rather than relying on raw frequencies alone. For example, words like 'and' and 'the' will likely have high frequencies in each of the count vectors, but are unimportant in differentiating between scripts and will have a low weight in the tf-idf vectors. Tf-idf stands for term frequency inverse document frequency, and equation 7 shows the tf-idf calculation for a word w, in document d within the corpus D.

$$tf(w,d) = f_d(w) \text{ frequency of word, w in document, d} \tag{5}$$

$$idf(w,D) = log\frac{1+|D|}{1+df(d,w)}$$ df is number of documents in the corpus the word appears in (6)

$$tf-idf(w,d,D) = tf(w,d) * idf(w,D) \quad (7)$$

These calculations were computed using scikit-learn's TfidfTransformer tool [10]. After this transformation, the corpus is completely transformed from a series of long documents to parse through to a comprehensive series of values that can be used to easily classify documents based on frequencies of differentiating terms.

4.2.3. Model Fitting The final step in prepossessing is to fit the training data to each of the models I have chosen. This was done by first splitting the data in to 80% training data, so as to maximize input since I only have about 100 of each rating. The remaining 20% of the scripts are reserved for testing and analysis. I did this splitting using scikit-learn's train_test_split tool [10], which randomly splits the data into training and testing vectors.

4.3. Model Selection

The goal of this study is to assess how objective the MPAA ratings are by demonstrating how a machine cannot accurately predict the ratings despite having ample information, because human bias is unpredictable. Thus, it is necessary to use a few different models to limit the possibility that the results are caused by one model that was simply ill-suited to handle the problem. I chose four different models that approach the problem in different ways: Naive Bayes, Linear Regression, and a One VS Rest classifier from scikit-learn.

4.3.1. Naive Bayes The first model I chose was Naive Bayes, which I considered as a kind of baseline, since it is among the simplest of the classification models, and it assumes all features are independent, hence the term 'Naive'. Naive Bayes is a generative classifier, so it takes into account all of the features present in a given document and uses that data to determine which class has the highest probability of being applicable to this document given the frequency of present features. The specific variant of Naive Bayes that I chose to use was Gaussian Naive Bayes, which assumes the features follow a normal distribution and calculates the posterior

probability of a given class x_i given feature y using the formula given in equation 8.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y{}^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \tag{8}$$

4.3.2. Linear Regression - Multinomial Unlike Naive Bayes, the linear regression model is discriminative rather than generative. This means that it generates classifications by eliminating classes based on features they $don't$ typically have. I chose this model as I thought this algorithm would perform better than Naive Bayes for this situation, as the ratings system functions on a gradient that eliminates certain audience members as certain features become present in the film. A G rated film will almost never have profanity, and thus those features will be a neat seperator to distinguish between a G rated film and one rated PG-13 or R. To construct this model, I used scikit-learn's LogisiticRegression tool from the linear_model library, [10] which is similar to the Maximum Entropy Markov Model used in the Blackstock and Spitz [2] study, since logistic regression is also a maximum entropy classifier. The probabilities for classification are calculated using the logistic function shown in equation 9. The solver I chose for this model is the Stochastic Average Gradient optimization, because it works particularly well with large data sets, which is helpful given that the tf-idf and count vectors for the scripts are very large and can contain as many entries are there are distinct words in all the 283 scripts combined.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{9}$$

4.3.3. One VS Rest Classifier I implemented this final model using scikit-learn's OneVSRest classifier, which functions by fitting a separate binary classifier for each class against all other classes. Each of the binary classifiers used on each individual class is implemented by a C-Support Vector Classifier. This model is the only one currently able to be graphed using scikit-learn's ROC-AUC tools and precision-recall tools, (as well as any other tools on nltk or other online resources I could find) because of this unique design which treats the problem as three

distinct binary problems. ROC-AUC and precision-recall are inherently binary problems, and are thus best implemented with binary data, and can only be expanded to multi-class data if it is binarized and interpreted as a series of separate binary problems. I did not expect this classifier to perform as well as the other two, because the binary treatment of the data oversimplifies it to the point where some information is likely lost or misconstrued. The problem with this treatment is that films that are in the middle, PG-13 films, will be compared to both R rated and PG rated films as if they are within the same, separate class from PG-13, which will likely confound the results, as these two classes are the most different among the three. However, this is the only model that allows for graphical visualization, so it was important to include in my research.

4.4. Testing

As I mentioned in section 4.2 each of the models I chose called for the same preprocessing, where I converted the text documents to count vectors and then to tf-idf vectors. I then split the corpus one time into 80% training and 20% testing, and used those same vectors and labels for each different model that I used so as to guarantee that different data did not contribute to the differences observed in the models' results. For each of these models, I intended to plot both a ROC-AUC curve and a precision-recall curve so I could visually compare the results from each of them, however I ran into an issue when I discovered that while scikit-learn has many models that support multi-class classification, the tools for ROC-AUC curves and precision recall curves only support their One VS Rest model for classification, so while I calculated precision recall scores and ROC-AUC scores for the other models, I was unable to graph them for this project.

5. Evaluation

As expected, none of the models performed particularly accurately, however the logistic regression had the highest area under the ROC-AUC curve, indicating the most successful machine. Interestingly however, the Naive Bayes classifier actually has the highest f1 score at .4962, which is comparable to the score of the logistic regression, .4754 and far greater than that of the One VS

Model	ROC-AUC-score	f1 score
Naive Bayes	0.4065	0.4962
Logistic Regression	0.5443	0.4754
One vs Rest	0.5066	0.3712

Table 2: ROC-AUC scores: area under ROC curve

Rest model which has an average f1 score of .3712. However, each of these metrics falls below the f1 score of a completely random classifier, which is .5, and therefore do not indicate a particularly strong ability for either precision or recall. The red lines in figures 3 and 4 indicate an average f1 score of .5, which is roughly the score of a completely random classifier. The ROC-AUC scores

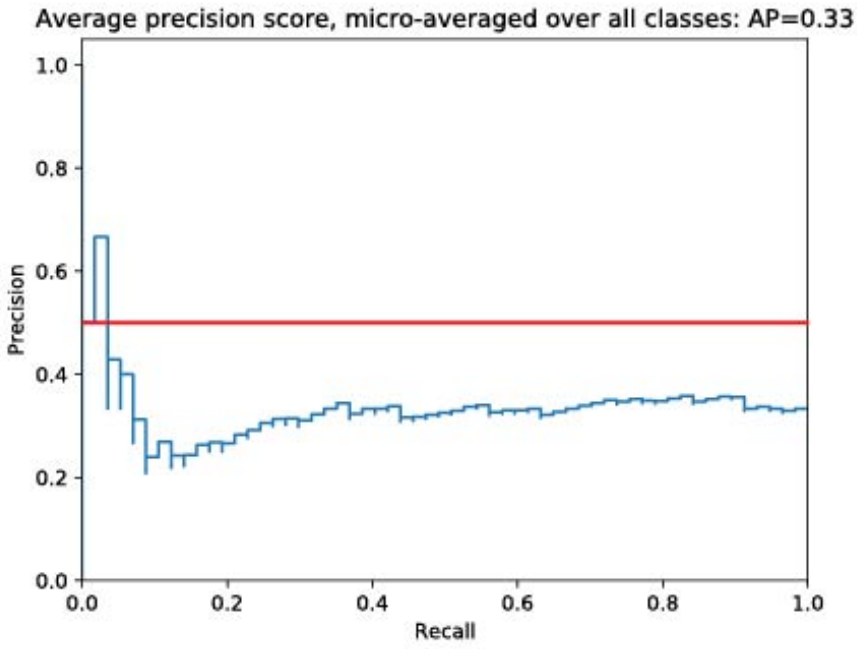

Figure 3: Average Precision Recall One VS Rest [10]

for each model are given in table 2, and although the logistic regression performed the best, it still has a ROC-AUC score of .5443, which is barely above the diagonal line in the ROC-AUC curve, which represents an equal number of true positives and false positives, which a completely random predictor could achieve. This line is illustrated in figure 5, which shows ROC-AUC curves for each individual class as predicted by the One VS Rest classifier. As we can see from this figure, the One VS Rest model is most successful at identifying the PG movies, with an individual area under the curve of .59, while its PG-13 predictions are about even with the middle line indicating

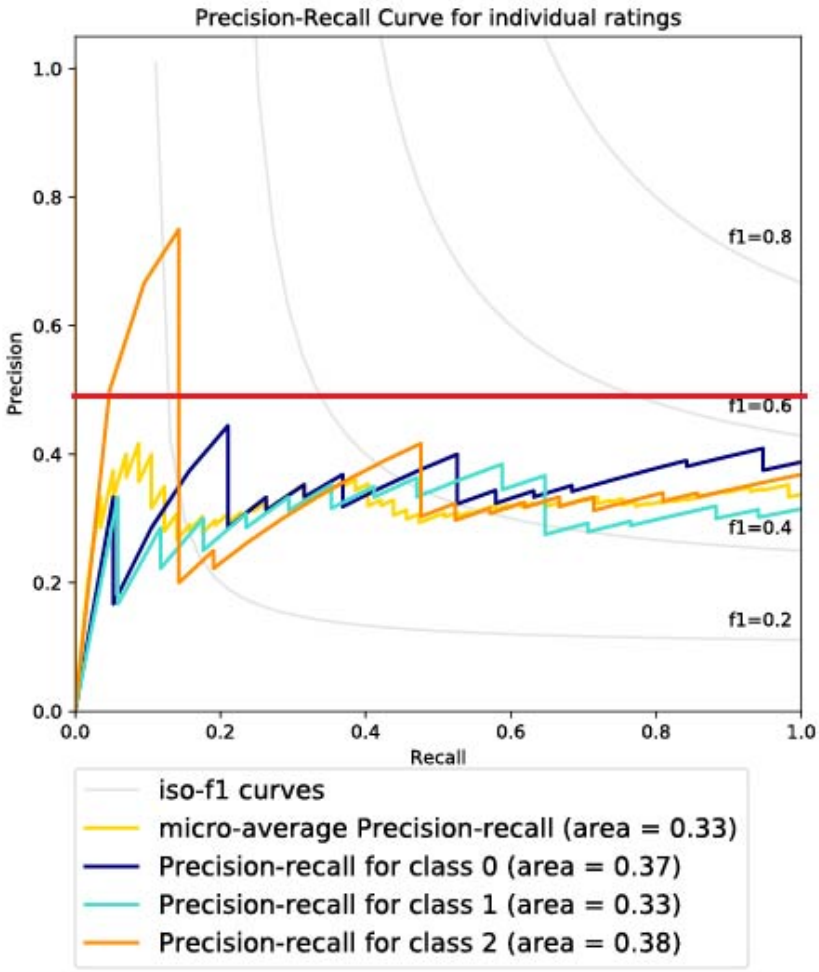

Figure 4: Precision Recall for Each Rating [10]

equal false and true positives. However the ROC curve for R rated films dips well below the middle line and indicates an area of only .41. PG films were perhaps easiest to differentiate due to their complete lack of strong language, violence and sexual themes, whereas PG-13 and R are allowed to have varying degrees and ratios of these things depending on the films, and were thus not so easy to correctly predict. Unlike the ROC-AUC curves, the precision recall curves are more similar amongst the different classes. There is a distinct spike in the precision for R rated films with a low recall, which indicates a small collection of features and indicators such as profanity that can easily distinguish R rated films from other films, however they are present in a small enough sample size that the drop off is very steep, and the curve for R rated films levels out at

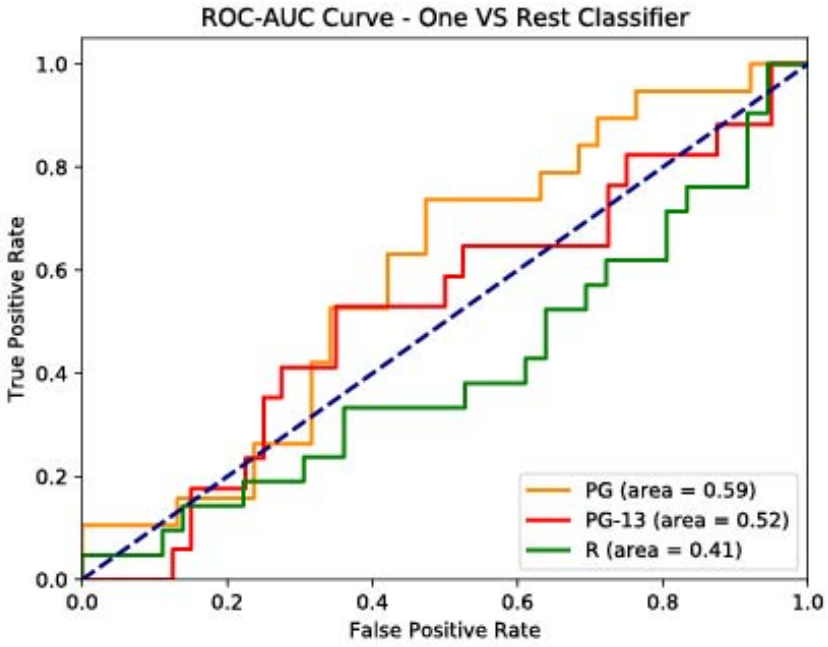

Figure 5: Multi-Class ROC-AUC Curve One VS Rest [10]

the same level of the other ratings. The average F1 curve - shown in figure 3 - is slightly higher, likely caused by this spike, however both graphs show the F1 curve falls well below the .5 score of a random classifier.

The most notable thing about the data gathered is that each of these models produced a success rate (by both metrics) of about 50% or less. Recall the results from the genre based studies discussed in section 2 of this paper; the lowest reported accuracy in any of these studies was about 50%. The models that I chose are very similar to the models used in those studies, with the Blackstock and Spitz [2] study using a MaxEnt model similar to the logistic regression model I chose as well as a Naive Bayes Classifier, and the USC study [6] using the same tokenization scheme that I adopted. Yet, the genre based classifiers have a much higher rate of success, regardless of the model chosen - the best success rate for any of my models are about 50%, which is also around the worst success rate for any of the genre classifiers. What this tells me is not that my models have failed, but rather that they have shed some light on the inconsistent and unpredictable nature of the MPAA maturity ratings.

It is difficult to directly compare the results of my study with the results from other studies

however, as I did not have the opportunity to directly control all of the variables and the data used. Ideally, I would have built my own genre classification models using the exact same algorithms as those used on the ratings, so that they could be more empirically compared to determine which is more objective. On the other hand, while genre is not as important to movie marketing and thus less susceptible to the persuasion of big film studios, it is a multilabel problem, which is a whole different issue to tackle. This means a singular film can fall into many different categories of genre, but can only have one MPAA rating, making it difficult to compare the two issues directly. Additionally, there are many more genres of movies than there are ratings, especially given that the only commonly used ratings today are PG, PG-13 and R. That being said, the ratings classifiers performed significantly worse than the genre classifiers in general, but further study is required to determine the exact cause of the disparity.

6. Conclusions and Future Work

The nature and scope of this project make it somewhat difficult to reach a definitive conclusion, but there are some notable pieces of information to take away. The first is that the maximum entropy logistic regression model seems to be best suited for this problem, as it produced by far the highest average ROC-AUC score at .5443, and the second highest f1 score, .4754. As I theorized in section 4.3, this is likely due to the fact that this model is built by identifying features that certain classes can't have, which I believe is particularly effective as things like swear words can easily differentiate between R rated and PG rated films. It's difficult to say whether the One VS Rest or the Naive Bayes classifiers is next best, as the One VS Rest has a much higher ROC-AUC score at .5066, compared to the .4056 of the Naive Bayes classifier, however the Naive Bayes has a significantly higher f1 score, .4962 where the One VS rest has a f1 score of only .3712. These results balance each other out to reflect almost equally poor performance by both models. At the very least, the results I have gathered reflect an inability by these three models to reliably predict the MPAA ratings of the test data. This suggests that the ratings may be influenced by outside data that are not reflected in the scripts. It is possible that this outside influence comes from the

bias of the MPAA, however, with the data that I have currently gathered, it is impossible to say so conclusively.

6.1. Future Work

The most important next step in continuing the work begun in this project would be to actually train these exact three models to classify based on genre rather than MPAA rating. This project is limited by a lack of ability to determine a cause for the results. It would greatly add to the analysis of the data if I were able to empirically compare the classification capability of these exact models on genre vs rating, as it would allow me to determine how much of he error is due to the models themselves. An increase in the number of models tested goes hand in hand with this idea, and would also help to strengthen the argument that no model of any type can predict rating based on script alone. If more models were used to predict both genre and rating, this would allow for even more precise determinations of which models work best on script data. So ideally, further work in this vein of study would be dedicated to finding and building more models to handle this issue, perhaps delving into some that are more specifically tailored to the exact data given.

7. Acknowledgements

Firstly, I'd like to thank Professor Christiane Fellbaum for her guidance and support throughout the development of this project, I would have been utterly lost without her. I would also like to thank Molly Pan for helping me with data processing and model implementation. The tools she gave me truly brought my project into being. I'd also like to thank my fellow classmates in COS 397, Natural Language Processing for their comments and critiques which made my project better and gave me direction every single week of the process.

8. Honor Code

I pledge my honor that this paper represents my own work in accordance with University regulations -Natalie O'Leary

References

[1] K. Afra, "Ratings creep, and the legacy of screen violence: The MPAA responds to the ftc's 'marketing violent entertainment to children'," *Cinema Journal*, vol. 55, pp. 40–64, 2016.

[2] A. Blackstock and M. Spitz, "Classifying movie scripts by genre with a memm using nlp-based features," 2008.

[3] S. Brin and L. Page, "The anatomy of a large-sclae hypertextual web search engine," *ScienceDirect*, vol. 30, pp. 107–117, 1998.

[4] T. Şengönül, "Negative effects of media on children and youth' socialization process: A study on violent and aggressive behaviors," *Faculty of Education Journal*, pp. 368–398, 2017.

[5] C. Danescu-Niculescu-Mizil and L. Lee, "Cornell movie-dialogs corpus," 2011.

[6] M. Fleischman and E. Hovy, "Recommendations without user preferences: A natural language processing approach," in *Proceedings of the 8th International Conference on Intelligent User Interfaces*, ser. IUI '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 242–244. Available: https://doi.org/10.1145/604045.604087

[7] D. Goldstein, "American moviegoers by age and ethnicity," 2017.

[8] R. Leone and L. Barowski, "MPAA ratings creep," *Journal of Children and Media*, vol. 5, pp. 53–68, 2011.

[9] MPAA, "Film ratings," 2019.

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] T. A. studentoflife, "Imdb top 250 lists and 5000 plus imdb records," 2017.

[12] M. Walker, "Film corpus 2.0," 2005.