Analysis of Movie Based on Rotten Tomatoes Review

Abstract

Movie industry is a huge revenue sector which in 2019 hits 32.6 billion US dollars record. It is the industry that movie producer is fighting to produce the best movies and getting the best reviews which relates to the higher number of audience and getting the award. There are a lot of movie reviews site that audience look up for the movie recommendation and ratings. One of the most popular platforms is Rotten Tomatoes. It is one of the most complete movie review platforms. It stores information from the multiple perspectives that includes critics and user's review and movie information that includes movie synopsis, movie director, list of casts and movie studios. In this project, the paramount aim is to provide useful information for the movie producer for their movies on getting the higher probability on the 'Fresh' category (term used as good quality in rotten tomatoes) and higher rating. There are 3 main studies being done in this project which are: Prediction analysis based on the movie information, prediction analysis based on the critics, and image classification based on the movie poster. Throughout our analysis, industrial talents, studio suggestions, guidance on utilizing critics' reviews and extra tips on poster design are provided based on results of our analysis (Check out our Github Page).

1. Problem Statement

The problem statement on this project is to assist movie producer on producing high quality movies. The tools that is provided from this project is as follows:

- Web Scraper to help the movie producer on extracting the data from the internet to gain additional information
- Prediction analysis based on the movie information to gain the insights on the movie information that get the higher chance of attaining 'fresh' category. Information include good genre, movie directors, actors, actress and movie studios
- Sentiment analysis based on the critic's review to gain insights the words that need to be avoided on creating the movie on getting the higher probability of 'fresh' category
- Image classification analysis on getting the pattern to be followed on getting the higher probability of 'Fresh' category

2. Dataset Description

The datasets are obtained from Kaggle website which includes the data from Rotten tomatoes website. The data are broken down into 2 categories which are movie information and movie reviews. For the movie information, this includes 16,638 movies with variables such as director name, writer name, cast, studio name and genre. On top of that, the movie information also includes movie poster URLs for our third study which is the image classification analysis.

Second category of the data is the critic's review and user ratings. There are in total 930,942 reviews. In rotten tomatoes, the critics review is classified as Tomatometer which is the score depicting the performance of the movie and the Tomatometer will be categorized in the end as 'Fresh' or 'Rotten'. Below is the breakdown of the definition from the critic's review and user rating:

Critics Review Categories	*Notes and Criteria*
Certified Fresh	At least 75% of critics' reviews are positive and 5 reviews come from top critics
Fresh	At least 60% of the critics' reviews are positive
Rotten	Less than 60% of the critics' reviews are positive
Audience Review Categories	
Upright	At least 60% of the audiences' reviews are positive
Spilled	Less than 60% of the audiences' reviews are positive

Table 1: Rotten Tomato Rating Criteria

For this study, we will treat Certified Fresh as Fresh in order to simplify the problem into binary classification and to improve calculation efficiency.

3. Prediction Analysis

For the prediction analysis, basic movie information was utilized to predict the Rotten Tomato review outcomes. This is to help movie producer understand what kind of movies are more likely to be rated positively on Rotten Tomato, so that they will likely produce high quality films by hiring appropriate talents and studios based on the findings of this study. The features include movie genre, PG rating, movie length, director, writer, cast member and studio name. Tomatometer status fresh or rotten was used as targeted variable for supervised categorical learning; audience rating and critics rating were used as supervised regression learning target variables. In total three models were built to analyze the effect of movie information on the audience and critics feedback.

3.1 Data Pre-processing

One-hot encoder was applied on movie genre, PG rating, studio name and all the human names for directors, writers and cast member. Only the first 5 cast members were included for this study; in addition, for director, writer, cast and studio, only those with at least 10 movies are included to reduce numbers of insignificant features and improve calculation efficiency. It is possible for one movie to have multiple directors/writers/genres. For the movie run time feature, standard scaler was applied to ensure its normal distribution. Null value for selected featured are marked as a separate one-hoc encoder for missing value for each feature category, i.e. "director NA" encoder feature for movie without director name. Records with targeting variable being null value were dropped. Eventually total 16386 valid records left, 5546 features were included, with 835 directors, 1050 writers, 3397 casts as one-hot encoders and others being genres, PG ratings one-hot encoders and standardized movie runtime.

3.2 Categorical Analysis

3.2.1 Model Training

For the categorical prediction on the Tomatometer status being Fresh or Rotten, 9 models were selected, out of which Decision Tree, Random Forest and Extra Tree Classifier are overfitting on the training dataset. Ultimately logistic regression was chosen as the final model with testing accuracy as 72%. Logistic regression was chosen also due to its interpretability and feasibility of further analysis which is shown in section 3.2.2. The training results are shown in Table 1.

Table 1. Model performance for categorical analysis

Models	*Training accuracy*	*Validation accuracy*
LogisticRegression	**0.82**	**0.72**
KNN	0.80	0.61
DecisionTree	1.00	0.64
RandomForest	1.00	0.69
AdaBoost	0.71	0.68
XGBoost	0.76	0.70
SGDClassifier	0.84	0.72
ExtraTreesClassifier	1.00	0.67
GaussianNB	0.66	0.59

3.2.2 Insights

Features importance was obtained using the logistic model tuned and trained. Results are shown below in Table 2.1 and 2.2.

Table 2.1 Top positive features for categorical analysis

Top Features	*Importance*
Cast	
cast Geena Davis	*0.47*
cast Jim Broadbent	*0.46*
cast Michael Cera	*0.40*
cast Meryl Streep	*0.40*
cast Ralph Fiennes	*0.39*
Directors	
director Steven Spielberg	*0.46*
director David Cronenberg	*0.46*
director Martin Scorsese	*0.42*
director Peter Weir	*0.41*
director Quentin Tarantino	*0.38*
Genres	
genre Documentary	*1.52*
genre Classics	*1.29*
genre Animation	*0.69*
genre Art House & International	*0.60*
genre Drama	*0.37*

Studios	
studio Sony Pictures Classics	*0.89*
studio Criterion Collection	*0.71*
studio Music Box Films	*0.70*
studio United Artists	*0.55*
studio Drafthouse Films	*0.51*

Table 2.2 Top Negative features for categorical analysis

Bottom Features	*Importance*
Cast	
cast Heather Graham	*-0.52*
cast Jennifer Lopez	*-0.44*
cast Danny Dyer	*-0.44*
cast Richard Burton	*-0.43*
cast Elvis Presley	*-0.42*
Directors	
director J. Lee Thompson	*-0.37*
director Tyler Perry	*-0.30*
director Sean McNamara	*-0.25*
director Otto Preminger	*-0.25*
director James Ivory	*-0.24*
Genres	
genre Faith & Spirituality	*-0.42*
genre Gay & Lesbian	*-0.31*
genre Action & Adventure	*-0.30*
genre Horror	*-0.24*
genre Mystery & Suspense	*-0.18*
Studios	
studio Freestyle Releasing	*-0.58*
studio Regent Releasing	*-0.55*
studio Screen Media Films	*-0.49*
studio Vertical Entertainment	*-0.44*
studio Freestyle Digital Media	*-0.39*

Several findings were made based on the results: Sony Picture Classics, Steven Spielberg, Documentary and Geena Davis are respectively the best studio, director, genre and cast in terms of Tomatometer status, they all showed strongly positive feature importance in the obtained logistic regression model; the top three genres are Documentary, Classics and Animation, which coincides with exploratory data analysis; overall, the top cast member and directors are generally active during 1990s era, suggesting that critics generally favors classics film over modern ones; action & Adventure is ranked third on the least favorite genre, and it is also the third most common genre in the dataset, suggesting that Action & Adventure, even though it is popular to make, it's very hard to win over critic's heart.

3.3 Regression Analysis

3.3.1 Model Training

6 models were trained for both audience rating and critics rating which both range from 0 to 100, out of which 1 is overfitted for both; XGBRegressor was chosen for its lowest mean square error over both target variables.

Table 3.1 Model performance for regression analysis – Audience rating

Models	*Training accuracy*	*Validation accuracy*
LinearRegression	160.25	2.36E+21
XGBRegressor	**227.68**	**310.31**
Ridge	170.29	372.07
PassiveAggressiveRegressor	218.48	494.12
HuberRegressor	184.51	412.09
MLPRegressor	327.33	398.36

Table 3.2 Model performance for regression analysis – Critics rating

Models	*Training accuracy*	*Validation accuracy*
LinearRegression	284.30	6.26E+22
XGBRegressor	**443.52**	**583.92**
Ridge	302.90	679.19
PassiveAggressiveRegressor	363.18	982.74
HuberRegressor	326.47	781.16
MLPRegressor	549.38	657.82

One observation is that for audience rating, the models generally have lower MSE than critics, suggesting that audience feedbacks are comparatively more predictable than critics using movie information.

3.3.2 Insights

Feature importance was obtained for both XGBRegressor models.

Table 4.1 Audience rating regression feature importance

Audience Rating Features	*Importance*
genre Horror	*0.02828*
genre Documentary	*0.02274*
genre Classics	*0.01646*
genre Special Interest	*0.01097*
genre Animation	*0.00834*
studio Criterion Collection	*0.00576*
genre Art House & International	*0.00511*
studio Sony Pictures Classics	*0.00441*
cast Akshay Kumar	*0.00439*
genre Mystery & Suspense	*0.00413*

Table 4.2 Critics rating regression feature importance

Critics Rating Features	*Importance*
genre Documentary	*0.01757*
PG Non-rated	*0.01475*
genre Classics	*0.01069*
genre Art House & International	*0.00562*
studio Sony Pictures Classics	*0.00495*
genre Animation	*0.00393*
genre Special Interest	*0.00387*
genre Drama	*0.00385*
director Joel Coen	*0.00358*
cast Jeff Fahey	*0.00287*

Some findings based on the obtained feature importance results are: genre Documentary both affects audience and critics rating significantly, followed by genre Classics, which again aligns with the exploratory data analysis that these two genres have higher average ratings; some other interesting facts are that genre Horror very significant on audience rating and PG rating Non-rated is also very significant for critics rating, which were both unexpected.

4. Sentiment Analysis

In this section, sentiment analysis would be leveraged on critics' reviews data. It would help movie producers identify audiences' perceptions on movies and therefore help with consolidating abundant movie reviews. The dataset consists of in total 60000 reviews that are randomly sampled from original 873206 valid reviews, ranging from 1941 to 2019. The features are basically raw texts from critics' reviews and the labels are critics icon which is either 'fresh' or 'rotten' submitted by critics for each review, which serves as a proxy to their sentiments.

4.1 Data Pre-processing

A standard procedure of text preprocessing is applied on the review contents. It includes but not limited to steps as follows:

- Lowercasing
- Noises removal including punctuations, urls, special characters
- Removal of digits and stop words except for sentiment-related words like 'not' or 'no'
- Lemmatization, e.g., 'running' -> 'run', 'has' -> 'have', 'hours' -> 'hour', etc
- Contraction mapping, e.g., 'i'm' -> 'i am', 'she'll' -> 'she will', etc
- Tokenization

Note that pertaining to dealing with digits, in addition to simply removing it, there is another normalization method which translates digits to English words such as '8' to 'eight' and '1000' to 'one thousand'. Concerning the time-consuming drawback of this method without obvious improvement on model performances, we decided to stick with removal of digits. Lemmatization, which is the process of converting a word to its base form, is more suitable to use compared with stemming, in the sense that lemmatization considers text and converts a word to its meaningful base form rather than simply erase the last few characters like stemming does. Nevertheless, its usefulness is not guaranteed for our task. In order to determine its effectiveness, we would use different model configurations with and without lemmatization and compare their performances.

4.2 Feature Engineering

As both machine learning models and deep learning models would be used for this task, the core features are somehow different for these two sets of models. Word features created from CountVectorizer are main features for machine learning models, whereas word embeddings and character embeddings are main features for neural network models. Compared with TfidfVectorizer, CountVectorizer was finally selected as we do not want to penalize words with high frequencies across different reviews. Pertaining to word embeddings, both pretrained word embeddings such as Glove and Fasttext and non-pretraining word embeddings from Keras embedding layer were experimented in order to find the most suitable approach for our problem (Jeffrey, Richard & Christoper, 2014; Piotr et al., 2016). Specifically, models with pretrained embeddings employed a stacked word embeddings of both Glove (100 dimensions) and FastText (300 dimensions) due to an motivation to somehow mitigate the out-of-vocabulary (OOV) problem since

FastText take cares of sub-word information. In addition, we also created a few review-level hand-crafted features to be used for both sets of the models. Specifically, they are review length (no. of words), number of sentences and number of URLs in the review.

4.3 Model Training and Selection

In this section, we would describe models in use and model results, which are shown in Table 1 below. Under machine learning categories, logistic regression, svm (linear-kernel) were selected due to their simplicity, interpretability and decent performances on high dimensional problem. Xgboost and Lightgbm were selected because of their inherent boosting algorithms. The bottom five models are under neural network categories with descriptions as follows:

- NN 1: lemmatization on reviews, non-pretrained word embeddings
- NN 2: lemmatization on reviews, pretrained word embeddings
- NN 3: no lemmatization on reviews, non-pretrained word embeddings
- NN 4: no lemmatization on reviews, pretrained word embeddings
- NN 5: no lemmatization on reviews, non-pretrained word embeddings, character embeddings

Pertaining to the structure of these five neural network models, there are intrinsically quite similar with each other. Specifically, NN1 to NN4 have almost same structure with the only change on the embedding layer. Hence, we would focus on discussing about the structure of NN3, the best model, which can be found in Appendix A-1. The length of input sentence is 50, as 99% percentile of the reviews have less than 49 words. The main structure is built on top of 1D CNN and BILSTM due to a motivation to combine both the feature detection power from CNN and long sequence learning capability from LSTM. As mentioned in Section 4.2, the hand-crafted features were also involved in the model to further improve the model capability. As for the structure of NN5, shown in Appendix A-2, has another branch from character inputs. Basically, we used 1D CNN to learn the hidden pattern stored at a character-level to resolve the out-of-vocabulary issue. This technique has been widely used and some seminal works, for example, involves using a hybrid bidirectional LSTM and CNN architecture to automatically detect word-level and character-level features in order to solve a name entity recognition (NER) task on Conll-2003, a standard corpus for NER task (Chiu & Nicolas, 2015; Sang & Meuler, 2003).

From the Table 5 below, it can be identified that Xgboost after hyperparameter tuning performs best among the machine learning categories with a test accuracy of 75.14% and neural network without lemmatization and without pretrained word embeddings notches the highest test accuracies of 78.23% among all the models.

Table 5. Sentiment analysis model performances

	Training Accuracy	*5-fold CV Accuracy*	*Test Accuracy*
Logistic Regression	0.8804	0.6958	0.7258
SVM (linear-kernel)	0.8091	0.6855	0.7043
Xgboost (base)	0.8799	0.7386	0.7473
Xgboost (tuned)	0.**8810**	0.**7403**	0.**7514**
Lightgbm	0.8172	0.6933	0.7395
*NN **1***	0.8490	-	0.7751
*NN **2***	0.9286	-	0.7712
NN 3	0.**8505**	-	0.**7823**
*NN **4***	0.8800	-	0.7815
*NN **5***	0.8352	-	0.7765

Some other findings include:

- Lemmatization might not be suitable for the neural network models in our context.
- Using pre-trained word embeddings has a potential to increase training accuracy but not test accuracy, which seems to be prone to overfitting.
- Although resolving the out-of-vocabulary problem, character embeddings might not boost up the model performance.

There are some potential ways to further boost up the model performances.

- Feature engineering: As there are only three hand-crafted features embedded in our models, increasing the number of effective hand-crafted features could help increase the accuracy scores, e.g., number of adjective words used in a review.
- Hyperparameter tuning: The results suggest neural network models are relatively superior to machine learning models. Calibrating the model by hyper parameter tuning would possibly improve the model performances further.

Nevertheless, as a tradeoff, the time consumed would also surges up tremendously, which might be applicable to movie producers in practice.

4.4 Discussion & Use Case

4.4.1 Discussion

After obtaining a fitted model, we can check the feature importance to learn what features contribute to the predictive decision. From the summary plot in Figure 1 based on the tuned Xgboost model via Shap(Lundberg & Lee, 2017), it shows some relevant words that have positive impact to 'fresh' label include sturdy, resist and explore. On the other hand, words like pointless, alas, tepid contribute to the 'rotten' label. What is noticeable is that almost all these important tokens have strong contribution towards either fresh or rotten, except for the word 'not'. The fact is that a review content with more occurrences of 'not' might mean a positive sentiment of the critics, for example as listed in Appendix A-3, "I can't recommend this film enough and can't imagine anyone not being bowled over by it."

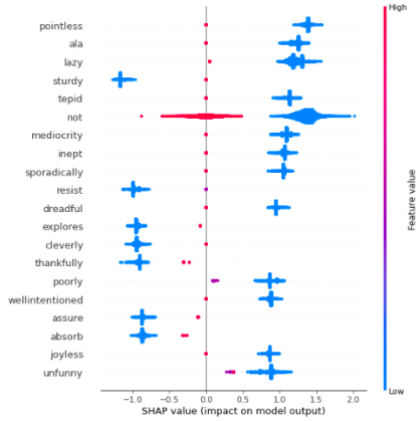


Figure 1: Sentiment Analysis - Feature Importance Plot

4.4.2 Use Case

There are many applications of our sentiment analysis model that movie producers could leverage on in practice. For example, as one simple use case, they could use the model to process a great amount of audiences' reviews and perform data analysis on the predicted sentiments or perceptions of audiences to obtain a general understanding on movie quality. In this section, we would demonstrate another particular use case based on feature importance plots. Basically, our assumption is that based on historical critics' reviews, movie producers can capture what features to focus on and what drawbacks to prevent for future productions. Here we handpicked a 5-year observation period from 2015 to 2019 and fitted a Xgboost model on these historical reviews in different genre. From the summary plot for 'Horror' genre as shown in Figure 2, some insights for 2020 productions are for example, that the movie plot should not be tedious nor pretentious and the screenplay is best to be sheer and efficient.

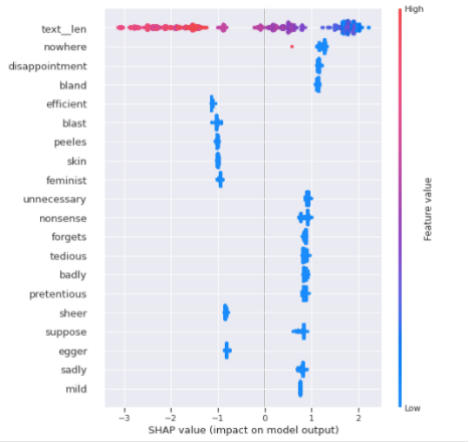


Figure 2: Sentiment Analysis - Use Case for Horror Movies

Likewise, for Comedies, some suggestions could be, for instance, the plot should be refreshing, funny but not depressing. In addition, there is a specific caution when exercising montage techniques in the movie.

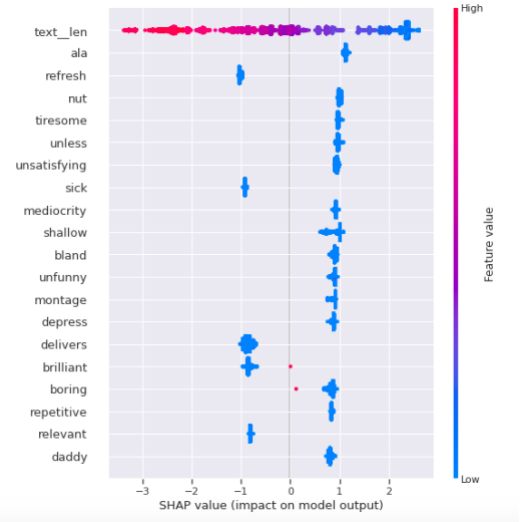


Figure 3: Sentiment Analysis - Use Case for Comedy Movies

5. Image Classification

5.1 Data Pre-processing

From the dataset, more than 16,000 valid posters were extracted using URLs with each one of size (305, 206, 3). Every poster was then resized to be (224, 224, 3) for the

convenience of deep learning model pipeline. Some features of the posters were explored: denoised the images to separate background from foreground, and identified different objects with marked segments (Getreuer, 2013). Examples of image exploration are shown in Appendix B-1. For the sake of interpretation, only the resized posters were then transformed into a large NumPy array of size (16000, 224, 224, 3), which were passed to the deep learning models, together with their corresponding movie critics rating, either being fresh or rotten as labels. This image classification task was objective to discover the correlation between movie posters and the probability of the movie being rated fresh, hopefully, providing insights for movie producers on the poster design.

5.2 Techniques used for boosting CNN performance:

5.2.1. Data Argumentation:

Image argumentation strategy aims to boost the performance of deep learning networks (Shorten & Khoshgoftaar, 2019). It is necessary due to two main reasons: one is that deep neural networks require a large amount of training data to prevent overfitting, the other one is that the orientation of the image could hinder the model performance, since model lacks of ability to generalize. Therefore, the image augmentation technique was applied here to artificially creates training images through different ways of processing or combination of multiple processing, such as random rotation, shifting, shearing, zooming and flipping of the existing posters. It was implemented using Keras package *ImageDataGenerator* API, two examples are shown in Appendix B-2.

5.2.2. Transfer Learning:

In transfer learning, one repurposes the learned features, or transfer the knowledge from a relevant trained task to a second target network wait to be trained (Torrey & Shavlik, 2010). Some pre-trained deep learning models are VGG-16 from Oxford (Simonyan & Zisserman, 2015), and Inceptions from Google (Szegedy et al., 2015) were applied in this project. Take the VGG-16 as an example (with its structure illustrated in Appendix B-3). It contains 13 layers: 5 blocks of convolution layers each with a max-pooling layer for down-sampling; 2 fully connected dense layers, and 1 output layer containing 1000 classes. For transfer learning, the last three layers of the VGG-16 were dropped, and replaced by our own fully connected dense layers to do the binary classification on whether the posters will be fresh or rotten using output layer with "sigmoid" as the activation function.

One strategy is to use VGG-16 as a feature extraction tool, where all blocks' weights will be fixed (non-trainable). While the often-used strategy is to replace and retrain the classifier on top of the pre-trained networks on the new dataset and also to fine-tune the weights of the pre-trained network by continuing the backpropagation. It is efficient to only fine-tune some high-level portion of the network while keep earlier layers fixed. Because it is discovered in research (Nogueira et al., 2017) (and later demonstrated with illustration in session 5.4) that the earlier features of a pre-trained network learn more generic features, whereas the deeper layers of the network become progressively more specific to the details of the classes contained in the original dataset. With this strategy, the network is more capable of recognizing the poster patterns. The structure of both strategies using VGG-16 pre-trained network is shown in Appendix B-4.

5.3 Model Training and Selection

The dataset was divided into train (7,680 images), validate (5,120 images) and test segments (3,200 images). A basic CNN model with three convolutional layers, coupled with max pooling for auto-extraction of features from the poster images and also down-sampling the output convolution feature maps. From the result, it can be seen that the model is overfitting after 5-7 epochs. Then another convolution layer with dense hidden layer was added to the basic CNN, with dropout of 0.3 after each hidden layer to enable regularization. Dropout randomly masks the 30% of units from the hidden dense layers and set the outputs to 0. However, the results still ended up overfitting around 75% (better than 99% in basic CNN, but still not good enough). Thereafter, image augmentation strategy was applied to the existing posters and fed them to the CNN. This quite improved the overfitting issue, and the accuracy also jumps from 55% to 58.5%. In the next step, pre-trained CNN models were leveraged to further boost up the performance with *transfer learning*. There were three different pre-trained models used in this project. They were VGG-16, InceptionResNetV2, and InceptionV3. The VGG-16 was initially used as a simple feature extractor by freezing all the five convolution blocks to make sure their weights were not updated after each epoch and only train the model on two more dense layers added after the VGG-16 pre-trained model. To further fine-tuned the pre-trained model, last two blocks of VGG-16 model were unfrozen (Block 4, and 5) and their weights were getting updated in each epoch as the model was trained. The results showed that with fine tuning, the best model was obtained with validation accuracy boosted up to 62%. Similarly, pre-trained model InceptionResNetV2 and InceptionV3 were also used with fine tuning on the convolution blocks. InceptionResNetV2 boosted up the training accuracy up to 66%, and validation accuracy above 62%, thus used as our final model. The model was tested on the untouched posters, gave us the test accuracy of 62.3% as the best performance, and the classification report with ROC plot were shown in Appendix B-5.

Table 6. Model performance and improvement

Models	*Training accuracy*	*Validation accuracy*
Basic CNN	0.99	0.55
Regularized CNN	0.75	0.56
Regularized CNN with Image Argumentation	0.59	0.58
Transfer Learning VGG-16 (Feature Extractor)	0.62	0.61
Transfer Learning VGG-16 (Feature Extractor) with Argumentation	0.62	0.61
Transfer Learning VGG-16 with Fine-Tuning and Argumentation	0.62	0.62
Transfer Learning InceptionResNetV2 with Fine-Tuning and Argumentation	**0.66**	**0.62**
Transfer Learning InceptionV3 with Fine-Tuning and Argumentation	0.64	0.61

The accuracy of 62.3% for a binary classification problem was not high. To better interpret the result and gain more insights, the visualization of intermediate layers of the best model and CAM techniques were applied afterwards to discuss the contribution of the image classification task to the achievement of the project.

5.4 Visualization of Intermediate layers:

The pre-trained deep CNN models used for transfer learning, like VGG-16 (visual geometry group) and inception architecture, are meant for achieving more effective feature extraction using the existing acknowledge with less data availability. To better understand the pre-trained model on how it is able to classify the input image, it is useful to look at the output of its intermediate layers. Take the 3 convolution layers and their coupled activation layer in Figure 4: (the other two layers displayed in Appendix B-6).

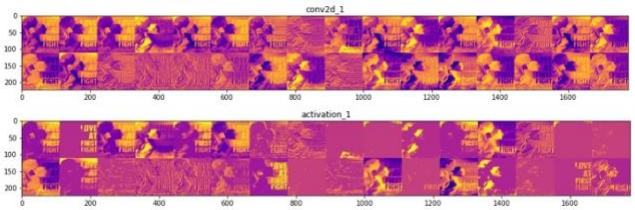


Figure 4: Visualization of 1st convolution (activation) layer

With the help of these visualizations of the intermediate layer, it is pretty clear to see how different filters in different convolution layers are trying to highlight or activate different parts of the input image. Some act as edge detectors, others detect a particular region of the poster like the title, darker colored portion, or background. It is easier to see this behavior of convolution layers in the starting layers (more general patterns), since as model went deeper the pattern captured by the convolution kernel become more and more sparse. Deeper the layers in the network, more training data specific features were visualized. This is consistent with the research finding stated in fine-tune strategy of transfer learning (in 5.2.2 session). In the next section, another visualization technique, Class Activation Maps (CAM) was used to deliver the insights of our output to movie producers.

5.5 Class activation map & Insights

Class Activation Maps (CAM) technique is widely used when interpretation of the output is crucial in the use case. For the poster classification, since the objective is to provide movie producers insightful information on the design of posters, CAM was an essential step of the project delivery. A class activation map for a particular category indicates the discriminative region used by CNN to identify the category. This is achieved by projecting back the weights of the output layer on the convolution feature maps obtained from the last convolution layer (GAP layer, which takes an average cross all the activation to find all the discriminative regions) (Zhou et al., 2016).

Based on the predicted probability of being fresh, the top 8 (most likely fresh) and bottom 8 (most likely rotten) posters were selected, and then displayed with CAM techniques applied to see what the model has learned to predict the posters output classes in Figure 5 and 6 (The rest selected posters are displayed in Appendix B-7).

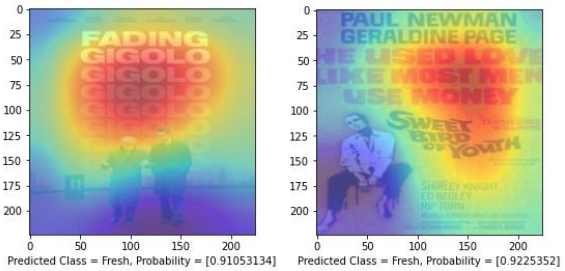


Figure 5: Predicted to be rated as "Fresh"

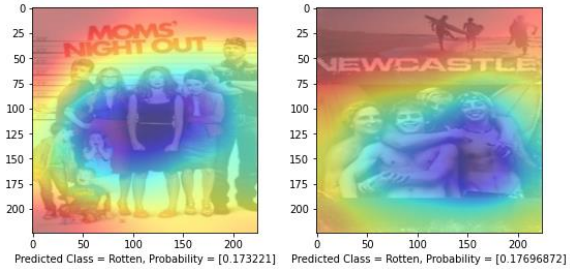


Figure 6: Predicted to be rated as "Rotten"

The heatmap shows the score (weight) that the location of the image possesses in predicting the output class of the image. Since this is a binary classification using "sigmoid"

function as the output layer, we can focus on the warm region for fresh class, and cold region for rotten class. What the model learned is impressive. For the top posters that are predicted fresh (with probability > 0.9), they all have very big title (some titles are stacking) occupying most of the poster content space; and there are very few characters in the poster. For those predicted to be rotten ($p < 0.2$), the posters tend to have a lot of characters filling up almost the whole poster, and most of these characters are standing in parallel or other regular patterns.

In terms of accuracy, some of these top predicted posters were mis-classified, since the prediction model only has an accuracy of 62.3% on the test posters. There are some limitations of the final image classification model: Title dominant rule. Titles are supposed to be part of the poster and it is the purpose of delivering what kind of font or layout are more positively catching audience's eyeballs. However, the model now seemed to decide as long as there are plenty of large font of texts inside, then it is fresh. Secondly, the accuracy is relatively low could be due to the effect of movie genre or year of release. The styles of posters could be similar and of mainstream within a period of time or of a certain genre. And a comedy poster tends to be fresh could look extremely different from an attractive thriller poster. Last but not least there are so many other factors that affect the feedback of the movie: like the ones we analyzed in the prediction analysis section: directors, casts, and most importantly the storytelling. The goal of the poster analysis is not giving director absolute insights barely from image, but to provide additional useful tips before the movie production to help them make better decisions.

6. Future Work

As for the recommendation on future work, the further study could use the web scraper to extract various movies' parameters to get the relevant information. Then, the study could include the new target variables which includes the movie revenue and movie's audience count. Lastly, the further work could extend our study for another movie's review platform such as IMDB.

7. Conclusion

In conclusion, the result from all the analysis are as follows. For the prediction analysis on the movie parameter result, the movie producer should choose documentary as genre, Steven Spielberg as director, Sony Classics as movie studio and Geena Davis as cast. On the prediction analysis for critic's review, for movie producer that would like to do horror movies, the movie producer should focus on the being efficient on the plot and not pretentious. If the director would like to do comedy movies, they should focus on not being depressing and cautious on using montage. Lastly, on the poster image classification analysis, for genre drama, the poster should have less character with big fonts. If the poster created for comedy, the poster should avoid on having too many movie actors in 1 poster.

However, the movie producer should take the analysis as a reference and to do more analysis from the multiple perspectives. Moving forward, movie producers could use our study for their inspirations including web scraper as a tool to extract more movie data from rotten tomatoes or prediction analysis of movie to understand the current trend for movie category. In addition, by leveraging the sentiment analysis on critic's reviews, movie producers could extract the relevant words to be incorporated to the movie. Lastly, the image classification could be employed to understand the trend of movie poster which is considered 'Fresh'.

References

[1] Bojanowski, Piotr & Grave, Edouard & Joulin, Armand & Mikolov, Tomas. (2016). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. 5. 10.1162/tacl_a_00051.

[2] Chiu, Jason & Nichols, Eric. (2015). Named entity recognition with bidirectional LSTM-CNNs. Trans. Assoc. Comput. Linguist.. 6. 10.1162/tacl_a_00104

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, Rethinking the Inception Architecture for Computer Vision (2015), http://arxiv.org/abs/1512.00567

[4] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations (2015)

[5] Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.

[6] Nogueira, Keiller, Otávio AB Penatti, and Jefersson A. Dos Santos. "Towards better exploiting convolutional neural networks for remote sensing scene classification." Pattern Recognition 61 (2017): 539-556.

[7]. Pascal Getreuer, A Survey of Gaussian Convolution Algorithms, Image Processing On Line, 3 (2013), pp. 286–310.https://doi.org/10.5201/ipol.2013.87

[8] Pennington, Jeffrey & Socher, Richard & Manning, Christoper. (2014). Glove: Global Vectors for Word Representation. EMNLP. 14. 1532-1543. 10.3115/v1/D14-1162.

[9] Sang, Erik & Meulder, Fien. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceeding of the Computational Natural Language Learning (CoNLL). 10.3115/1119176.1119195.

[10] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

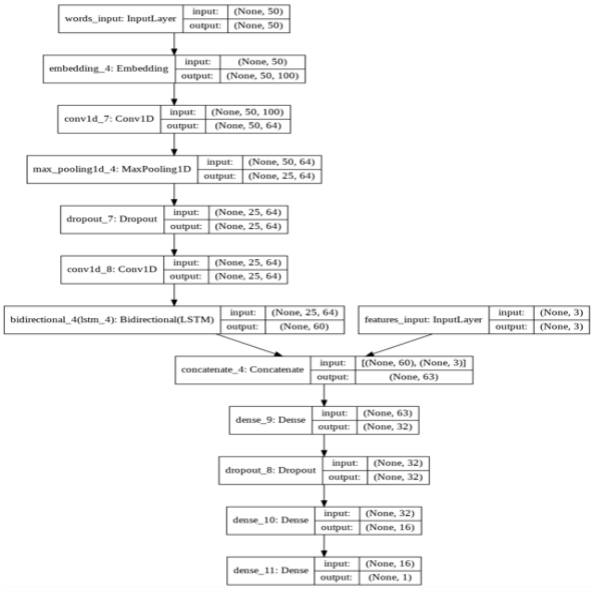
[11] Torrey, Lisa, and Jude Shavlik. "Transfer learning." In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242-264. IGI Global, 2010.

[12] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921-2929. 2016.

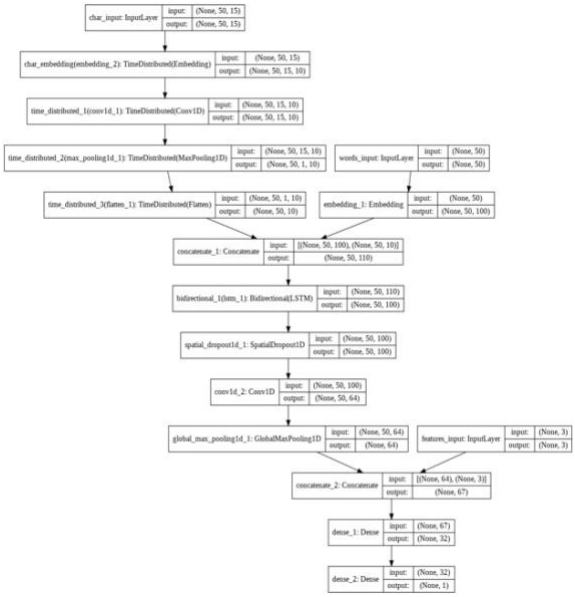
Appendices

A. Sentiment Analysis

1. NN3 Neural Network Model Structure



2. NN5 Neural Network Model Structure

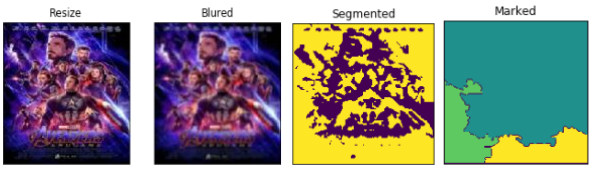


3. Fresh Reviews with a Few Occurrences of 'not'

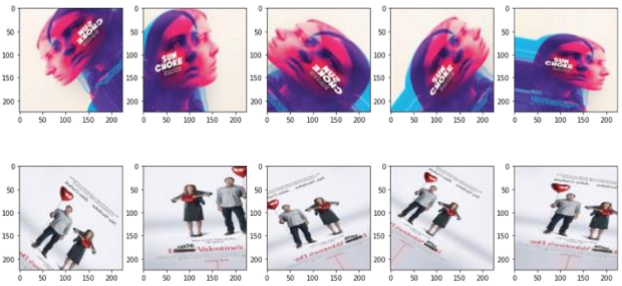
Movie	Fresh Reviews with High Occurrences of 'not'
A STAR IS BORN (Tomatometer: 90)	"I can't recommend this film enough and can't imagine anyone not being bowled over by it."
MYSTIC RIVER (Tomatometer: 88)	"I can't figure out why, but it didn't really stick with me once I left the theater. It's very good, but not transcendent."
STAR WARS: EPISODE VII - THE FORCE AWAKENS (Tomatometer: 93)	"It's precisely what you expect. Nothing more. Nothing less. And there isn't anything wrong with that."

B. Image Classification

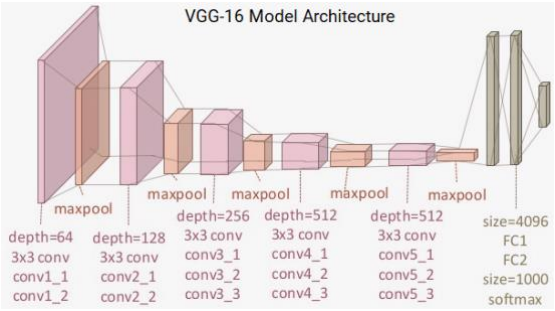
1 Image preprocessing (resizing, denoising, separating background, detecting objects)



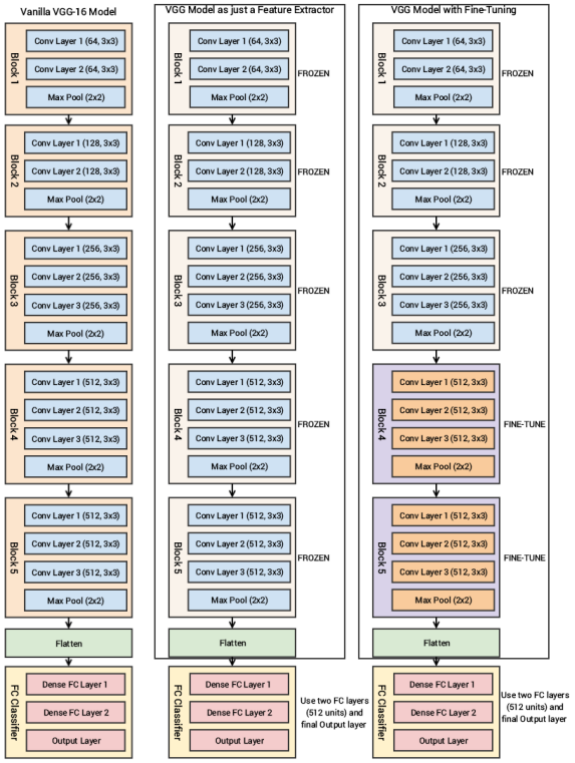
2. Image Augmentation samples



3. Pre-trained VGG-16 model architecture

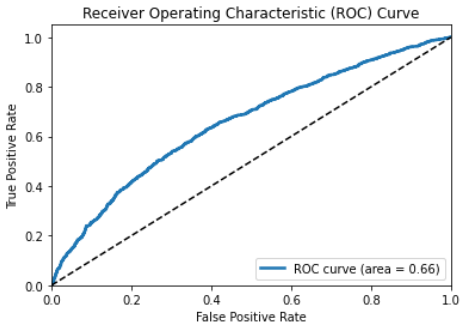


4. Transfer learning strategies (Feature extraction & Fine-tuning)

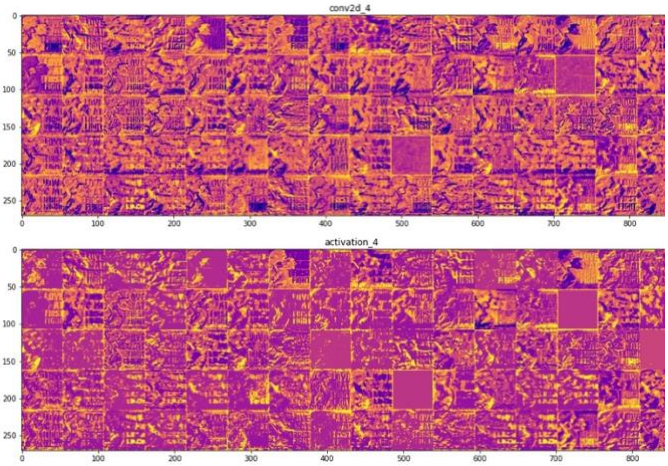


5. Test matrix with ROC plot

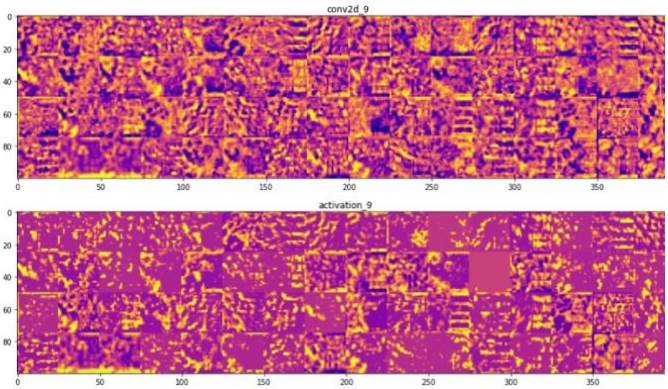
```
Model Performance metrics:
------------------------------
Accuracy: 0.6234
Precision: 0.6135
Recall: 0.6234
F1 Score: 0.6082

Model Classification report:
------------------------------
              precision    recall  f1-score   support

       Fresh       0.65      0.79      0.71      1870
      Rotten       0.57      0.40      0.47      1330

    accuracy                           0.62      3200
   macro avg       0.61      0.59      0.59      3200
weighted avg       0.61      0.62      0.61      3200
```



6. Visualization of the intermediate convolution layers

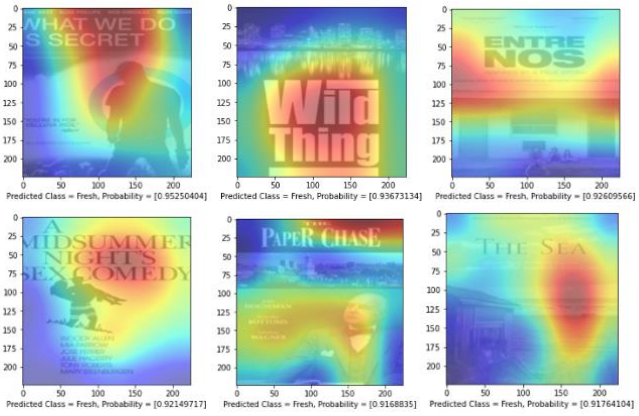


Visualization of the 4th convolution (activation) layer

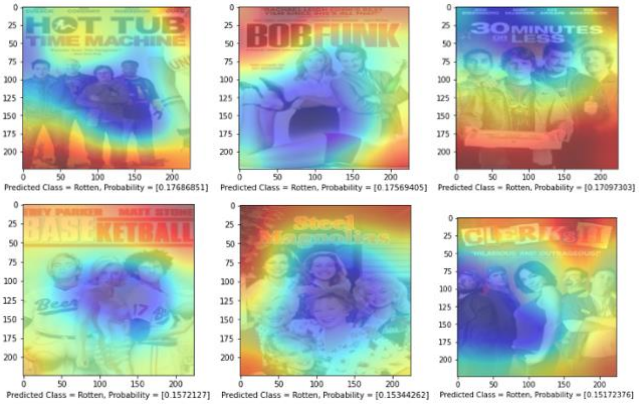


Visualization of the 9th convolution (activation) layer

7. The rest CAM heatmap for top predicted posters



Predicted Fresh



Predicted Rotten