DeepIMDB: Multimodal Movie Rating and Revenue Prediction

Shikha Asrani, Samir Char, Srujan Esanakarra

Advanced Topics in Deep Learning

The production of a film is an expensive, risky investment for producers. The majority of films fail to recoup their production budgets at the box office and incur significant losses. Existing decision support systems that aim to analyze and predict the performance of movies prior to their release using pre-release information lack accuracy and are do not meaningfully benefit the stakeholders. These models primarily use post-production and post-release data, such as opening weekend data, that is only available immediately prior to or immediately after release. Such forecasts, however, are of little value to filmmakers as they can only influence late-stage adjustments and not the significant levers in the production process.

For filmmakers and the movie industry, the ability to accurately predict a movie's rating and box-office revenue before its release is immensely valuable as it can guide their decision making and execution and reduce financial risk. This is not simple, however, as ratings and revenues are a result of objective and subjective factors and there is a complex relationship between the data available prior to release and the ratings and revenues. Since the subjective factors cannot be measured prior to release, the objective factors must be used for prediction. In this paper, we propose a multimodal deep neural network architecture that uses categorical data such as title and synopsis, numerical data such as runtime, budget, and year of release, and visual data in the form of posters for rating and revenue prediction. We experiment with various approaches to the multimodal architecture to process the various types of inputs for the prediction task. To evaluate the performance of the proposed model, we carried out comparative analyses with other prediction techniques. Experimental results demonstrate the potential of the proposed multimodal model for movie rating and revenue prediction.

Multimodal Model | Movie Ratings and Revenues Prediction

Link to Code

Introduction

The movie business is one of the riskiest endeavors for investment as it is fraught with difficulty and unpredictability in forecasting the demand (1). Only thirty to forty percent of movies break even and just about one out of every ten movies becomes profitable at the box office (2, 3). Movies require significant upfront capital and run the risk of swift swings in demand as they are likely to only be viewed once in a theater by any given individual and many are influenced by reviews and early talk (4). This makes the opening weekend crucial for a movie's performance, as it accounts for nearly 40% of total box office revenue on average and is a strong predictor of a movie's overall financial success (5). Thus, existing models rely on post-production or post-release data and are able to forecast a movie's performance with great accuracy; however, such forecasts are of little use as they can only influence late-stage adjustments to advertising or distribution strategies with little effect on the overall success.

It has been said that "Hollywood is the land of hunch and the wild guess" (6) and Jack Valenti, president and CEO of the Motion Picture Association of America, once stated that "no one can tell you how a movie is going to do in the marketplace...not until the film opens in darkened theatre and sparks fly up between the screen and the audience" (7). This sentiment is echoed across the various movie industries around the world and in trade journals, interviews, and analyses. However, there is immense value in being able to predict how a movie will perform prior to its release, which has been a long sought ability.

The movie industry is characterized by frequent losses and infrequent blockbusters that offset the many failures. In 2019, the most recent year unaffected by the COVID-19 pandemic, 83.2% of gross box office revenues came from just 68 movies or 10% of the films released in that year (8). In the five year period between 2015 and 2019, the average production budget for a movie was $39.3M, and more than half of the movies in this period failed to recoup their production budgets in box office revenues, with an average deficit of nearly $19M (8). Box office performance increasingly depends on a small number of blockbusters, all the while the number of filmmakers, ideas, and projects continue to increase. Thus, investors face the need to be more selective in their choices of which projects to green-light and have an ever-growing need for accurate decision support systems.

The success of movies is largely measured by their box office revenues and, to a lesser extent, their ratings by critics and audiences. While movies are distributed through many media today, they are still primarily released through theaters first, which influences their success in the secondary channels, and thus, the box office revenue is still a crucial indicator of success (9). While these metrics are imperfect and incomplete, more often than not, they guide the decisions of studios and producers and determine which movies are viable investments. Existing decision support systems for pre-production and green-lighting analysis are not accurate enough to meaningfully guide filmmakers. This leads to a warped perception of movies in popular culture and the suppression of good stories and ideas if they do not have a franchise tag or big stars. Therefore, the ability to predict a movie's rating and revenue prior to its theatrical release is highly valuable to producers and filmmakers, who can balance risk and make more informed decisions on which movies to produce. Such predictions will also inform a movie's scale, advertisement budget, special effects, and all stages of the production process.

The complexity and challenge of the problem has attracted various attempts to develop models for forecasting the demand and financial success of movies. Early studies used statistical forecasting techniques and found that box office revenues tend to tail off after the opening week and that 25% of revenues come from the first two weeks, so once the first week's revenue is known, the total revenue can be predicted with high accuracy (10). Thus, the challenge lies in predicting revenue, especially for the first week, prior to the movie's release. Over the years, there have been numerous attempts to tackle this problem. One approach framed it as a classification problem categorizing movies into 'hits' and 'flops' using a multi-layer perceptron and achieved 36.9% accuracy (10). Others used clustering but incorporated both pre-release and post-release features (11), social media sources including IMDb, Twitter followers, and YouTube comments with three categories for classification (hit, neutral, or flop) (12), convolutional neural networks using movie posters (13), and even the level of activity on movies' Wikipedia pages (14). These studies achieved varying degrees of success in the prediction of box office revenue, but are limited in their applicability to real-life scenarios.

Movies have a diverse set of features associated with them that can be useful in prediction to varying degrees. Among these are title, synopsis, cast, director, language, country, production company, genre, run time, year of release, and poster. The International Movie Database (IMDb) (15) is a well-known online database of information related to films, television series, home videos, video games, and streaming content online. IMDb provides users with information about cast, production crew, personal biographies, plot summaries, trivia, ratings, fan and critic reviews, and more. This data has previously been used in tasks such as movie genre classification using posters, since they are highly related (16). However, posters do not tell the whole story, and other data should be included for more accurate results.

In this paper, we propose a multimodal neural network architecture to predict the ratings and revenues of movies using data available prior to release. We explore various approaches to this problem and tackle it as an ablation study to develop the architecture. The proposed model takes as input categorical, numerical, and visual features about movies and predicts their ratings and box office revenues as outputs. It extracts the most relevant features for the outputs and combines the best information from the various types of data to improve prediction performance. This model improves the forecasts of a movie's performance to better guide the decisions of stakeholders and benefits those who have a significant vested interest in the success of films.

Related Work

The prediction of movie ratings and revenues can be approached in numerous ways that differ in the data features used, the timing of the data, and the prediction techniques. The features can include information such as genre, runtime, posters, etc. and different experiments select different subsets. The timing of the data can be pre-production, post-production but pre-release, post-release, or any combination of these. The prediction techniques range from simple statistical modeling to machine learning to deep learning.

A study using factor analysis and multiple linear regression modeling predicted gross box office revenue for movies using both pre-production and post-release data and found that statistical modeling could not accurately predict box office revenue with just pre-production data, but achieves a significant improvement in accuracy when post-release data is included (17). Another study uses cosine similarity and clustering with k-Nearest Neighbors (kNN) to predict movie ratings using pre-production data and achieves reasonable accuracy, measured as 83% of the predicted ratings are within a range of 1 from the actual ratings (18). A third approach predicts both rating and revenue using XGBoost regressors, but only achieves reasonable accuracy for revenue prediction and is not generalizable as a lot of the data is thrown out and preprocessing is not thorough (19).

Others have used machine learning techniques to approach this problem. One study used movie trailers as inputs for a linear support vector machine (SVM) classifier to predict the opening-week box office revenues, in which audiovisual features were extracted manually from trailers to use for the prediction task (20). The study found that these features help in improving prediction performance. In another study, post-release data and data from social media was used in several nonlinear regression algorithms to build forecasting models for box office revenue (21). These approaches, however, are limited by the need for careful manual feature extraction and do not work well to reveal complex relationships that are inherent in movie-relate data.

A third set of studies approach the problem using deep learning techniques. The first of these extracts features from movie subtitles, posters, and metadata and uses classification and regression models to predict ratings, achieving 0.7 mean square error for the regression model (22). A second study uses multimodal data for the prediction of ratings, including visual, text, and categorical features and approaches the problem using an array of techniques, including linear and ridge regression, decision tree and random forest, SVM, and neural networks (23). The study found that random forest achieved the best results with a test mean square error of 0.85, but the $R\hat{2}$ was only 0.42. This shows that there is room for improvement for this task using better models. This was demonstrated in a third study that also used a multimodal approach incorporating a CNN for movie posters and a multilayer perceptron for the other data. This approach significantly outperformed the random forest, SVM, and neural network techniques in the prediction of box office revenues (13).

In this paper, we are building off of this work, but taking it further by performing a series of ablation studies to identify the best model for each modality of data, which will then be combined to produce the best multimodal model for both rating and revenue prediction. We differ from prior studies in our use of state-of-the-art deep learning models and approaches. While these approaches have been extensively used for other problems, their use in movie rating and revenue prediction is sparse and remains to be explored. We aim to leverage these techniques to improve on existing models and better predict movie ratings and revenues using pre-release data.

Methods

The IMDb website has a subset of the data publicly available for free download at (15). Additionally, we use The Movie Dataset (TMD) from Kaggle (24) to include variables not available in the IMDb dataset. However, TMD is 4 years old and there is missing and outdated data, which is why we also use the IMDb official API Cinemagoer (25) to fill in missing values and update the outdated information. Between these three data sources, we have most of the information except for the movie revenue, which we only have for a small number of movies. To solve this problem, we web scrape Box Office Mojo (26), an official IMDb website that specializes on box office data. We scrape the data because Box Office Mojo does not have publicly available data nor an API. After cleaning and merging the data, we decide to narrow down the scope of the project by applying the following filters:

- We only focus on movies. The IMDb database also has TV series, video games, etc.
- We keep movies whose original language is English. This does not mean that the movie might be available in other translated languages.
- We keep movies that were released on or after 1980.

After this process, we get to a dataset of 11,694 movies with 14 variables that are described in table 1.

We tackle this project as an ablation study:

1. Establish a performance baseline with Random Forest Regressor using only tabular data (categorical,numerical, and date)
2. Train single modality models to answer two questions:
   - What is the best performance we can achieve using only text features (synopsis and title)?
   - What is the best performance we can achieve using only the movie posters?
3. Combining text and tabular: identify the best way of combining tabular and text data (two modalities, without the image poster)
4. Combine all three modalities into one model

We attempt to do each of the previous steps for both rating and revenue prediction.

Measuring Performance. Movie ratings are between 1 and 10 and with only 1 decimal. The distribution of movie ratings is close to normal and bounded, so we simply use Mean Squared Error (MSE) to measure performance.

On the other hand, the distribution of movie revenues has a very long tail on both ends. There are movies that only made a couple of hundred dollars of revenue, while there are movies like Avatar that made close to 3B USD (27). This makes this target particularly challenging, so we decide to train the model on *log(revenue)* to mitigate the effect of the distribution and the outliers. Nevertheless, on the test set, we measure performance based on the Mean Absolute Error (MAE) of the inverse transform of the model predictions (i.e. exp(predictions)), which yields predictions on the same units as revenue (US Dollars).

Baseline. The first set of experiments we perform are to establish a baseline for prediction of movie rating and revenue. This is the baseline we use to evaluate how our single modality models (posters and text) or multi-modal models (combining all text,poster and tabular data) perform on the prediction tasks. The baseline model we use is a simple Random Forest Regressor. We perform extensive hyperparameter search and ablation studies on this model to establish a baseline.

The baseline is established on tabular data. Tabular data refers to all features in the extracted data that is not textual (title and synopsis) or image (poster). The features are namely Budget, Runtime, Release Date, Available Countries, Available Languages, Directors, Top 3 production companies, Top 3 Genres and Top 3 cast members.

As evident, the tabular data consists of numerical and categorical features. Dealing with categorical features was one of the first design decisions here as the categorical features like genre, cast etc. are also multi-value (i.e. each movie has 0-3 categorical values for these features). To deal with the multi-value categorical features, we index all unique values for the feature, split the feature into multiple features (for example, 'Top 3 Genres' is split into Genre1, Genre2, and Genre3), and replace the categorical values in these newly created features with the index we created.

Now that all the data is converted into numerical values/indexes, we can input it to random forest regressor. We search through multiple hyperparameters for the regressor (28):

- min_samples_leaf: min number of samples at a leaf node.
- max_features: number of features to be considered
- N_estimators: number of trees in the forest
- max_depth: max depth of a tree in the forest.

After establishing a baseline using all the features (tabular), we move to performing ablation studies to understand the contribution of each feature to the prediction. We also try

Feature	Type	Example
Title	Text	Dedication
Synopsis	Text	A modern love story in which a misanthropic, emotionally complex author of a hit children's book is forced to team with a beautiful illustrator after his best friend and collaborator passes away. As Henry struggles with letting go of the ghosts of love and life, he discovers that sometimes you have to take a gamble at life to find love
Poster	Image	Refer to Fig. 2
Top 3 genres	Categorical	comedy, drama, romance
Top 3 cast members	Categorical	billy crudup, mandy moore, tom wilkinson
Top 3 production companies	Categorical	first look international, hart-lunsford pictures, plum pictures
Director	Categorical	justin theroux
Available Languages	Categorical	English
Original Language	Categorical	English
Available Countries	Categorical	United states
Release Date	Date	2007-01-22
Runtime (minutes)	Numeric	95
Budget (dollars)	Numeric	13,000,000

Table 1. example of data related to a movie: Dedication

using multiple random combinations of features based on the study to come up with the baseline.

Single Modality Models. In this subsection, we discuss the methods used to train a model only text features (synopsis and title) and another model only on the image posters.

Text Model. We use pre-trained DistilBERT (29) to fine tune the model to predict movie rating and revenue. The input to this model is just the movie title and synopsis, concatenated as a string and separated by a [SEP] token. For example, an input might look like this:

"Avengers: Endgame [SEP] After the devastating events of Avengers: Infinity War (2018), the universe is in ruins. With the help of remaining allies, the Avengers assemble once more in order to reverse Thanos' actions and restore balance to the universe"

Image Poster Model. We use two approaches for the prediction models using just posters. The first involves a simple convolutional neural network (30) architecture outlined in Fig. 1.
The second approach involves transfer learning with pre-trained ResNet50 to fine tune the model to predict ratings and revenues (31). We experiment training with both freezing all but the output layers and unfreezing all the layers. The input to both approaches is a movie poster fit to 224 x 224 dimensions. An example input is shown in Fig. 2.

Combining Text and Tabular Data. We decided to combine text and tabular data first before including image posters. There are many ways of combining tabular and text data. In this subsection, we present the different strategies we tested to combine these two modalities.

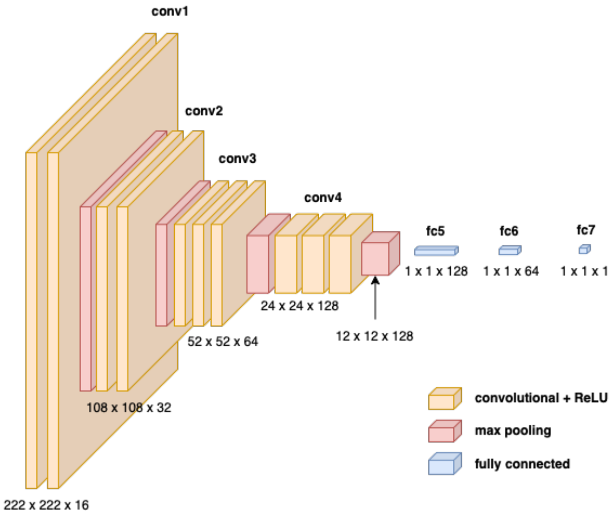


Fig. 1. Architecture for Simple CNN

Fig. 2. Example Movie Poster

The Simplest Way. The most obvious way of combining text and tabular data is to treat everything as text and feed it to DistilBERT. To do this, we take the following steps:

1. Convert all features (e.g., budget, runtime, date) to string

2. Concatenate all features with [SEP] token as a separator
3. Make sure the order of the features remains the same
4. Give it to DistilBERT just like any other text

After doing this, an input to the model could look like this:

"Dedication [SEP] A modern love story in which a misanthropic, emotionally complex author of a hit children's book is forced to team with a beautiful illustrator after his best friend and collaborator passes away. As Henry struggles with letting go of the ghosts of love and life, he discovers that sometimes you have to take a gamble at life to find love. [SEP] billy crudup, mandy moore, tom wilkinson [SEP] united states [SEP] justin theroux [SEP] comedy, drama, romance [SEP] english [SEP] first look international, hart-lunsford pictures, plum pictures [SEP] 13000000 [SEP] 95 [SEP] 2007-01-22"

Multimodal Transformers Inspired Models. We came across a toolkit(32) that can be used to process textual, categorical, and numerical data through transformer models (DistilBERT, Roberta, etc.). It provides various options to combine/convert/manipulate the data before passing it through a model of choice and evaluating (classification/regression) it with a metric of choice.

Instead of using the toolkit directly we decided to use a model inspired by the main model used in the source code of the toolkit. We leverage the idea of combining the data within a model inherited from transformer model (i.e. DistilBERT in our case).

The data we use here is textual data and tabular data, where the categorical features of tabular data are combined with the textual data, using the method mentioned earlier (everything as text separated by [SEP] token), while the numerical data is processed separately. Fig. 3 shows an overview of how the model works.

The transformer model we use to obtain the CLS embedding for the text (textual+categorical) data is DistilBERT. Before we combine the embedding(features) with numerical features, we can pre-process the numerical features by passing them through a hidden layer. Numerical features with/without preprocessing are then to be combined with the CLS embedding. This concatenation could be simple, i.e. concatenation of vectors or other complex variations. We can use weighted sum of features, where weights for all features are learned or gated sum of features, which allows or blocks features selectively.

The concatenated features can be passed through a single output layer or a complex MLP architecture to predict revenue/rating for the movie.

We use multiple configuration options to choose between the available options of preprocessing and concatenation. We also use hyperparameters to evaluate performance by varying the learning rate, batch size, weight decay, and repeats for the model. We choose the model that performs best on the validation set to evaluate the result on test set.

Text and categorical data:
"Dedication [SEP] A modern love story in which a misanthropic, emotionally complex author of a hit children's book is forced to team with a beautiful illustrator after his best friend and collaborator passes away. As Henry struggles with letting go of the ghosts of love and life, he discovers that sometimes you have to take a gamble at life to find love. [SEP] billy crudup, mandy moore, tom wilkinson [SEP] united states [SEP] justin theroux [SEP] comedy, drama, romance [SEP] english [SEP] first look international, hart-lunsford pictures, plum pictures"

Numerical Data:
Budget : 13000000 ; Runtime : 95 ; Year : 2007 ; Month : 01 ; Date : 22

Combining All Modalities. Fig. 4 and Fig.5 show an overview of how the models for combining all modalities work.

Fig. 4 is a representation of how we combine all modalities for predicting ratings. The image is passed through ResNet, which was chosen according to the single modality study we discussed earlier. Transfer learning with pre-trained ResNet performed best for posters and, hence, we use this for extracting image features from posters. As we learned from the experiments done for combining text and tabular data, combining all features as a text works best for rating prediction. We choose to use this as input for DistilBERT model to extract an embedding from these features. We then need to combine the image features and embedding. Before combining the features, we could choose to preprocess the features, passing them through a MLP architecture or perform dimension reduction (feature selection). Finally, after combining the extracted features from all modalities, we proceed to the output layer, which could be a simple linear layer or a multi-layer architecture.

Fig. 5 is a representation of how we combine all modalities for predicting revenues. The difference between the architectures of all-modality revenue and rating prediction models is how we combine text and tabular data. The experiments we performed showed that combining numerical features with textual embedding using the weighted sum of features in comparison instead of treating all features as text as we did for rating model works better for revenue prediction.

Experiments

For all experiments we use DistilBERT as the language model. This decision is purely due to time and resource constraints. We acknowledge that a non-distilled language model (e.g. RoBERTa) could achieve better results, but we assume DistilBERT can give very close results based on the literature. For the vision model, we tested different sizes of ResNet.

All models have the following hyper parameters/settings in common:

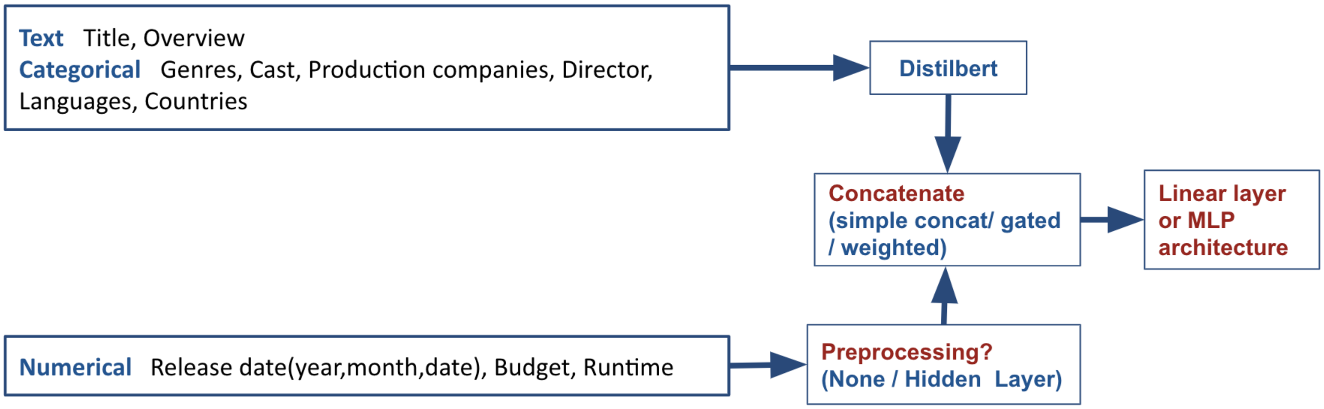
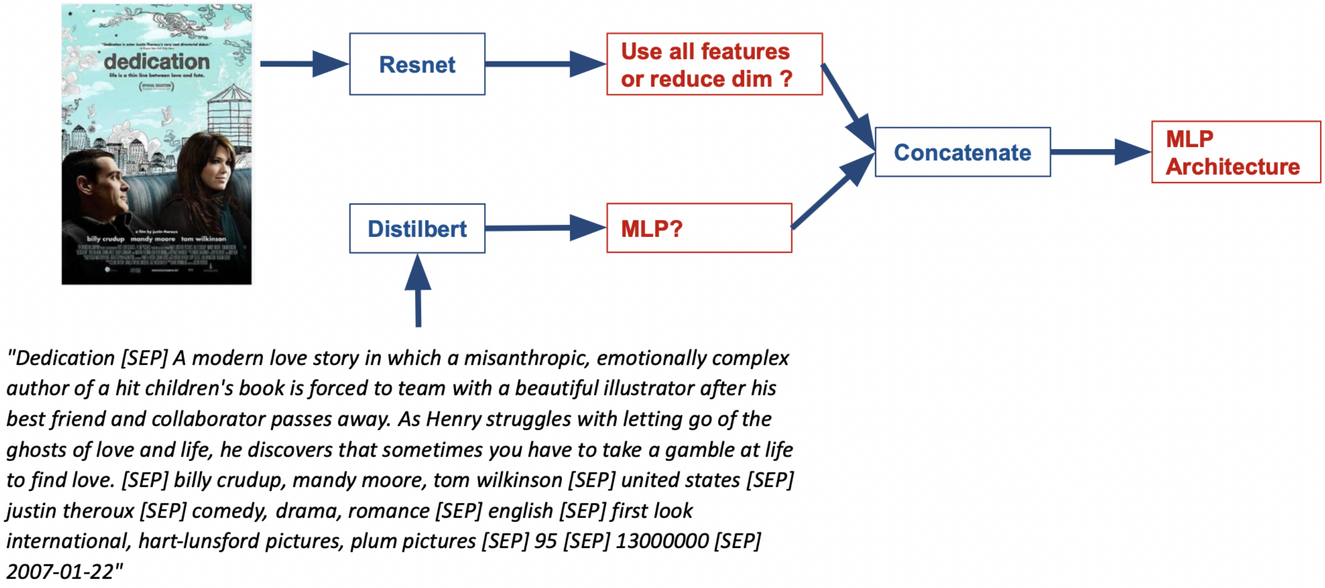
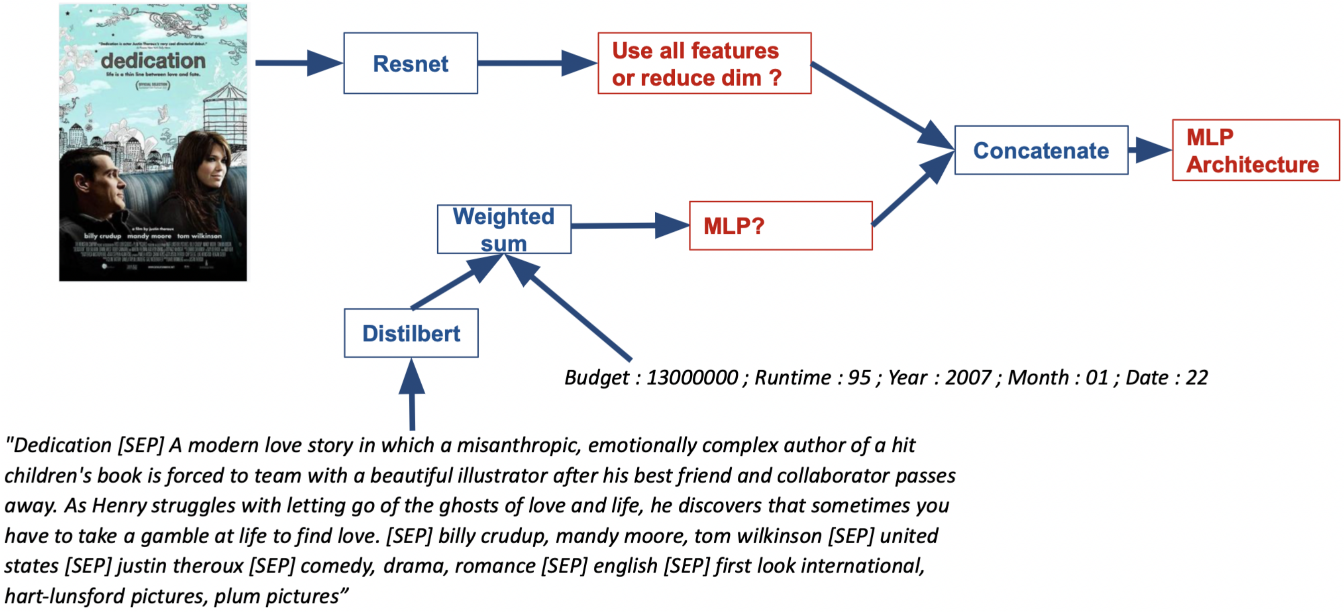

Text Title, Overview
Categorical Genres, Cast, Production companies, Director, Languages, Countries
Distilbert
Concatenate (simple concat/ gated / weighted)
Linear layer or MLP architecture
Numerical Release date(year,month,date), Budget, Runtime
Preprocessing? (None / Hidden Layer)

Fig. 3. Model for combining text and tabular



"Dedication [SEP] A modern love story in which a misanthropic, emotionally complex author of a hit children's book is forced to team with a beautiful illustrator after his best friend and collaborator passes away. As Henry struggles with letting go of the ghosts of love and life, he discovers that sometimes you have to take a gamble at life to find love. [SEP] billy crudup, mandy moore, tom wilkinson [SEP] united states [SEP] justin theroux [SEP] comedy, drama, romance [SEP] english [SEP] first look international, hart-lunsford pictures, plum pictures [SEP] 95 [SEP] 13000000 [SEP] 2007-01-22"

Fig. 4. Model for combined modalities



Budget : 13000000 ; Runtime : 95 ; Year : 2007 ; Month : 01 ; Date : 22

"Dedication [SEP] A modern love story in which a misanthropic, emotionally complex author of a hit children's book is forced to team with a beautiful illustrator after his best friend and collaborator passes away. As Henry struggles with letting go of the ghosts of love and life, he discovers that sometimes you have to take a gamble at life to find love. [SEP] billy crudup, mandy moore, tom wilkinson [SEP] united states [SEP] justin theroux [SEP] comedy, drama, romance [SEP] english [SEP] first look international, hart-lunsford pictures, plum pictures"

Fig. 5. Model for combined modalities

- Max epochs: 15
- Model evaluation every 100 steps
- Early stopping with a patience of 30 evaluation steps.
- Linear learning rate decay with warmup ratio = 0.06
- seed = 42, the magic number
- GPU: Nvidia T4 with FP16 Mixed Precision

We mainly follow RoBERTa (33) grid search for fine-tuning but do not limit ourselves that. Specifically, we use the coarse-to-fine methodology: if after the search the best trial has a hyperparameter at a boundary of the search space, we explore higher or smaller values depending on the situation. For example, if the learning rate of the best model is 5e-5 and that is highest learning rate we tested, then we search for higher learning rates (e.g., 7e-5, 1e-4). Therefore here we just specify the parts that deviate from this strategy.

Baseline. We use Random Forest Regressor from the Scikit-learn library to build a baseline model. The model uses tabular data and predicts either the rating or revenue of a movie. While predicting revenue, we use the logarithm of the revenue to make evaluation of the model easier and deal with skewed data. We tune the model using various hyperparameters:

- min_samples_leaf: [1,2,4,8,16]
- max_features: ['auto', 'sqrt', 'log2']
- n_estimators: [50,100,150,200,250,300]
- max_depth: [10,12,14,16,18,20]

Once we have our best model after tuning the above mentioned hyperparameters, we perform ablation studies. For these, we remove a feature from the set of inputs and determine the degradation in performance, which reflects the importance of the feature. We plot a graph to understand how much each feature contributes to the performance of prediction. The higher the degradation, more the important the feature is to the prediction task.

Single Modality Models. Now, we discuss the experiments we perform with each data type separately. We aim to answer how well we can predict the ratings and revenues of movies by using just text data (title and synposis) alone and using posters alone.

Image Poster Model. With only using the movie posters, we run several experiments to see how well we can do with just image data. We first establish a very rudimentary benchmark for comparison, taking the average of all the ratings and revenues and using those values as the predictions for all the test data. While this is imperfect for revenue as the data is not normally distributed, we use it as a starting point for the experiments.

Throughout the experiments with the posters, we tune the models using various hyperparameters in a grid search manner. The values used are as follows:

- batch_size: [8, 16, 32]
- learning_rate: [1e-5, 1e-4, 1e-3]
- weight_decay: [0.01, 0.1]

We first test a simple CNN model to see how well it performs. The architecture of this model is depicted in Fig. 1.

To improve on the performance of the simple CNN, we perform several experiments using transfer learning. Initially, we replace the output layer of ResNet50 with a new output layer for the current problem and take an educated guess at the hyperparameters and test the model (31). We freeze all the layers except the output layer, train the model, and then unfreeze all the layers and further fine tune the model.

In the next experiments, we perform transfer learning by directly unfreezing all the layers and making them trainable instead of the two step approach. We also tune hyperparameters with this approach.

We then proceed to combining the various data modalities and models to try and improve the performance of our models.

Combining text and tabular data.

Multi modal Transformers Inspired Models. We experiment with a modified transformer that includes tabular data along with text data. Before we use numerical data, it is important to normalize the data as we use features that work at very different scales and are also skewed. We split the 'date' feature into year, month, and date features.

We also try to leverage the cyclic nature of dates by using sine and cosine transformations to better represent the cyclical nature of the features date and month. However, this does not significantly improve the performance.

We also try using MLP preprocessing for the numerical data before combining it with the textual embedding. The embedding and the numerical data are then combined using one of the three methods: simple concatenation, gated sum of features, and weighted sum of features.

After our features are combined, we can choose the complexity of our output layer, which could be either of two: a simple linear classifier or a MLP with a chosen number of hidden layers.

All these experiments were performed while also tuning hyperparameters for DistilBERT. For each different combination, we identified the best performing hyperparameter settings on the validation set to use on the test set.

Combining All Modalities. Aside from typical DistilBERT hyperparameters that we have discussed, we also treat the following decisions as hyperparameters: (1) whether to augment the poster images, (2) the amount of dimensionality reduction of the ResNet features with an additional hidden layer, (3) the amount of dropout for the image features, and (4) whether to concatenate features with the CLS token embedding first and then pass that through an MLP or pass the embedding through an MLP, concatenate, and then run another MLP?

Results and Analysis

We begin with establishing a baseline for the task as mentioned earlier. We perform hyperparameter tuning and ablation studies. The hyperparameter setting that gives the best result on the baseline of both rating and revenue prediction are a min_samples_leaf of *1*, a max_features of *'auto'*, a max_depth of *10*, and a random_state of *0*.

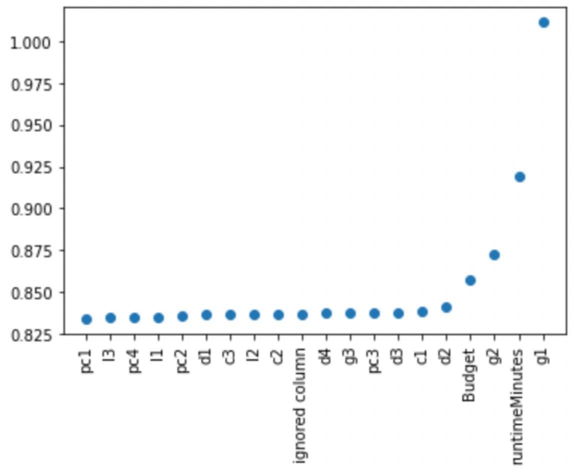


Fig. 6. Ablation study for rating

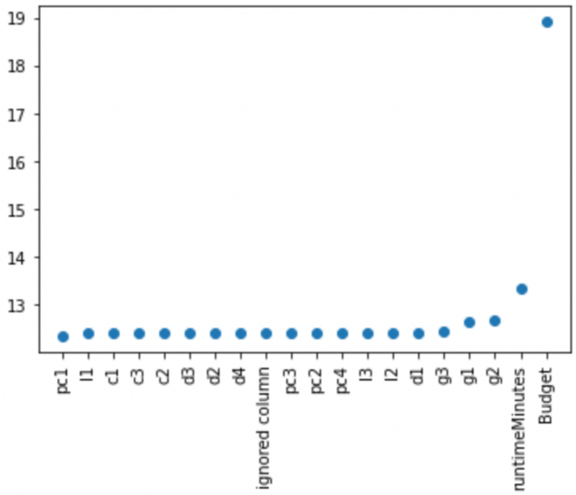


Fig. 7. Ablation study for revenue

The results of the ablation study performed during baseline model determination show that genre is an important feature for rating, as shown in Fig. 6, while budget is an important feature for revenue, as shown in Fig. 7. On x axis are the features that are removed one by one, and y axis represents the error - MSE(rating)/MAE(log(revenue)). The higher the error on removing a feature from input, more relevant the feature is to the prediction task.

Results for baseline experiments were as given in Table 2 for rating and 3 for revenue.

The results for the rest of the experiments are summarized in Table 4 for the prediction of ratings and Table 5 for the prediction of revenues.

Features used	MSE
Using all features	0.920
Using Budget, Genre, Prod. Studios	0.736

Table 2. Results of experiments for baseline of rating prediction

Features used	MAE
Using all features	44370461.7
Using Budget, Genre, Prod. Studios	44362493.63

Table 3. Results of experiments for baseline of revenue prediction

For the experiments with just using the posters, the simple CNN performed the same as average prediction for ratings, but improved for revenues. The two step transfer learning with pre-trained ResNet50 performed better than the simple CNN for ratings, but worse for revenues. After hyperparameter tuning and by directly training all the layers of pre-trained ResNet50, the best model performed significantly better than average prediction, but not as well as with using other data, as seen in Tables 4 and 5. We posit that poster data alone is not sufficient to predict these metrics as audiences are influenced by many more factors that must be accounted for. We also see that posters are more predictive of revenue than rating, which is expected as ratings depend on the reception, word-of-mouth, and consensus of viewers well into the post-release period, whereas revenue is largely accounted for within the opening weekend and driven by advertising and promotions, of which posters are a significant component.

The best model we got for rating prediction turned out to be one of the simplest, which is the Text and Tabular as Text. However, the best model for revenue is the Text and Tabular Weighted sum, and by a large margin. Finally, using image posters did not improve either of our targets.

The under performance of the multimodal models that include images can be explained by multiple reasons including:

- It is possible that the image posters are not good indicators of movie rating or revenue, and adding this feature to the model mostly introduces noise.
- We are using late fusion, but it is possible that another fusion method (e.g, early fusion with a pre-trained visual-language model) could have achieved a better performance.

In addition, our initial hypothesis was that the Text with Tabular as Text model was going to be an easy baseline to surpass. We were proven wrong as it turns out this is highly dependent on what we are trying to predict.

Conclusion

We have performed a range of experiments and analysed our results in the previous section. In this section, we summarise a few major findings or unexpected results. While performing experiments for unimodal models (text and images), we

Model Name	Modalities	MSE
Simple CNN with Posters	Image	1.168
Transfer Learning using ResNet50 with Posters	Image	1.106
Only Text	Text	0.901
Text with Tabular as Text	**Text, Tabular**	**0.626**
Text and Tabular Concatenation	Text, Tabular	0.686
Text and Tabular Weighted Sum	Text, Tabular	0.6836
Text and Tabular Gated Sum	Text, Tabular	0.68
All Modalities	Text, Tabular, Image	0.726

Table 4. Results of experiments for rating prediction

Model Name	Modalities	MAE (USD)
Simple CNN with Posters	Image	43,873,931
Transfer Learning using ResNet50 with Posters	Image	43,135,964
Only Text	Text	43,634,880
Text with Tabular as Text	Text, Tabular	34,272,932
Text and Tabular Concatenation	Text, Tabular	27,890,194
Text and Tabular Weighted Sum	**Text, Tabular**	**27,749,394**
Text and Tabular Gated Sum	Text, Tabular	31,239,150
All Modalities	Text, Tabular, Image	54,888,430

Table 5. Results of experiments for revenue prediction

realized that using pre-trained models significantly improved performance. The image models performed well for revenue but not for rating, understandably so as posters are known to boost revenue more than rating (involves critics of the movie). While combining text and tabular data, to our surprise, combining everything as text gave better results than using numerical features separately (for rating). For revenue though, combining numerical separately performs slightly better perhaps because numerical features are relevant to revenue. Finally, while combining all modalities, we realized that the late fusion of modalities is not a trivial task and leads us to several open ended questions. We have tried to perform trials using different manipulations possible. However, the project has scope for further experimentation. Nonetheless, we have shown that a multimodal approach to predicting movie ratings and revenues achieves significant improvement over existing models and benefits the decision support systems of movie stakeholders.

Limitations / Future Work

We recognize that our work is not without limitations and would like to acknowledge those. First, the data we used was limited to English language movies released after 1980. However, movies are made in numerous languages around the world and have a global reach in the modern era and including data about those movies would make the model more robust. Another limitation is that we chose features based on the available data and manual analysis. The model could be improved by using deep learning to automatically learn features from raw data by stacking several nonlinear modules (34). Lastly, we evaluated our models on movies already released, separated into a test set, as we were limited on time. A better approach would be to take movies that are going to release over the next few months and evaluate the models on those movies' post-release data.

This work can take several directions to reach a more robust and predictive model. As previously mentioned, an expanded dataset with automated feature extraction and the use of unreleased movies for testing is the first natural step. Further, explanatory analysis can be performed to understand how the models are working to achieve the predictions and identify the features that are most important for predicting the success of a movie. Prior research has shown that the number of screens, high technical effects, and high star value are major contributors to the prediction of the success of a movie and further work in the determination of contributing variables can be valuable to filmmakers (10). Lastly, the work can be expanded to implement multi-task learning and combine rating and revenue prediction into the same model. We would also try an early fusion method with a pre-trained vision language model. These steps can increase the usefulness of the work for stakeholders and enable exploration of its applications to other media product demand forecasting.

Contributions

Shikha Asrani: Random Forest Regression Model, Ablation study on tabular data, Multimodal Transformer Inspired Model, Combining All Modalities Models (with Samir Char)

Samir Char: Data Collection and Preprocessing, Exploratory Data Analysis, Only Text Models, Text with Tabular as Text Model, Combining All Modalities Models (with Shikha Asrani)

Srujan Esanakarra: Simple CNN Model, Transfer Learning Models

References

1. Arthur S De Vany and W.David Walls. Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. *Journal of Economic Dynamics and Control*, 28(6):1035–1057, 2004. ISSN 0165-1889. doi: https://doi.org/10.1016/S0165-1889(03)00065-4.
2. Thorsten Hennig-Thurau, Mark B. Houston, and Gianfranco Walsh. Determinants of motion picture box office and profitability: An Interrelationship approach. *Review of Managerial Science*, 1(1):65–92, 2007. doi: 10.1007/s11846-007-0003-9.
3. Harold L Vogel. *Entertainment industry economics: A guide for financial analysis*. Cambridge University Press, 2020.
4. David A Edwards. Platonism is the law of the Land. *Notices of the AMS*, 60(4):475–478, 2013.
5. Liran Einav. Seasonality in the US motion picture industry. *The Rand journal of economics*, 38(1):127–145, 2007.
6. Barry R. Litman and Linda S. Kohl. Predicting financial success of motion pictures: The '80s experience. *Journal of Media Economics*, 2(2):35–50, 1989. doi: 10.1080/08997768909358184.
7. Jack Valenti. *Motion pictures and their impact on society in the year 2001*. Midwest Research Institute, 1978.
8. Movie market summary for year 2019.
9. Anita Elberse and Jehoshua Eliashberg. The drivers of motion picture performance: the need to consider dynamics, endogeneity and simultaneity. In *Proceedings of the Business and Economic Scholars Workshop in Motion Picture Industry Studies. Florida Atlantic University*, 2002.
10. Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.*, 30(2):243–254, 2006. doi: 10.1016/j.eswa.2005.07.018.
11. P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N H Shwetha. An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In *International Conference on Computing, Communication Automation*, pages 933–937, 2015. doi: 10.1109/CCAA.2015.7148530.
12. Krushikanth R. Apala, Merin Jose, Supreme Motnam, C.-C. Chan, Kathy J. Liszka, and Federico de Gregorio. Prediction of movies box office performance using social media. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 1209–1214, 2013. doi: 10.1145/2492517.2500232.
13. Yao Zhou, Lei Zhang, and Zhang Yi. Predicting movie box-office revenues using deep neural networks. *Neural Comput. Appl.*, 31(6):1855–1865, 2019. doi: 10.1007/s00521-017-3162-x.
14. Márton Mestyán, Taha Yasseri, and János Kertész. Early Prediction of Movie Box Office Success based on Wikipedia Activity - Big Data. *CoRR*, abs/1211.0970, 2012.
15. IMDb: Ratings, Reviews, and Where to Watch the Best Movies TV Shows. https://www.imdb.com/.
16. Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. Movie posters classification into genres based on low-level features. In *2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1198–1203. IEEE, 2014.
17. Sharmistha Dey. Predicting Gross Movie Revenue. *arXiv preprint arXiv:1804.03565*, 2018.
18. Kevin Chen. Predicting IMDb Ratings of New Movies, May 2021.
19. Ryan Anderson. What makes a successful film? Predicting a film's revenue and user rating with machine learning, Aug 2019.
20. Adarsh Tadimari, Naveen Kumar, Tanaya Guha, and Shrikanth S Narayanan. Opening big in box office? Trailer content can help. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2777–2781. IEEE, 2016.
21. Taegu Kim, Jungsik Hong, and Pilsung Kang. Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting*, 31(2):364–390, 2015. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2014.05.006.
22. Mahsa Shafaei, Adrián Pastor López-Monroy, and Thamar Solorio. Exploiting Textual, Visual, and Product Features for Predicting the Likeability of Movies. In Roman Barták and Keith W. Brawner, editors, *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019*, pages 215–220. AAAI Press, 2019.
23. Yichen Yang, Ruoyun Ma, and Min Haeng Cho. Predicting Movie Ratings with Multimodal Data.
24. Banik Rounak. The Movie Dataset: Metadata on over 45,000 movies. 26 million ratings from over 270,000 users., 2017.
25. Cinemagoer: Python package for retrieving and managing the data of the IMDb movie database about movies and people. https://cinemagoer.github.io/, 2022.
26. Box Office Mojo by IMDb Pro – an IMDb company. https://www.boxofficemojo.com/.
27. Avatar Box Office data. https://www.boxofficemojo.com/release/rl876971521/, 2009.
28. Random Forest Regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
29. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.
30. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.
31. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.
32. Multi Modal Transformer. https://github.com/georgian-io/Multimodal-Toolkit.
33. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
34. Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nat.*, 521(7553): 436–444, 2015. doi: 10.1038/nature14539.