Movie Reviews Sentiment Prediction and Topic Modeling

Group Members: Yingying Qian, Xiao Xiao, Qingnan Wang, Hanchen Liu, Runfeng Zhang

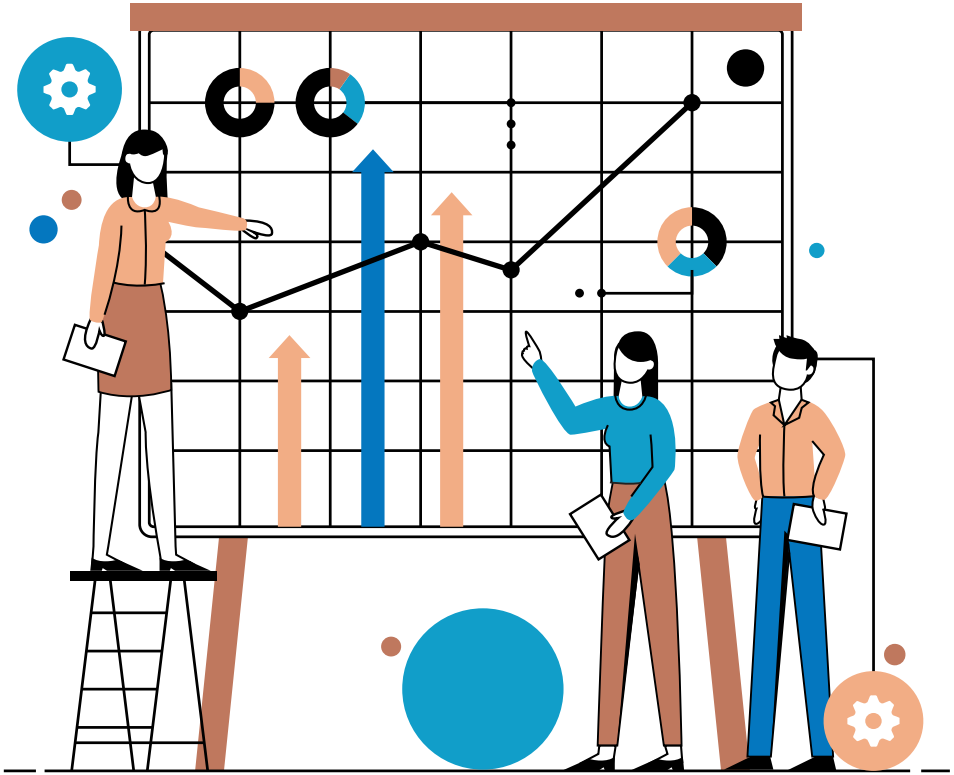
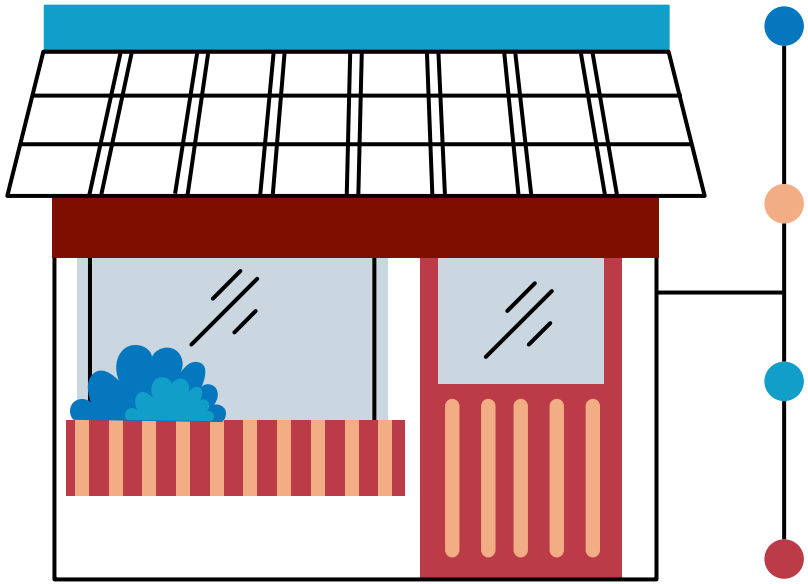


Table of Contents



Problem Statement

Data Scientist Team at Warner Brothers

Task 1

WB has released some movies recently. However, currently we do not have enough movie reviews on the movies review website so we are not sure about how our audience react to the movie. However, there are a lot of reviews posted on the social media (Twitter, Facebook, etc.) and we have already scrapped these reviews. We want to DS team to help us build models to predict audience sentiment.

Task 2

Topic models can help us analyze the movies that are discussed by our audience. By doing topic modeling, we know the hot topics about movies, which are useful to our marketing strategies. By analyzing these topic, we can make better decision on future movie topics. Besides, we can classify the reviews by different topics. Then users can easily find the reviews of the topics that they are interested in.

Problem Statement

Datasets we use: IMDB dataset having 50K movie reviews. It contains 25,000 highly polar movie reviews for training and 25,000 for testing, which is a good dataset for building sentiment classification models.

Sentiment Prediction

By training classification models using the IMDB movie reviews dataset, we can predict the underlying sentiment of movies reviews collected from social media

Topic Modeling

Using the IMDB movie reviews, we can find out different topics these reviews cover and group them by those topics. Then, do some analysis to make decision on future movie topics.

Sentiment Prediction: Logistic Regression

Data Preparation

- Remove HTML tags and punctuation using Regex
- Stemming

Model Building

- Tokenized text using both Count Vectorizer, TFIDF Vectorizer and Spacy
- Split training (37500 reviews) and testing dataset (12500 reviews)
- Train Logistic Regression model
- Accuracy on testing set:

Count Vectorizer	TFIDF Vectorizer	Spacy
85.2%	85.2%	85.6%

Sentiment Prediction: RNN & LSTM & Hugging Face

RNN & LSTM

- Stopwords removal using Spacy
- Tokenized Text using tokenizer by keras.preprocessing.text, num_words=5000
- Encoded documents and padded sequence length
  - max sequence length = 150, decided using histogram and percentiles statistics
- Train test split: test = 0.2
- Used glove6b100d vectors

Hugging Face

- Tokenized Text using BertTokenizerFast
- Predict sentiment with a pretrained model ‘sentiment-analysis’ using pipeline

LSTM	RNN	Hugging Face
87.4%	67.6%	82%

Model Implementation

Model Summary

LR using Count Vectorizer	85.2%
LR using TFIDF Vectorizer	85.2%
LR using Spacy	85.6%
LSTM	**87.4%**
RNN	67.6%
Hugging Face	82%

How we implement this model in the future

We could find that LSTM achieved highest accuracy. In the future, for all movies that have released for only a small period of time, we can scrape reviews from social media, and then use the preprocessing pipeline to clean these reviews, and finally apply the pre-trained LSTM model to predict sentiment of each review. We expect this model can classify sentiment of reviews with more than 85% accuracy. In addition, we will re-train this model monthly using more reviews collected in hope to increase prediction accuracy.

Business Use Case

Application of Sentiment Prediction

Marketing/Advertising Strategies

Adjust marketing strategies of each movie timely based on early stage review sentiment.

Media Monitoring

In the future, we can build an automate media monitoring process to track audience reaction to the movie overtime, and monitor mentions or reviews of the movies on different platforms.

Marketing Budget Optimization

Average cost of a movie

$ 65 million Production Cost

$ 35 million Marketing and distribution Cost

By predicting sentiment at the early stage of a movie release, we can adjust and optimize our marketing spendings based on audience reactions, and thus generate a higher return on advertising budget.

Topic Modeling

Step 1: Add New Stopwords

Step 2: Vectorize the Corpus

Step 3: Fit NMF Model

Step 4: Report Results for Each Topic

Step 5: Get the Top Documents for Each Topic

Step 6: WordCloud for Each Topic

Our goal is to find out the topics of the reviews. So we need to remove words about sentiments. We also need to remove some verbs and adverbs that has no relationship to topics of reviews.
So , we update stopwords here.

Examples:

Words about sentiments: `"worst","bad","good", "better", "like", "great", "best", "ever", "seen", "waste", "wast", "time", "money", "well", "worth", "recommend"...`

Adverbs and conjunctions: `"really", "even", "though", "although", "highly"...`

Topic Modeling

Step 1: Add New Stopwords

Step 2: Vectorize the Corpus

Step 3: Fit NMF Model

Step 4: Report Results for Each Topic

Step 5: Get the Top Documents for Each Topic

Step 6: WordCloud for Each Topic

Then we vectorize the corpus using TFIDF.

- We tried ngram=2 and ngram=3 and we found ngram=3 can bring us more meaningful results.
- Although infrequent words are useful for us to analyze the texts, they may be less useful in topic modeling because we want to find topics in common. “min_df=10” can help us keep the tokens appear at least 10 times.
- “max_df=0.4” help us get rid of common words appear in too many (more than 40% of) documents.

Topic Modeling

Step 1: Add New Stopwords

Step 2: Vectorize the Corpus

Step 3: Fit NMF Model

Step 4: Report Results for Each Topic

Step 5: Get the Top Documents for Each Topic

Step 6: WordCloud for Each Topic

We tried different numbers of topics and found 5 is reasonable. We get the top documents for each topic to better explore the 5 topics.
Here are top tokens for each topic:

TOPIC 0
new york citi (58.9%)
live new york (1.8%)
street new york (1.6%)
set new york (1.0%)
citi new york (1.0%)

TOPIC 1
world war ii (64.4%)
post world war (1.9%)
save privat ryan (1.3%)
kristin scott thoma (1.3%)
sit back enjoy (0.8%)

TOPIC 2
base true stori (69.7%)
stori base true (1.1%)
real life stori (0.7%) top notch perform (0.6%)
dream come true (0.6%)

TOPIC 3
sci fi channel (53.6%)
fi channel origin (3.2%)
john rhi davi (1.8%)
made sci fi (1.7%)
horror sci fi (1.7%)

TOPIC 4
texa chainsaw massacr (21.2%)
low budget horror (14.8%)
blair witch project (7.0%)
night live dead (2.6%)
budget horror flick (2.0%)

Topic Modeling

Step 1: Add New Stopwords

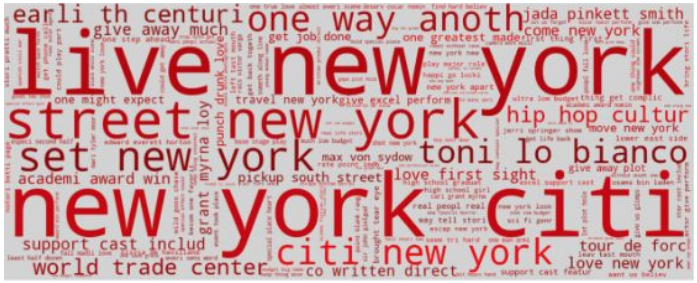
Step 2: Vectorize the Corpus

Step 3: Fit NMF Model

Step 4: Report Results for Each Topic

Step 5: Get the Top Documents for Each Topic

Step 6: WordCloud for Each Topic



Topic 0:
The reviews in topic one are those about New York. Besides tokens of NY and NYC, we can also find something like “World Trade Center” that are related to NYC.

Topic 1:
This topic is about Wars, especially world war II. We can also see some actors’/actresses’ names such as Kristin Scott Thomas, who appeared in movies about WWII (eg. *Suite Française*).

Topic Modeling

Step 1: Add New Stopwords

Step 2: Vectorize the Corpus

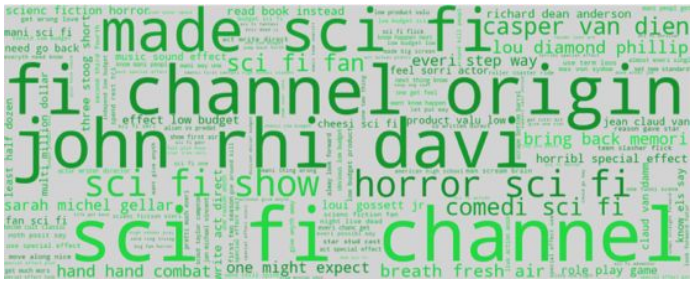
Step 3: Fit NMF Model

Step 4: Report Results for Each Topic

Step 5: Get the Top Documents for Each Topic

Step 6: WordCloud for Each Topic

Topic 2:
This topic is about the movies based on true stories. We can also find tokens about the characters such as “two young men” and “high school student” in the stores.



Topic 3:
This topic is about fictions, especially science fictions. We can also see a name – John Rhys-Davies, who portrayed Gimli in *The Lord of the Rings* trilogy.

Topic Modeling

Step 1: Add New Stopwords

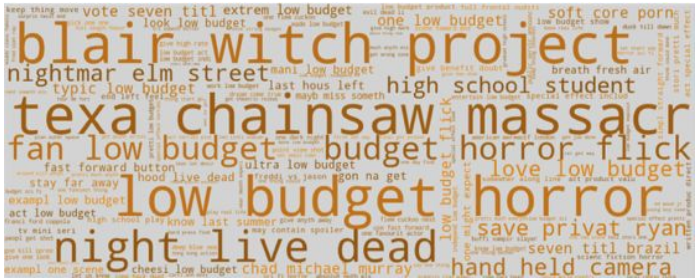
Step 2: Vectorize the Corpus

Step 3: Fit NMF Model

Step 4: Report Results for Each Topic

Step 5: Get the Top Documents for Each Topic

Step 6: WordCloud for Each Topic



Topic 4:

This topic is about horror movies. We can see names of horror movies like *Texas Chainsaw Massacre* and *The Blair Witch Project*.

Many horror movies are low-budget movies so we can see many tokens about budget here.

Besides, words that are related to horror movies also appear, such as “nightmare”.

Return on investment

Cost:

- Assume an experienced data scientist takes a total of 1 month of time on this model (including model building, modeling tuning, building and maintaining pipelines, etc)
- Assume his annual salary is $180,000 per year. That's $15k per month
- Assume other cost such as HR support, hiring cost, etc. totaled 1k per month

Benefit

- The result of this project can help movie producer to target the right audience better and utilize their advertising budget better
- Average movie box office pre-covid is $12,426,863 (in 2019)
- If the result can increase 1% increase in box office revenue, the producer gains $124,268 of extra evenue

Net

- 108k of net gain
- The return on investment is 675%