Movie Genre Classification

X

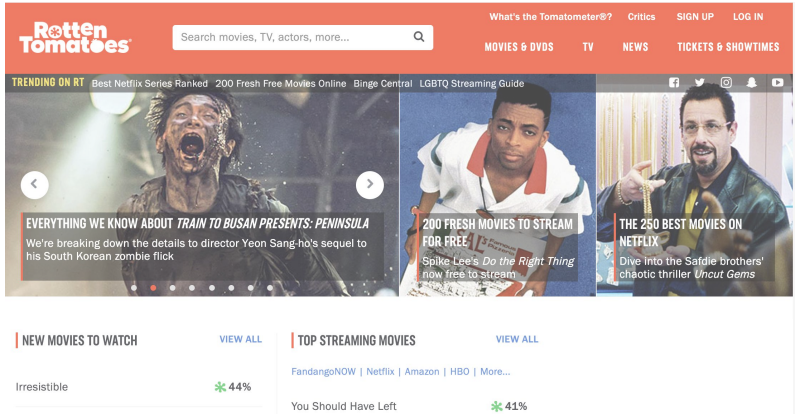
By: Pallavi Chhaparwal, Wong Yong Tian, Desmond Poo, Dominic Teo

Rotten Tomatoes®

Rotten Tomatoes and the Tomatometer score are the world's most trusted recommendation resources for quality entertainment. As the leading online aggregator of movie and TV show reviews from critics, they provide fans with a comprehensive guide to what's Fresh – and what's Rotten – in theaters and at home.

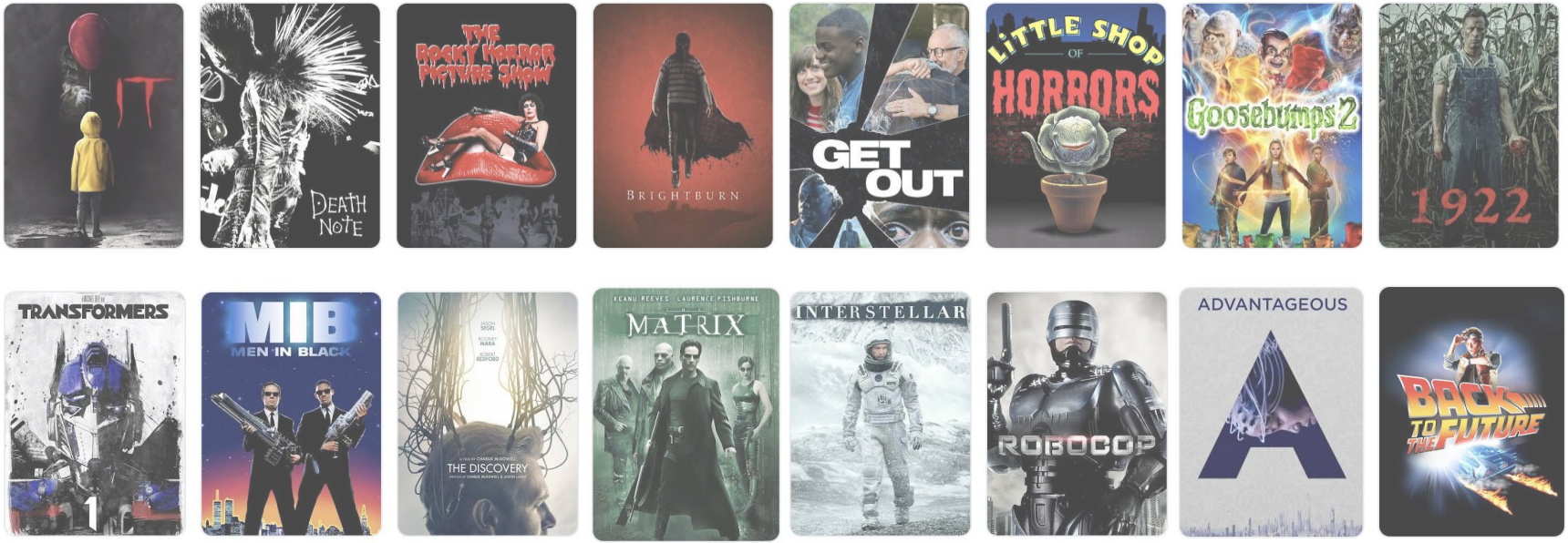
Lately, feedback from customers on not being able to find or match certain titles to respective genres led the Rotten tomato to discover some inaccuracy in existing movie genre classification process.

To improve on genre accuracy, the Management at Rotten Tomatoes has tasked an exploratory team to build a pilot model focusing on two popular genres: **Science Fiction** and **Horror** using Reddit as a data source.



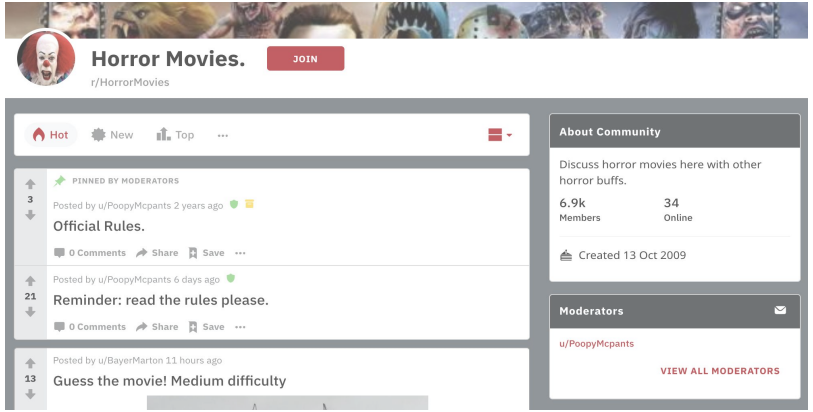
Problem Statement

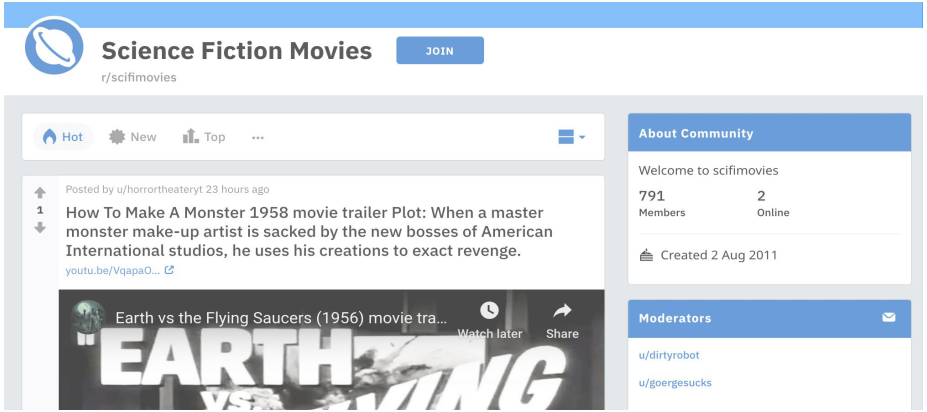
How can we predict a Genre using keywords from subreddit posts?



Rotten Tomatoes®

Subreddits





/r/HorrorMovies

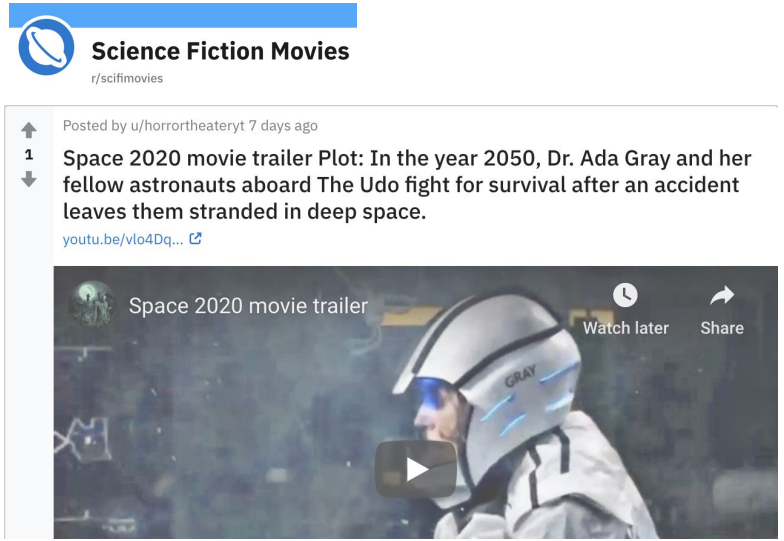
- **892** post collected
- < 10 new post a day
- No advertisements
- Text and media mix
- 6.9k members
- Has Official Rules
- High member interaction

/r/scifimovies

- **345** post collected
- < 10 new post a day
- No advertisements
- Text and media mix
- 791 members
- No Official Rules
- Low member interaction

Rotten Tomatoes®

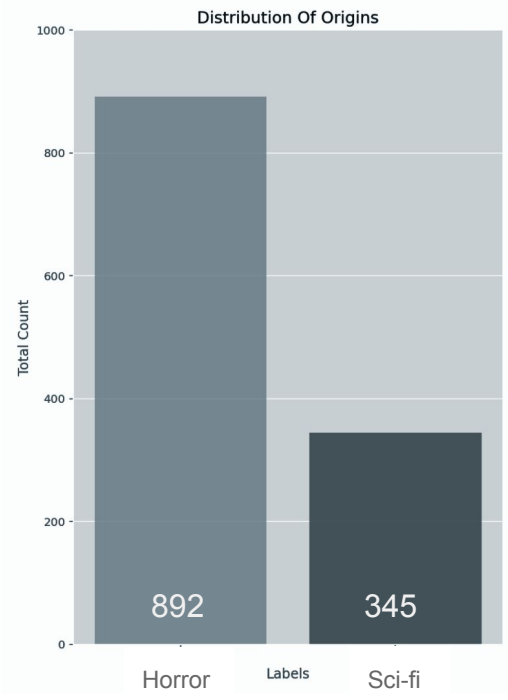
Subreddits

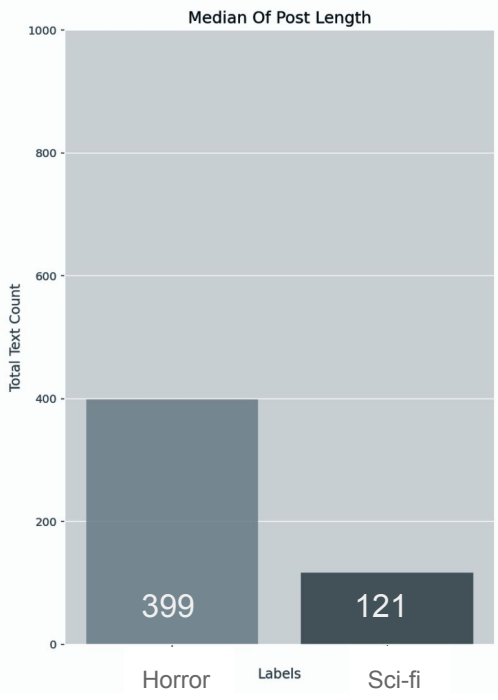


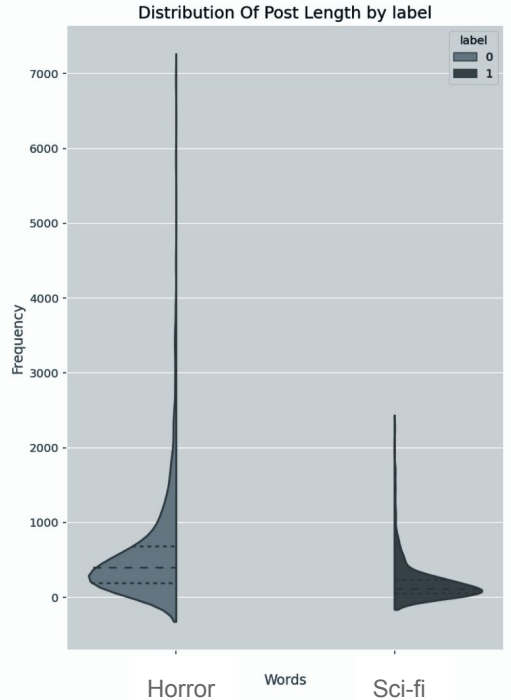
Visually inspecting our subreddits:

- High members and high interactions can explain the number of post collected from Horror
- We also found that Sci-fi post are more condensed, plot related and have fewer noise

Subreddits


Distribution Of Origins
Total Count
1000
800
600
400
200
0
892
345
Horror
Labels
Sci-fi


Median Of Post Length
Total Text Count
1000
800
600
400
200
0
399
121
Horror
Labels
Sci-fi


Distribution Of Post Length by label
label
0
1
Frequency
7000
6000
5000
4000
3000
2000
1000
0
Horror
Words
Sci-fi

Our Data

There is a total of 1238 post.

893 belonging to Horror and 345 belonging to Sci-fi.

Issue: Imbalanced data

How we can overcome this:

- We can mediate this by picking the right models : Logistic Regression, Naïve Bayes Classifiers
- Scoring methods which are favourable towards imbalanced data :
  Sensitivity-Specificity Metrics, Precision-Recall Metrics, ROC_AUC, F1
- Others : Stratified Cross Validation

Rotten Tomatoes

Our Baseline

1	0.720032
0	0.279968

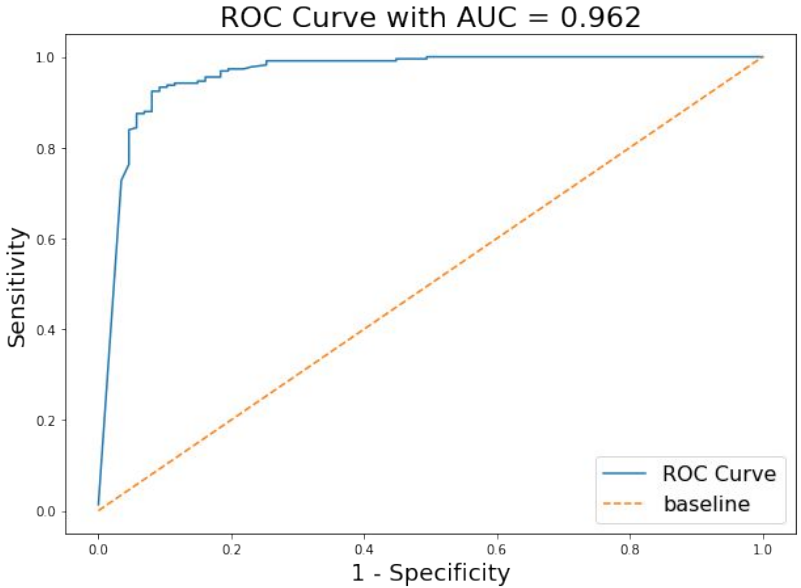
- Baseline model
  - 1 represent Horror movie
  - 0 represent Sci-fi movie
- Our baseline model accuracy is **72.0%** for predicting Horror movie posts.

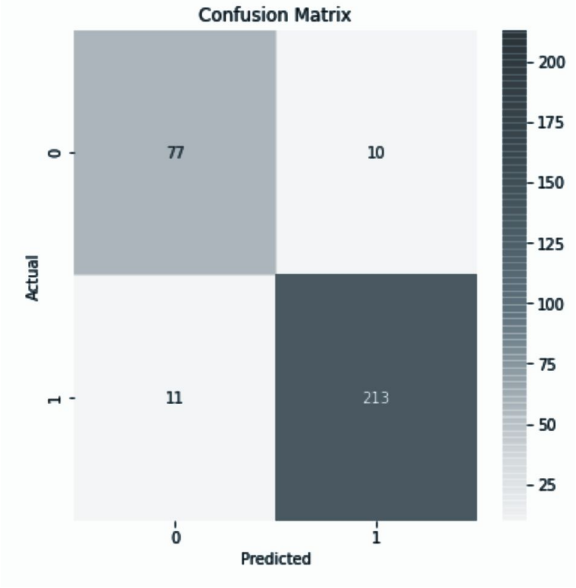
Our Models

- Tested 4 models:
  - **Logistic Regression** vs **Multinomial Naive Bayes**
  - **Count Vectorizer** vs **TFIDF Vectorizer**
- Evaluated Sensitivity, Specificity, Precision and ROC AUC

Classifier	Logistic Regression with CountVectorizer	Logistic Regression with TFIDFVectorizer	Multinomial NB with CountVectorizer	Multinomial NB with TFIDFVectorizer
Sensitivity	93.8	98.2	99.1	99.1
Specificity	88.5	77.0	67.8	55.2
Precision	95.5	91.7	88.8	85.1
ROC AUC	96.2	97.4	97.3	95.5

Our Final Model





- **Logistic Regression model with Count Vectorizer** performed best overall
- Best Specificity and Precision scores
- High Sensitivity and ROC AUC

Rotten Tomatoes®

Our Words

- Best predictive words for Sci-fi movies:
  - **"Sci", "fi", "robot", "alien", "war", "ship", "earth", "scifi", "star", "arrival", "planet", "space" and "clone"**

- Best predictive words for Horror movies:
  - **"Horror", "scary", and "haunting"**

Word	Coefficient
fi	-1.609759
sci	-1.609759
trailer	-1.402131
robot	-0.942322
alien	-0.931292
war	-0.917861
plot	-0.905834
review	-0.874224
ship	-0.844599
earth	-0.834473
scifi	-0.809075
else	-0.771269
star	-0.732366
arrival	-0.721598
planet	-0.683273
short	-0.682440
space	-0.597077
clone	-0.568945
two	-0.553655
possibly	-0.544175

Word	Coefficient
day	0.622345
worth	0.626837
looking	0.630703
next	0.645085
little	0.653564
point	0.701007
best	0.725860
haunting	0.739609
opinion	0.790999
favorite	0.811128
guess	0.841198
thought	0.944435
scary	0.967569
good	0.987151
guy	0.991624
watch	1.007974
anyone	1.054510
watched	1.103585
seen	1.321873
horror	2.369752

Rotten Tomatoes®

Our Words

Horror Movies

SciFi Movies

Limitations

- Insufficient data
  - small sample
  - unaccounted words

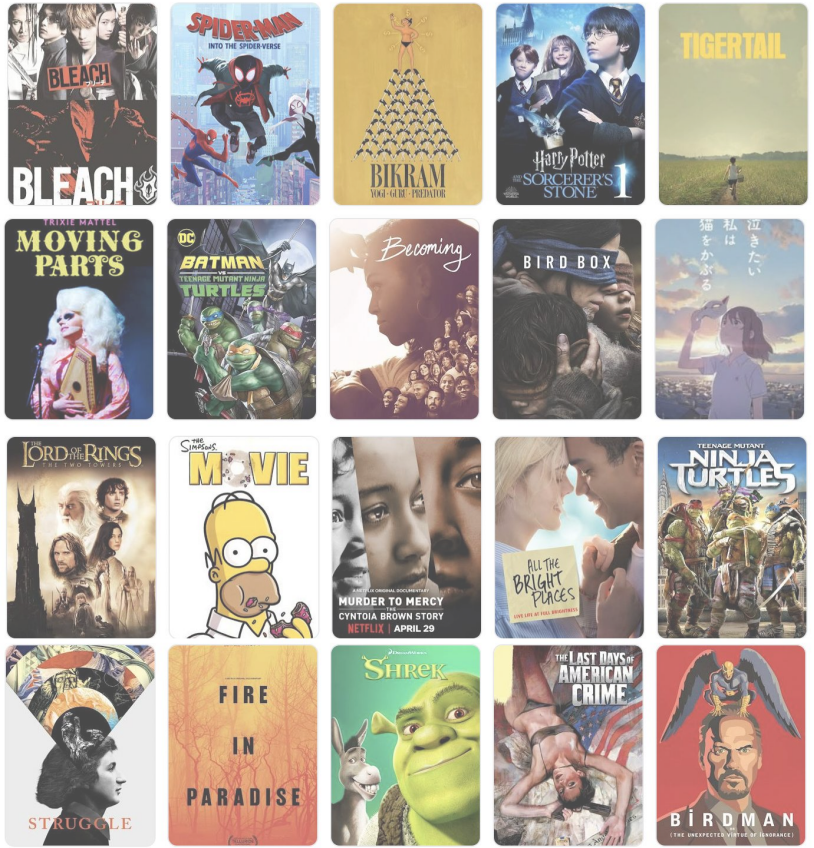
- Only classifies two genres

Recommendations & next step

1. Obtain more data: Sci-fi and Horror related, and others

2. Train a model with multiple target variables

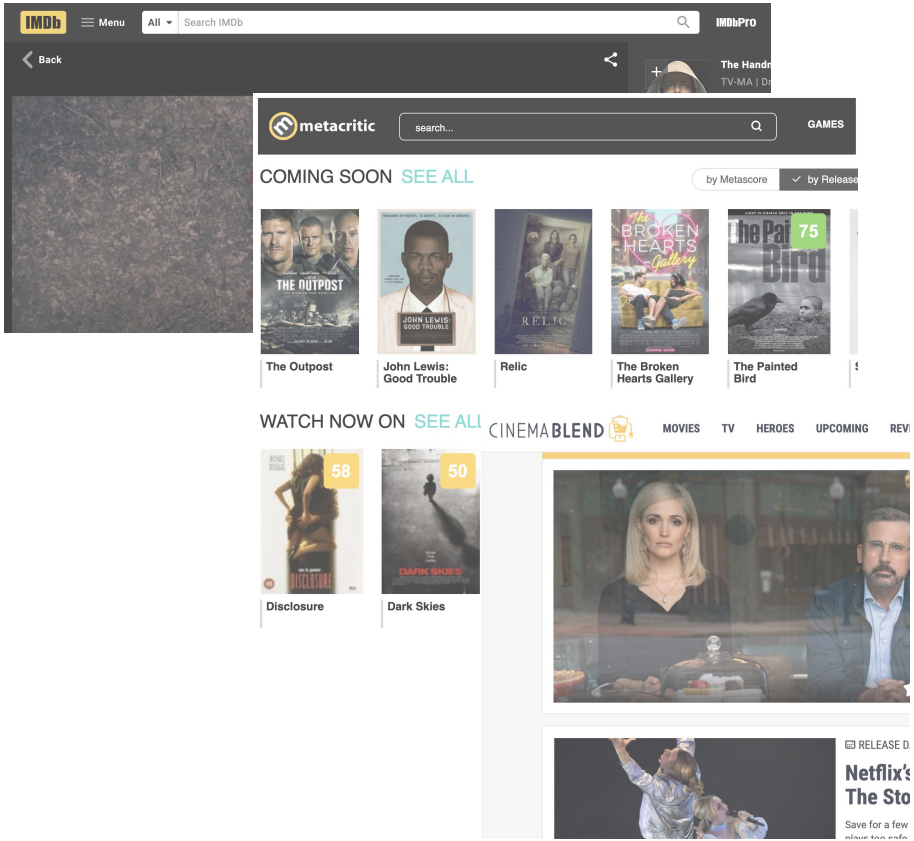
3. Expand the data acquisition to include all existing genres listed in Rotten Tomatoes



Takeaways & Improvements

- Model accurately predicts subreddit source
- Learned associative words
- Horror subreddit is less descriptive than Sci-fi (dataset is key)

- Change source of data (reviews/websites)



Thank you! Questions?

Rotten Tomatoes®