Multi-Label Movie Genre Classification

Springboard

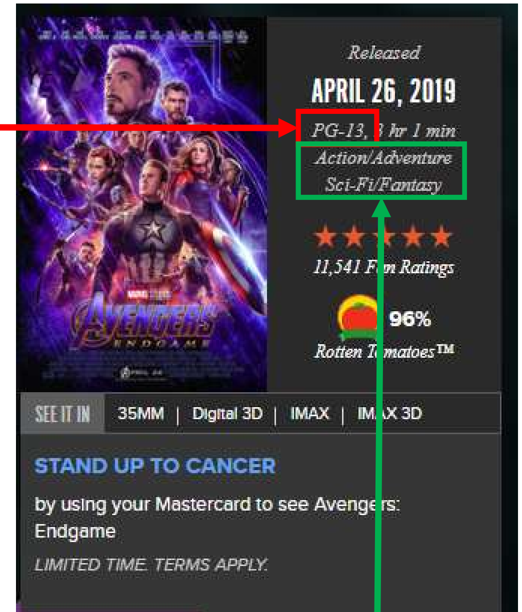
Shashank V. Maiya

Data Science Intensive Capstone Project

October 1, 2018 Cohort

Types of Classification Tasks

Multi-class Classification:
Each observation is assigned to one and only one class



Binary Classification

- Spam
- Not spam

Binary Classification:
Only 2 choices

Multiclass Classification

- Dog
- Cat
- Horse
- Fish
- Bird
- ...

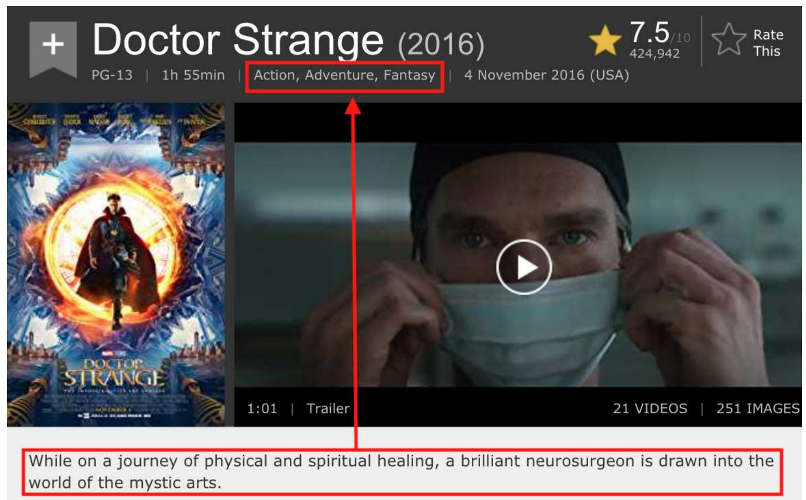
Multi-label Classification

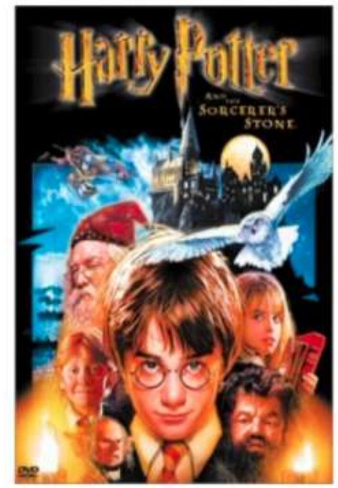
- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label Classification:
Each observation can be classified into multiple classes

Multi-Label Classification - Applications

Multi-Label Classification of Movie Genre using Plot





Horror: 0.02%
Romance: 0.02%
Adventure: 99.96%
Documentary: 0.0%

Multi-Label Classification for Image/Scene Categorization

- Multi-Label Classification of Genres based on movie Posters

Prediction Problem

- Total of 27 Possible Genres
- Movies are classified anywhere from 1 to 12 genres

- Given the plot of the movie, what are the genres they fall into?

Who might care?

Online Streaming Companies

NETFLIX

amazon prime video

hulu

HBO

YouTube

Movie Review Websites

IMDb

Rotten Tomatoes

YAHOO! MOVIES

Flixster

Data Overview

- Data set obtained from IMDB
- Column Description
  - Movie Title
  - Movie Plot
  - Plot Language
  - 27 Movie Genres

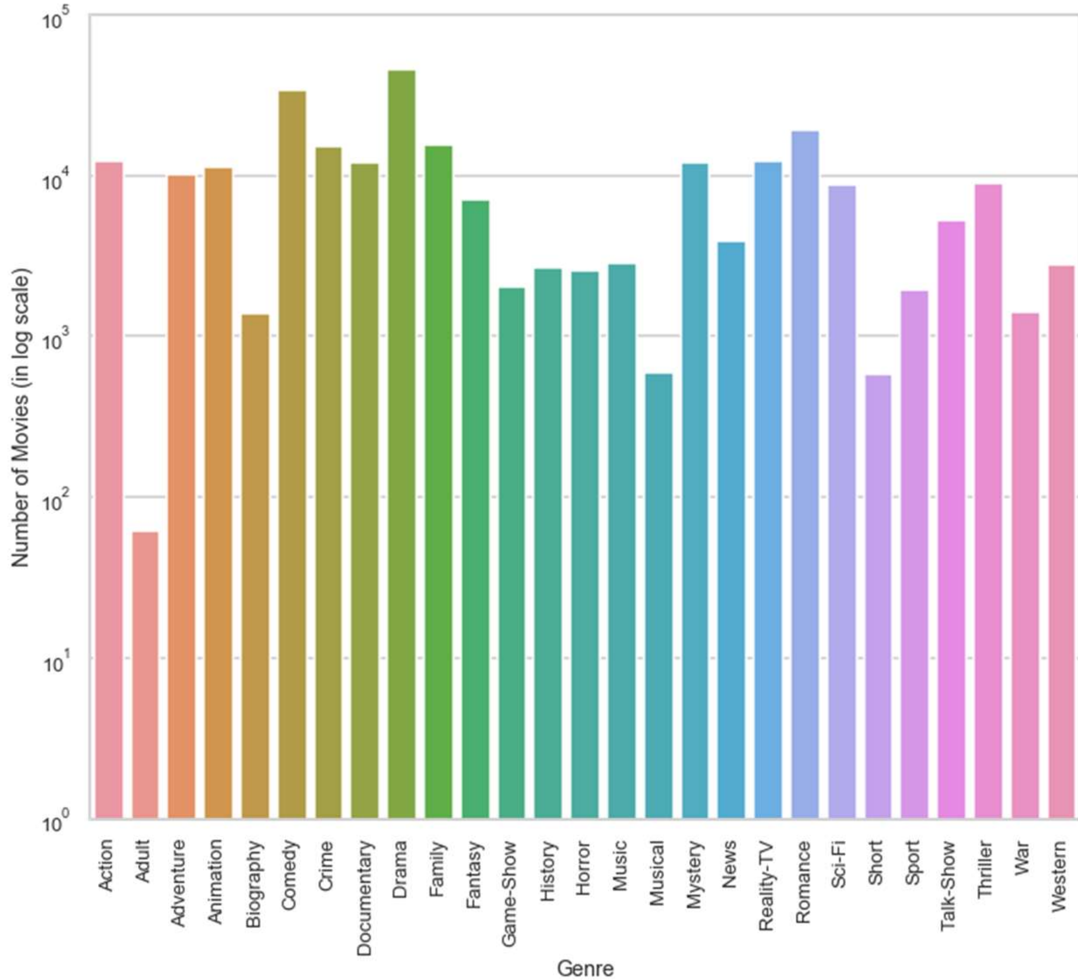
	title	plot	Action	Adult	Adventure	Animation	...	Sport	Talk-Show	Thriller	War	Western	plot_lang
0	"#7DaysLater" (2013)	dayslater interactive comedy series feature en...	0	0	0	0	...	0	0	0	0	0	en
1	"#BlackLove" (2015) {Crash the Party (#1.9)}	week leave workshops women consider idea ladie...	0	0	0	0	...	0	0	0	0	0	en

2 rows × 30 columns

Exploratory Data Analysis (EDA)

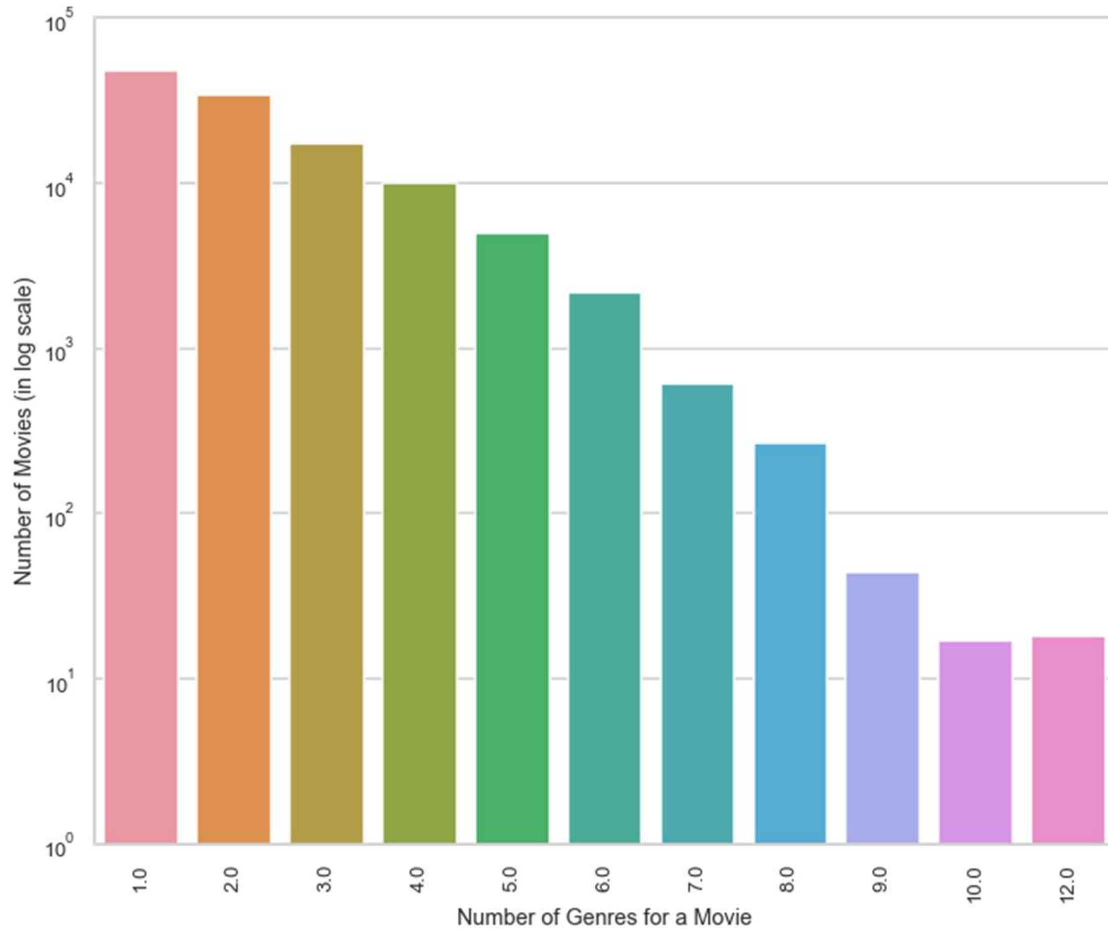
Number of Movies per Genre

- Highest Genre Movies = Drama (45891) followed by Comedy (33870)
- Lowest Genre Movies = Adult (61)

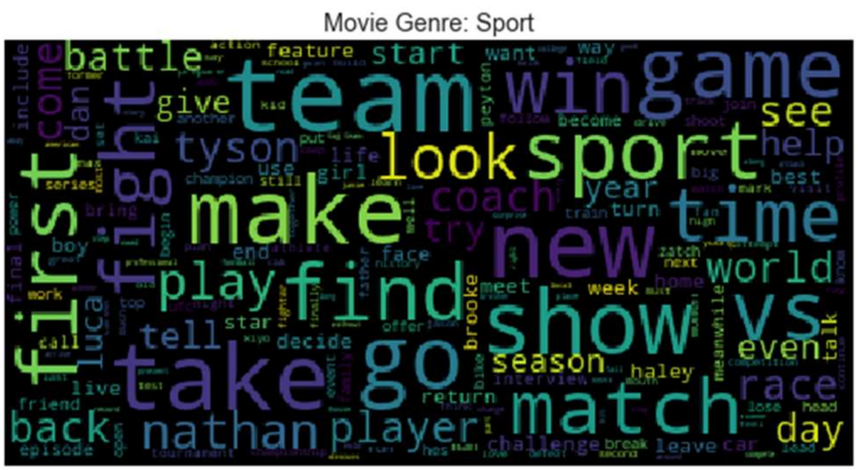


Number of Genres per Movie

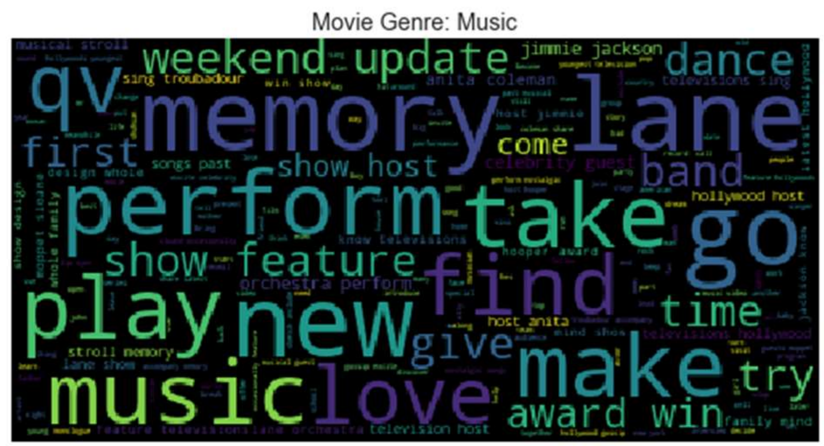
- Average of 2.1 genres per movie
- 18 movies are classified using 12 genres!



Word Cloud plots – [Sports, Music]



Relevant words – ‘team, vs, sport, first, match, player, race’



Relevant words – ‘memory, lane, perform, play, love, award, show’

Word Cloud plots – [Biography, War]

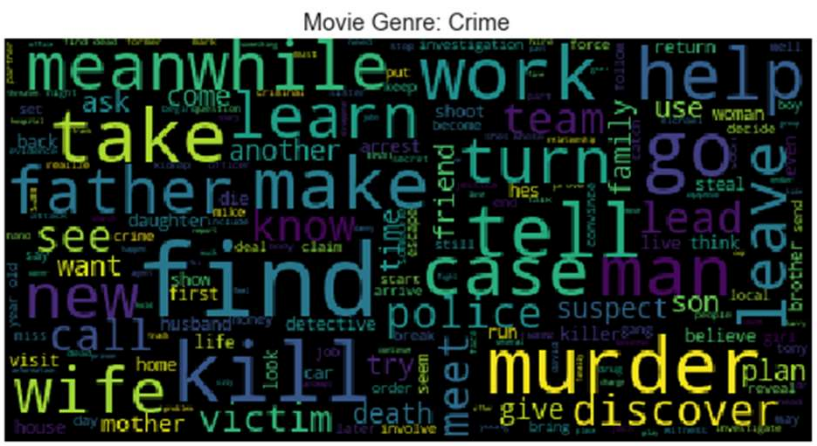
Movie Genre: Biography

Relevant words – ‘life, career, world, story, interview’

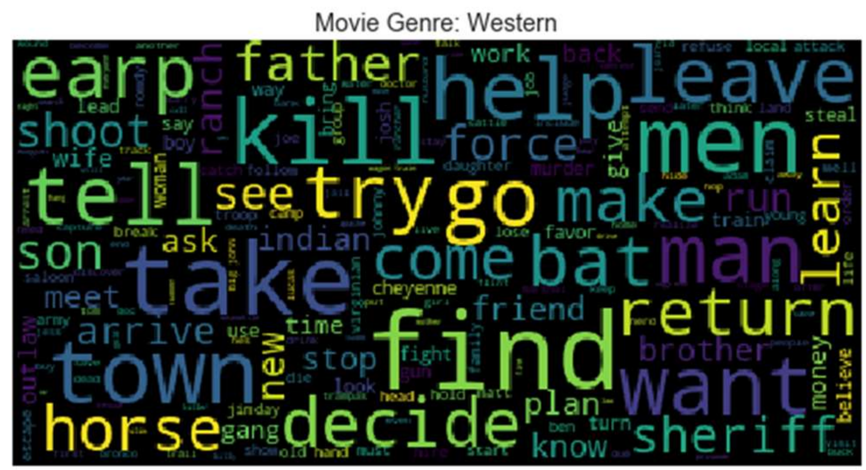
Movie Genre: War

Relevant words – ‘german, order, plan, kill, mission, american’

Word Cloud plots – [Crime, Western]



Relevant words – 'kill, murder, case, police, victim'



Relevant words – 'horse, sheriff, town, kill, Indian, gang'

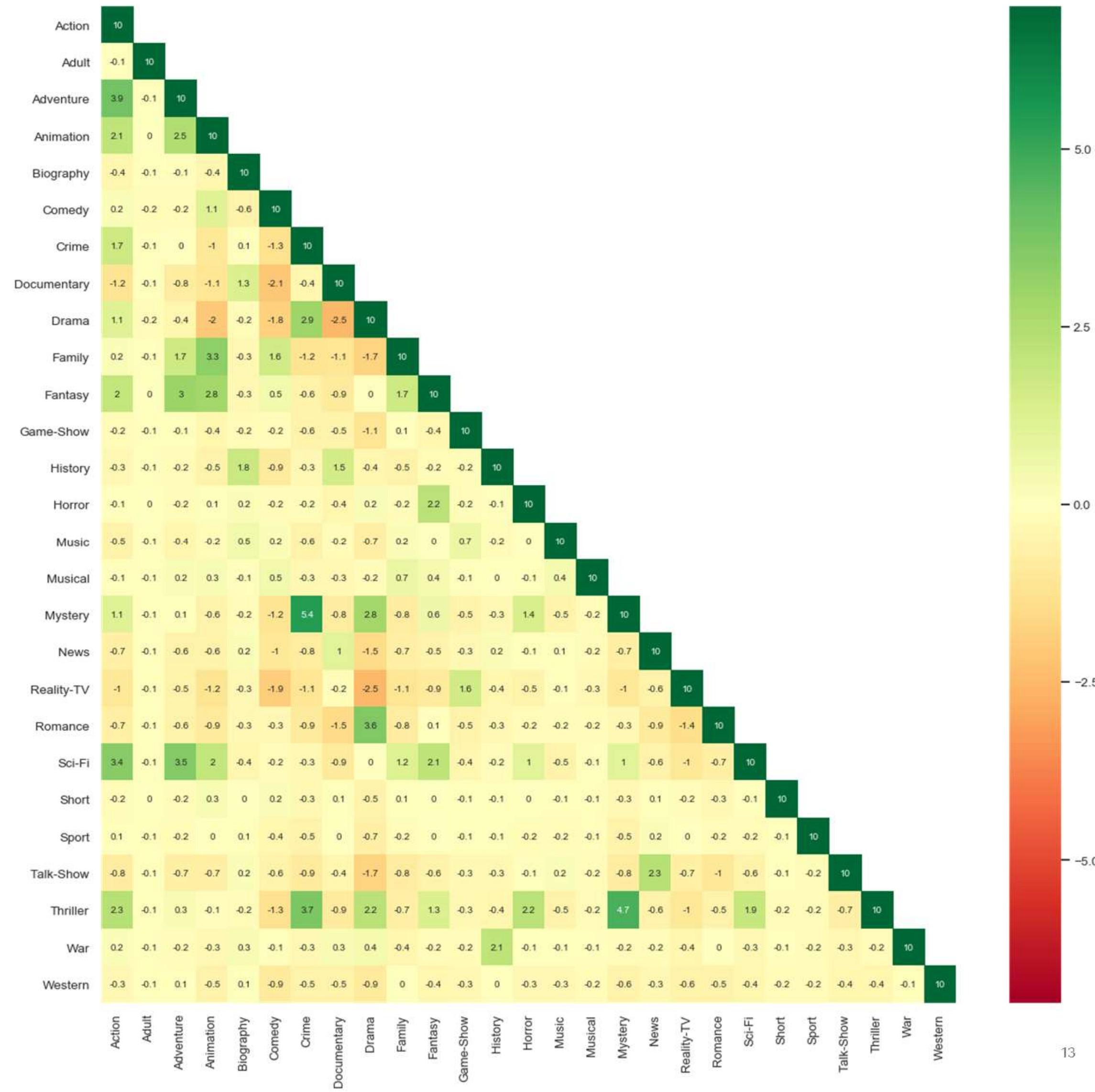
Correlation Analysis - Heatmap

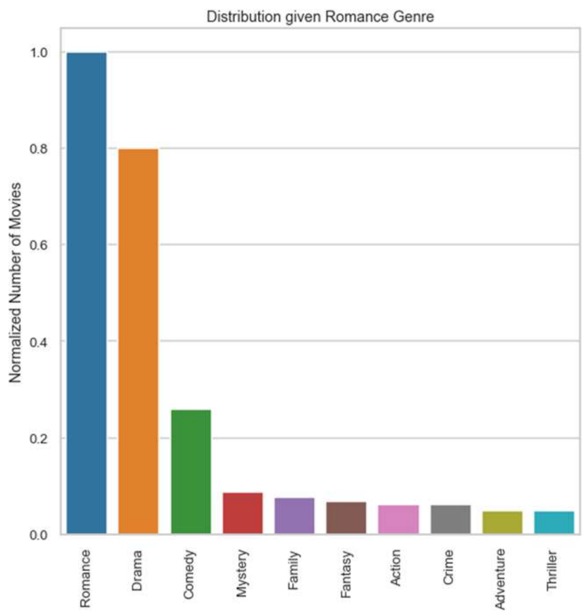
Genres with strong positive correlation

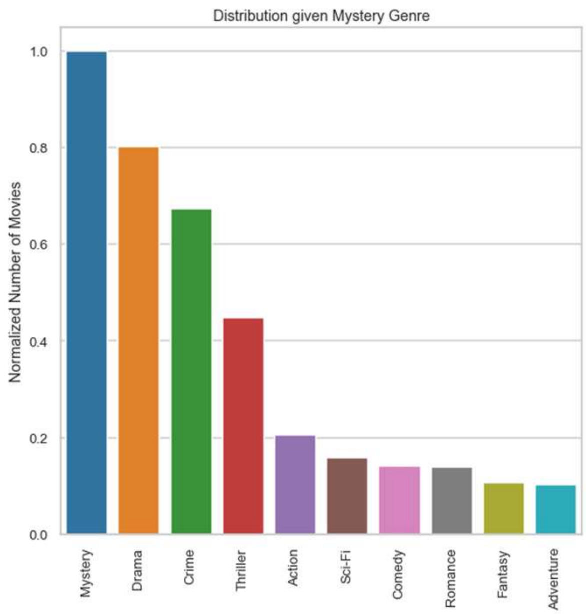
- Action, Adventure & Sci-Fi
- Animation, Fantasy & Family
- Crime, Thriller, Mystery & Drama
- Drama & Romance
- Game-Show & Reality-TV
- War & History

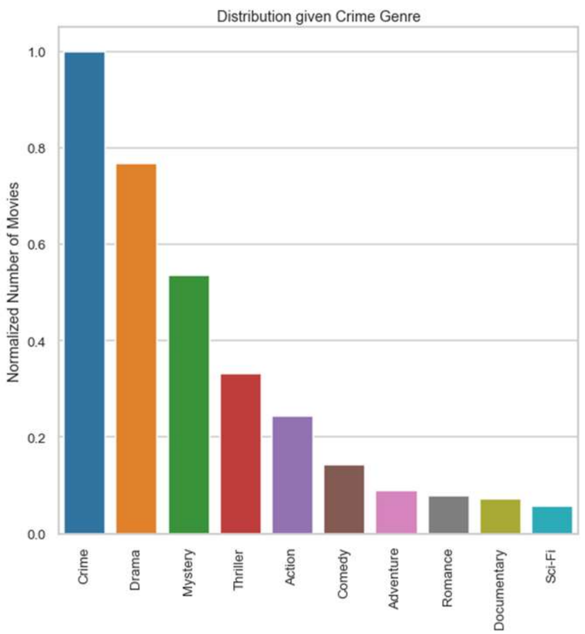
Genres with strong negative correlation

- Animation & Drama
- Comedy & Documentary



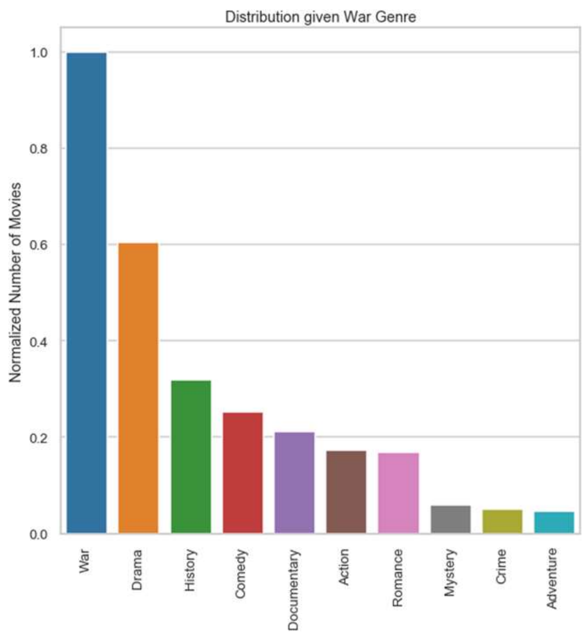


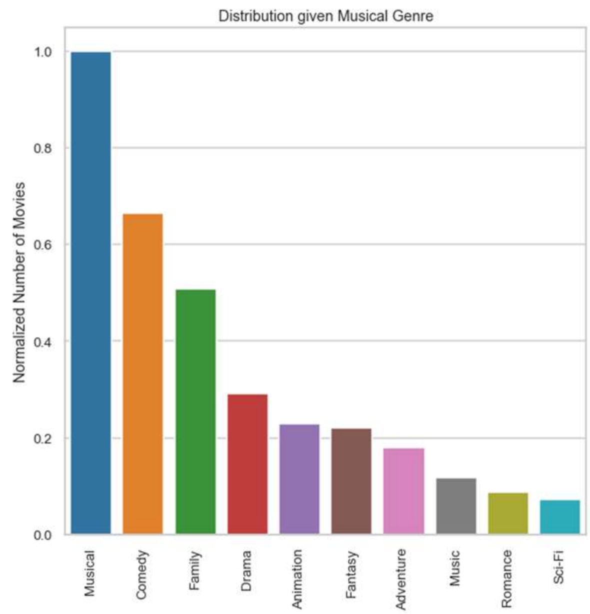


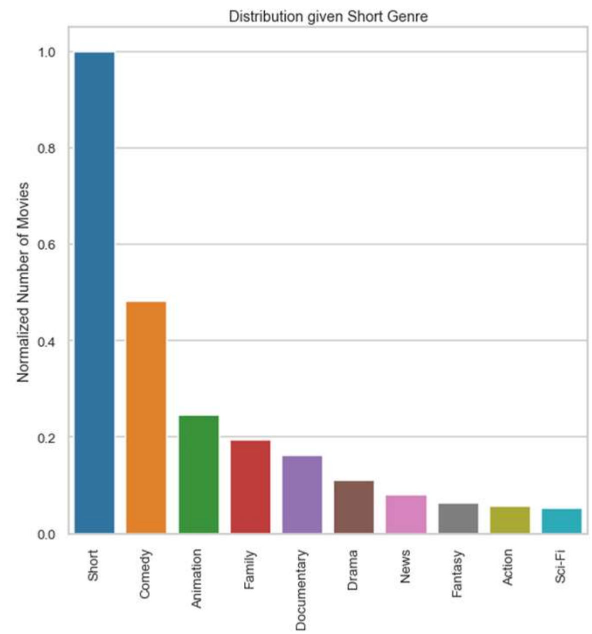


Multi-Genre Distribution Plots

80% of the Crime, Mystery & Romance movies are also categorized as Drama

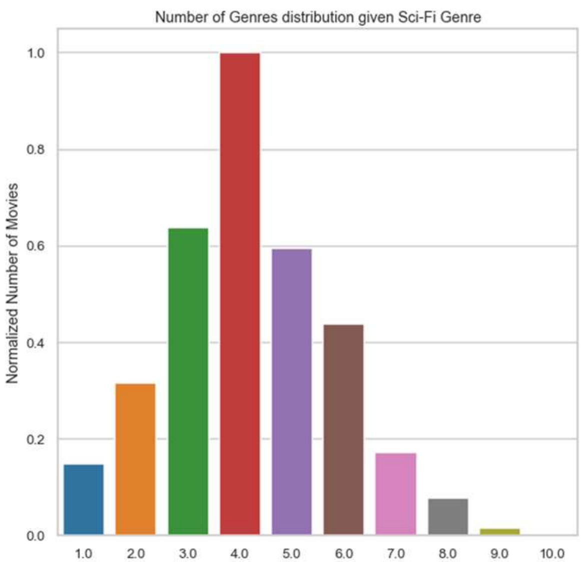


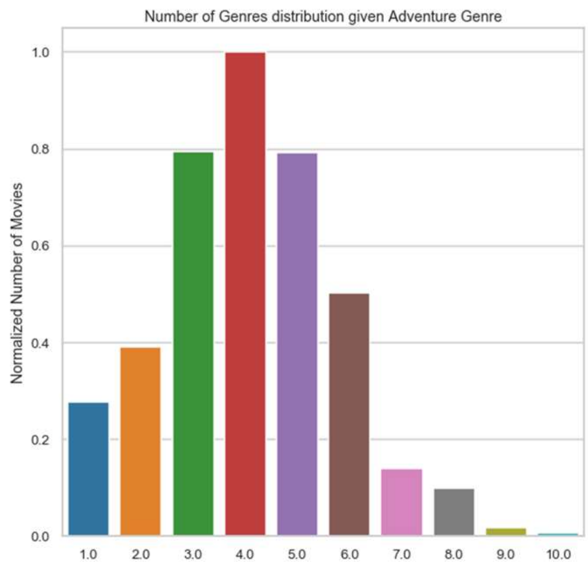


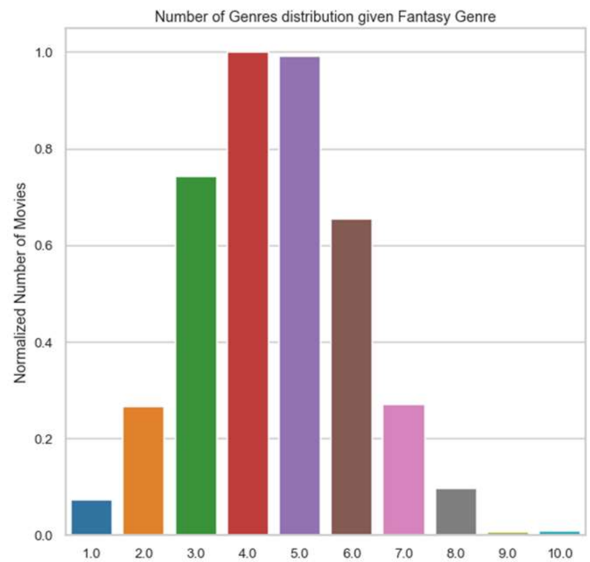


Multi-Genre Distribution Plots

65% of Musical movies are Comedy, 60% of War movies are Drama, Half of Short movies are Comedies

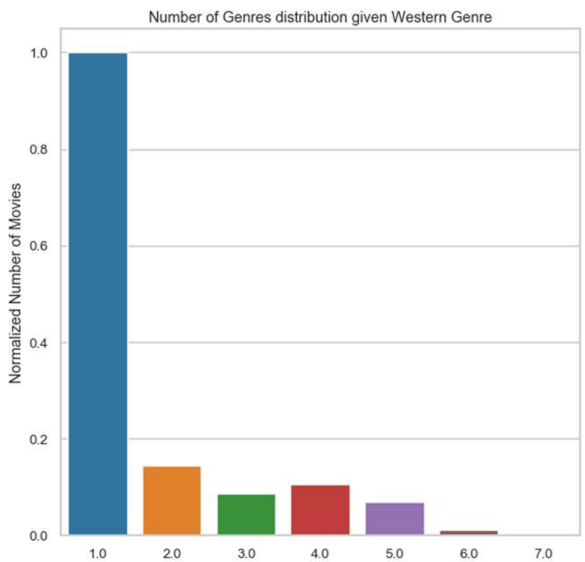


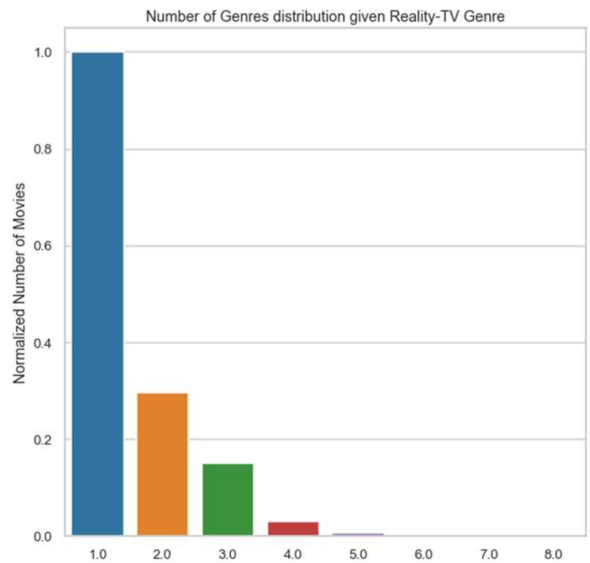


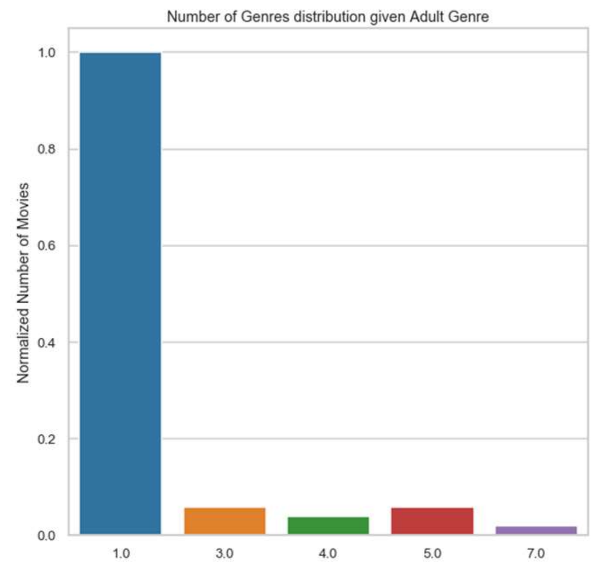


Number of Genres for [Sci-Fi, Adventure, Fantasy]

Adventure, Fantasy, Sci-Fi have 3 to 6 categories







Number of Genres for [Adult, Reality-TV, Western]

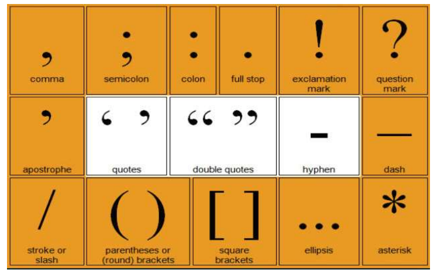
Adult, Reality-TV, Western are typically categorized to a single genre

```
print(remove_tags('<html><h2>Learning NLP</h2></html>'))
print(remove_tags(' <a>Movie-Plot</a>'))
```

executed in 3ms, finished 15:08:11 2019-04-28

```
Learning NLP
Movie-Plot
```

Remove HTML Tags



Remove Punctuations

áàâäãå ÁÀÂÄÃÅ çÇ
éèêë ÉÈÊË íìîï ÍÌÎÏ
ñÑ óòôöõ ÓÒÔÖÕ
úùûü ÚÙÛÜ ÿŸ

Convert accented characters to ASCII

```
def keep_alpha(sentence):
    alpha_sentence = re.sub('[^a-z A-Z]+', ' ', sentence)
    return alpha_sentence
```

Keep only alphabetic strings

Text Preprocessing

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Stop words removal

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Lemmatization

Raw	Lowercased
Canada CanadA CANADA	canada

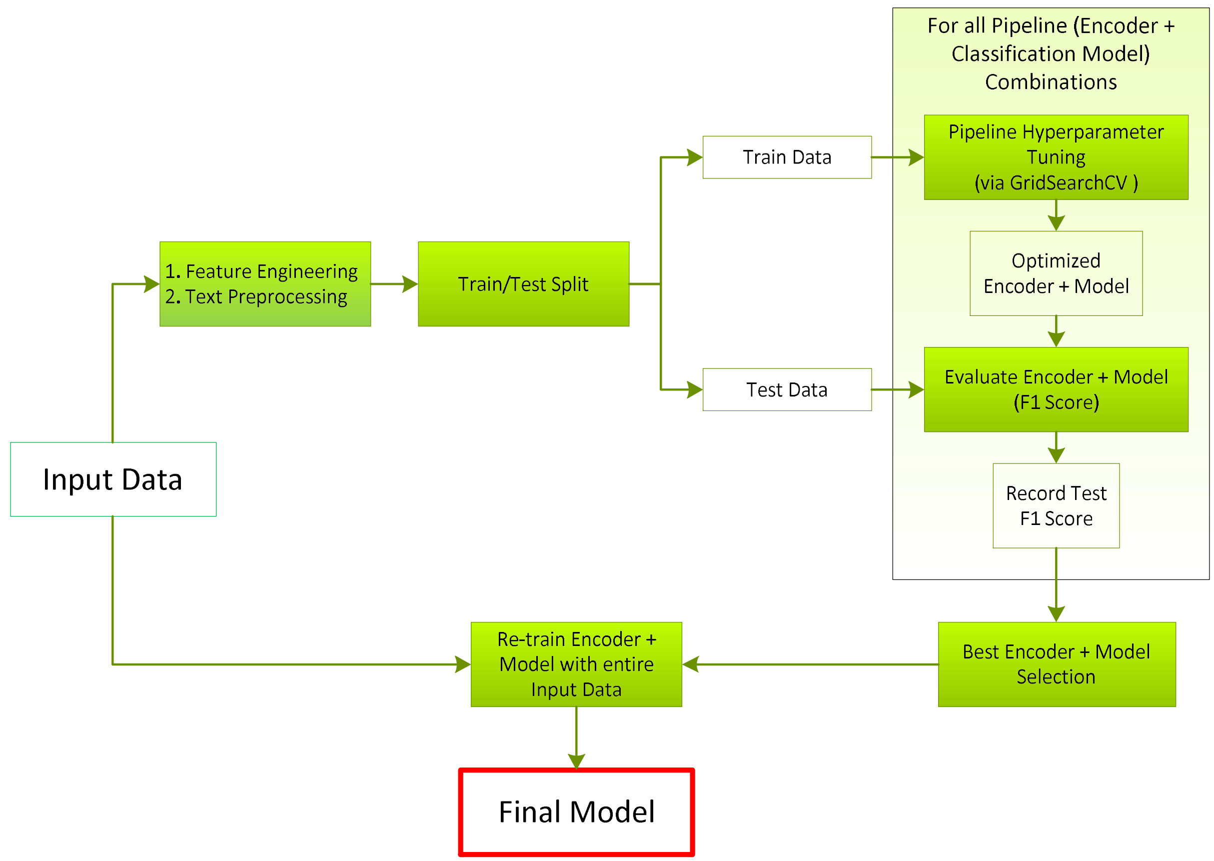
Lower casing all the words

Text Preprocessing

Modeling

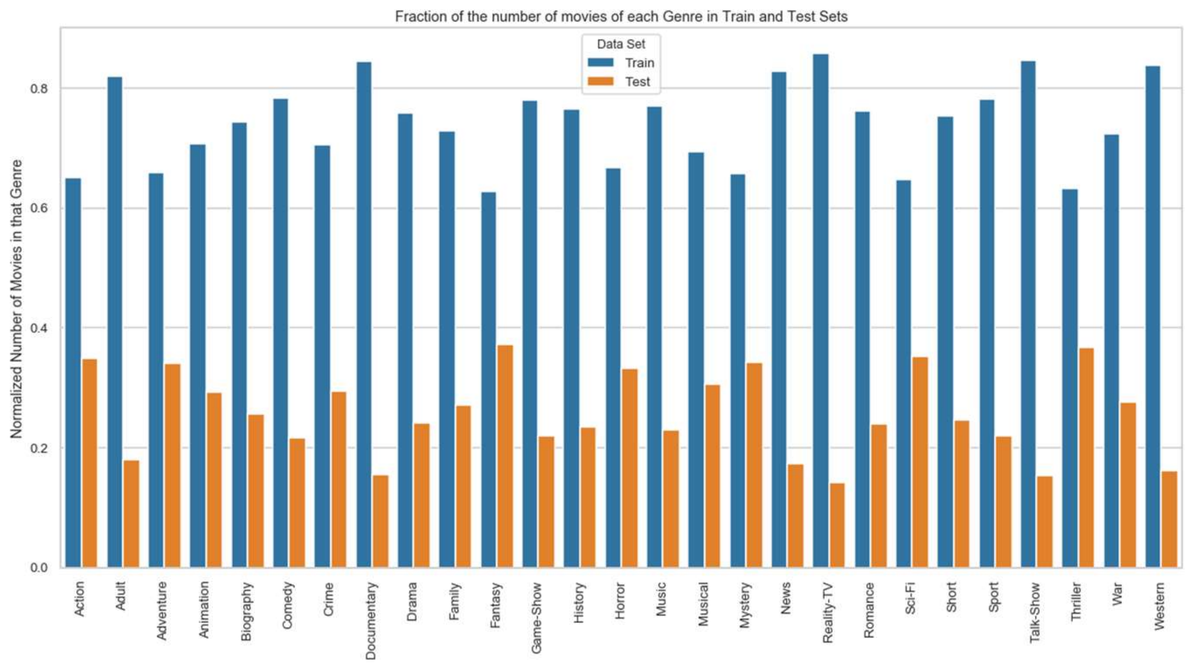
Modeling Overview

- Type: Supervised Learning
- Classification (Multi-label) Problem with 27 labels
- Pipeline consists of
  - Encoder
  - Classification Model



Train/Test Split

- Imbalanced Data with genre occurrences ranging from 61 (Adult) to 45891 (Drama)
- Train/Test split such that
  - At least 60% of the samples in the training set
  - At least 15% of the samples in the test set



For each Genre

$$\text{Precision (Action)} = \frac{\text{Number of movies 'correctly' identified as Action Genre}}{\text{Total number of movies that have been identified as Action Genre}}$$

$$\text{Recall (Action)} = \frac{\text{Number of movies 'correctly' identified as Action Genre}}{\text{Total number of Action Genre movies in the data set}}$$

$$\text{F1 score (Action)} = \frac{2 * \text{Precision(Action)} * \text{Recall(Acation)}}{\text{Precision(Action)} + \text{Recall(Action)}}$$

Overall F1 Score = Weighted Average of individual Genre F1 Score

Evaluation Metric – Overall F1 Score

Multi-Label Algorithm

- Binary Relevance
- Label Powerset
- Label Powerset with Clustering

Text Encoder

- Count Vectorizer
- TF-IDF
- Sentence Embedding

Classifiers

- Logistic Regression, Linear SVC, Naïve Bayes
- Cosine Similarity
- Neural Networks

Multi-Label Algorithm + Text Encoder + Classifiers

Binary Relevance

- Treat each label as a separate single class Classification
- 27 Genres ➔ 27 Binary Classifiers

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Label Powerset

- Treat each of the unique genre combinations found in the training data as a possible class
- 1505 Unique genre combinations ➔ Multi-class classification with 1505 classes

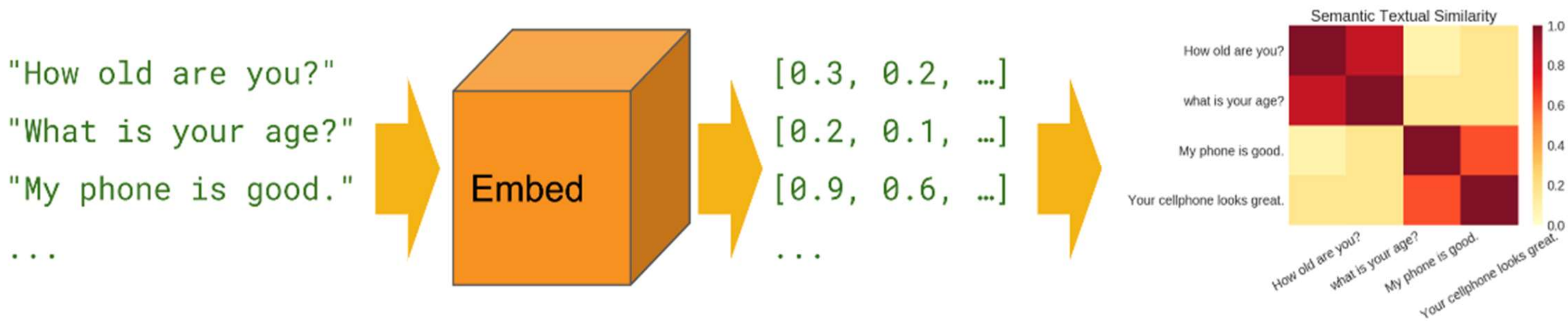
X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Label Powerset with Clustering

- Reduce the number of genre combinations by clustering (from 1505 to 75)

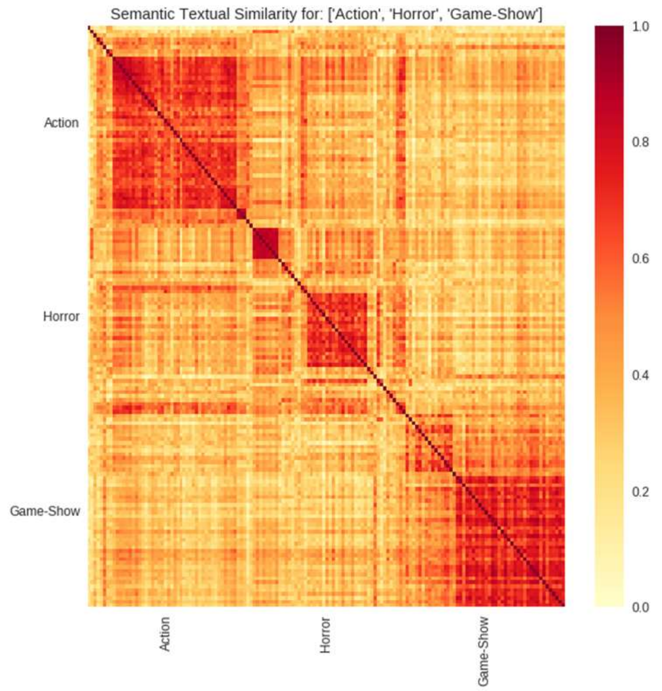
Multi-Label Classification Algorithms



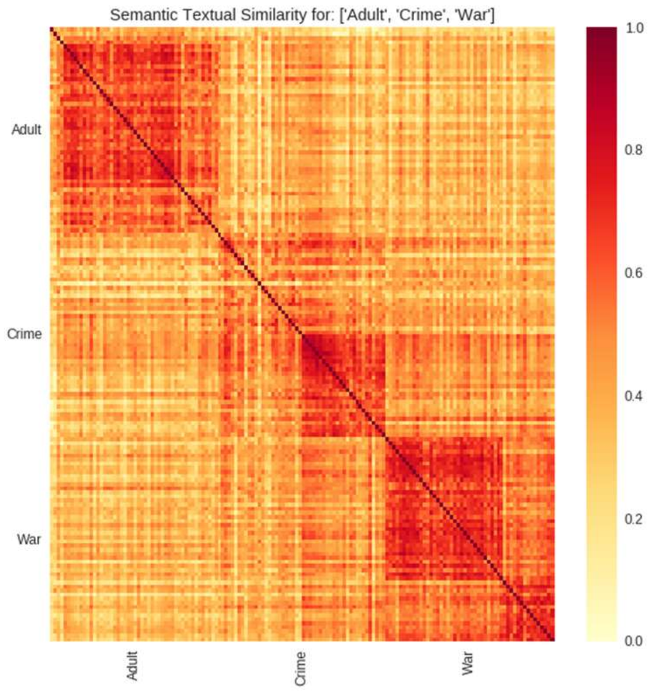
Text Encoder – Sentence Embedding

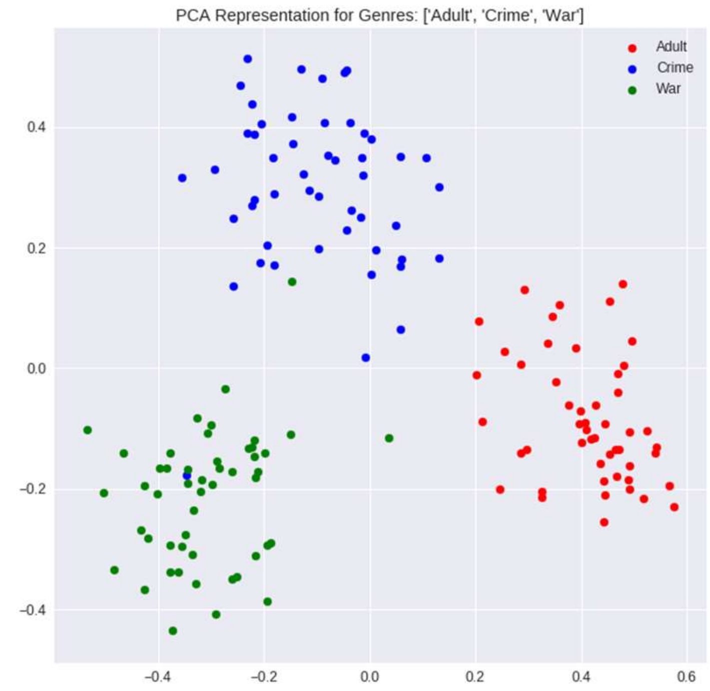
- Google Universal Sentence Encoder (USE) converts input sentence to 512 dimension numeric vectors
- Preserves the semantic meaning of the sentence

Sentence Embedding – [Action, Horror, Game-Show]



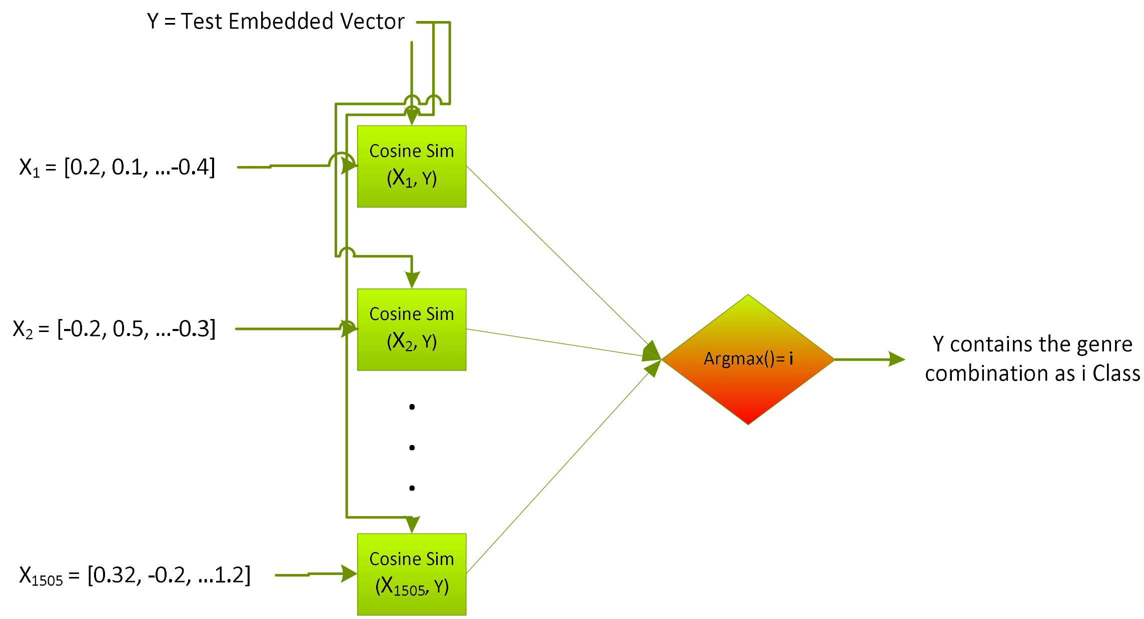
Sentence Embedding – [Adult, Crime, War]



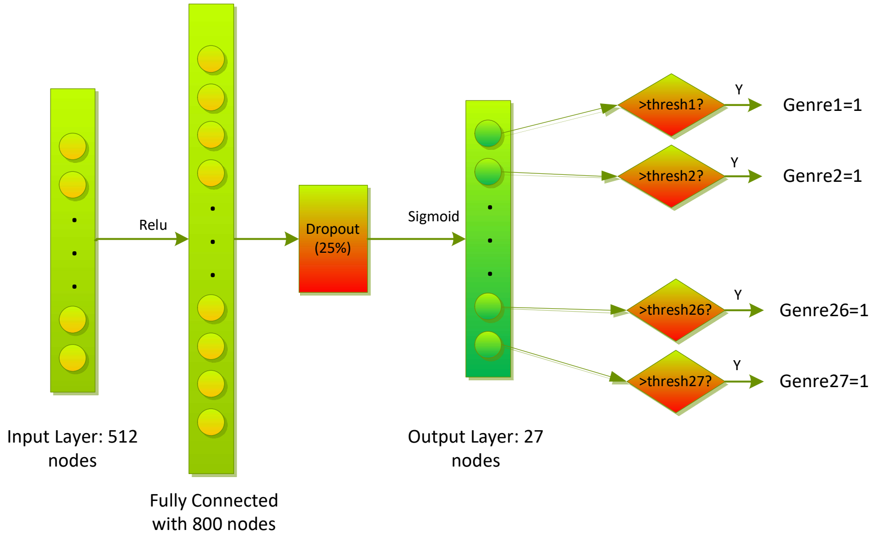


Classification Models – Cosine Similarity

- Sentence Embedding of similar genres cluster together
- Cosine Similarity $(\vec{x}, \vec{y}) =$
- $\frac{\vec{x} \cdot \vec{y}}{||\vec{x}||\ ||\vec{y}||} = \frac{\sum_{i=1}^{512} x_i y_i}{\sqrt{\sum_{i=1}^{512} x_i^2}\ \sqrt{\sum_{i=1}^{512} y_i^2}}$

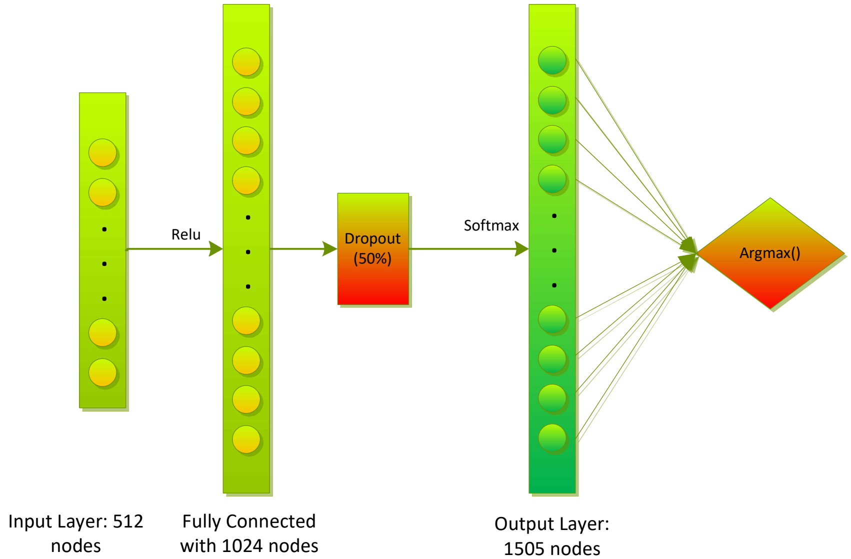


Classification Models – Neural Network



Using Binary Relevance

- Each Genre predicted separately ➔ Output Layer = 27 Neurons
- Use sigmoid Activation Function



Using Label Powerset

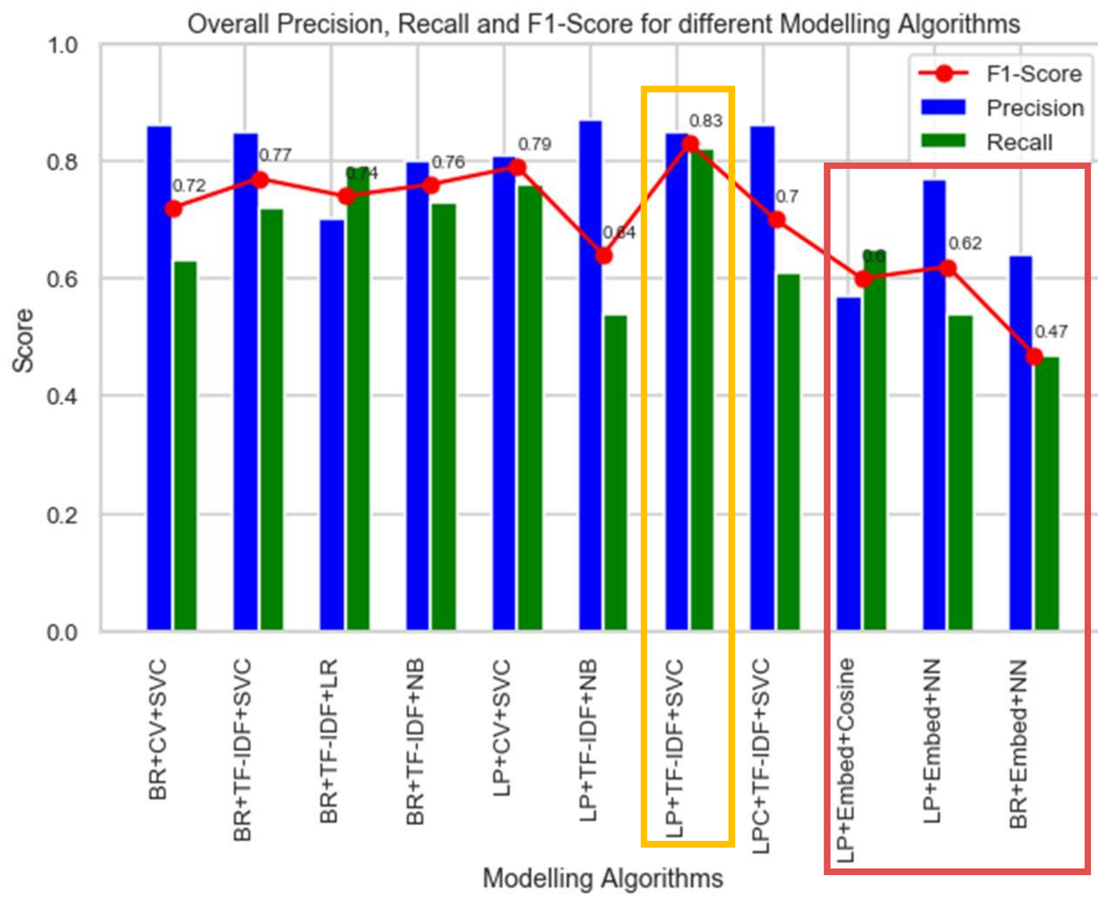
- Output Layer = 1505 Neurons
- Use softmax Activation Function since a single class is predicted

Result Summary

- Best predicting Model: Label Powerset + TF-IDF + Linear SVC (Overall F1-Score = 0.83)
- Models using Sentence Embedding performance is worse compared to other Vectorizers and ML models

BR: Binary Relevance
LP: Label Powerset
LPC: Label Powerset with Clustering
CV: Count Vectorizer
Embed: Sentence Embedding via USE

SVC: Linear Support Vector Classifier
LR: Logistic Regression
NB: Naïve Bayes
Cosine: Cosine Similarity
NN: Neural Networks



Label Powerset + TF-IDF + Linear SVC

- Hyperparameters
  - TF-IDF
    - Ngram = (1, 2)
    - Min_df = 2
    - Max_df = 0.5
  - LinearSVC
    - C = 10
- Overall F1-score = 0.83
- Poorest performing Genre = Adult
  - F1-score = 0.32, but only 11 samples

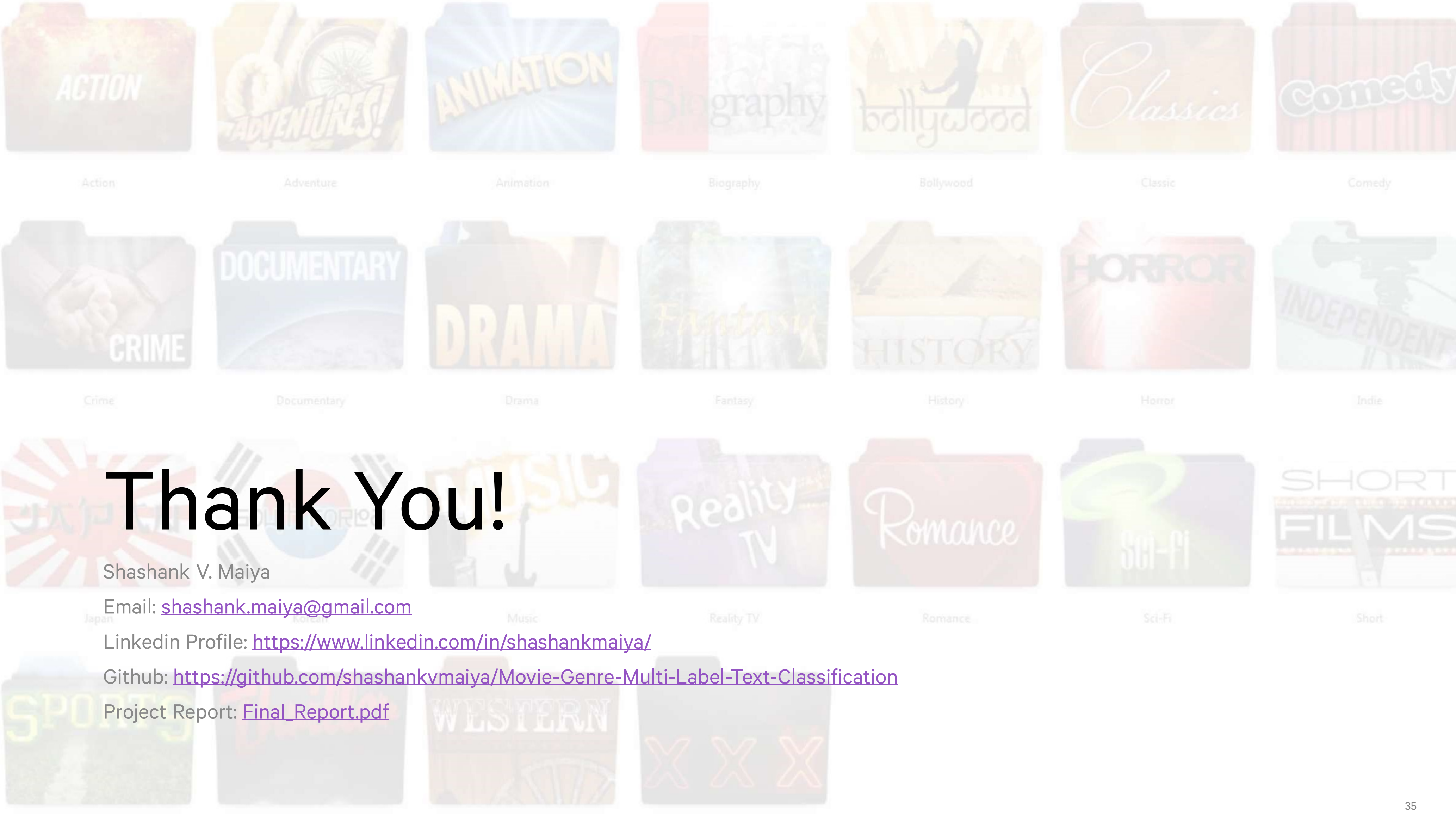
	Precision	Recall	F1-Score	Support
Action	0.87	0.82	0.84	4321.0
Adult	0.38	0.27	0.32	11.0
Adventure	0.85	0.80	0.82	3496.0
Animation	0.86	0.84	0.85	3333.0
Biography	0.54	0.43	0.48	354.0
Comedy	0.87	0.84	0.85	7320.0
Crime	0.86	0.86	0.86	4453.0
Documentary	0.72	0.73	0.73	1863.0
Drama	0.91	0.87	0.89	11067.0
Family	0.82	0.84	0.83	4173.0
Fantasy	0.86	0.78	0.81	2643.0
Game-Show	0.79	0.90	0.85	450.0
History	0.64	0.65	0.65	623.0
Horror	0.75	0.60	0.67	854.0
Music	0.76	0.80	0.78	654.0
Musical	0.75	0.65	0.70	182.0
Mystery	0.85	0.80	0.82	4114.0
News	0.76	0.80	0.78	681.0
Reality-TV	0.80	0.78	0.79	1748.0
Romance	0.87	0.86	0.87	4581.0
Sci-Fi	0.88	0.80	0.84	3055.0
Short	0.53	0.40	0.46	142.0
Sport	0.75	0.78	0.76	426.0
Talk-Show	0.77	0.86	0.81	809.0
Thriller	0.84	0.73	0.78	3254.0
War	0.74	0.71	0.72	388.0
Western	0.68	0.85	0.75	445.0
Avg/Total	0.85	0.82	0.83	65440.0

Conclusions

- Out of the 11 models considered, the best predicting model uses TF-IDF Vectorizer, Linear Support Vector Classifier and Label Powerset approach to achieve an overall F1-score of 0.83
- Sentence Embedding doesn't provide any benefit
- EDA specific Observations
  - Drama and Comedy the most popular Genre
  - On an average, a movie is classified into 2 genres (and a maximum of 12 genres)
  - Few strongly correlated genres include – a) Crime, Mystery & Thriller, b) Drama & Romance
  - 80% of Crime, Mystery and Thriller movies are also categorized as Drama

Limitations and Ideas

- Improving Sentence Embedding
  - Sentence embedding doesn’t require the sentence lemmatization, or stop word removal, or in fact any of the text preprocessing steps. Use the original text before preprocessing to obtain sentence embedding



Thank You!

Shashank V. Maiya

Email: shashank.maiya@gmail.com

Linkedin Profile: https://www.linkedin.com/in/shashankmaiya/

Github: https://github.com/shashankvmaiya/Movie-Genre-Multi-Label-Text-Classification

Project Report: Final_Report.pdf