



Multimodal Deep Learning To

# Predict Movie Genres

# The problem

01

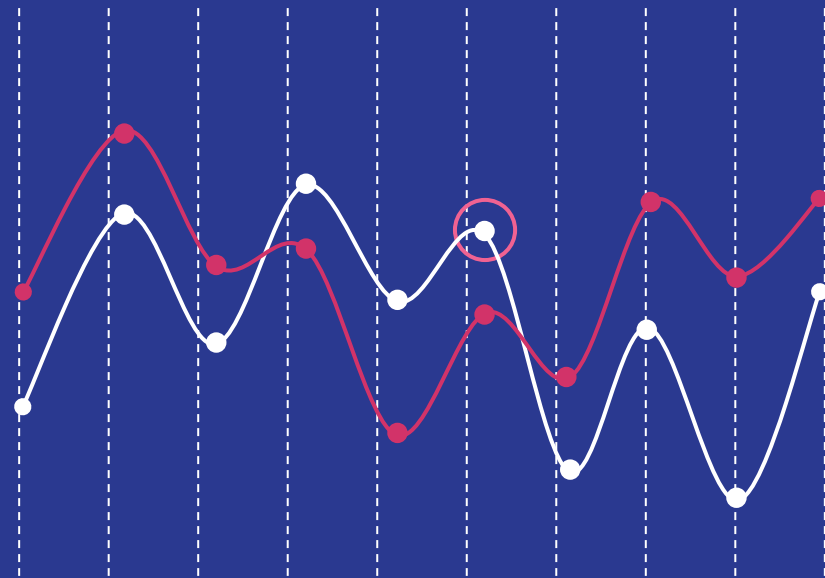
Our aim is to implement a multimodal framework for multilabel classification of movies (according to movie genres) with the help of movies' description and poster.

## OUR APPROACH

In this project, we will concatenate the features extracted from images and text sequences using a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network, respectively.

These features will be used to try and predict movie genres.

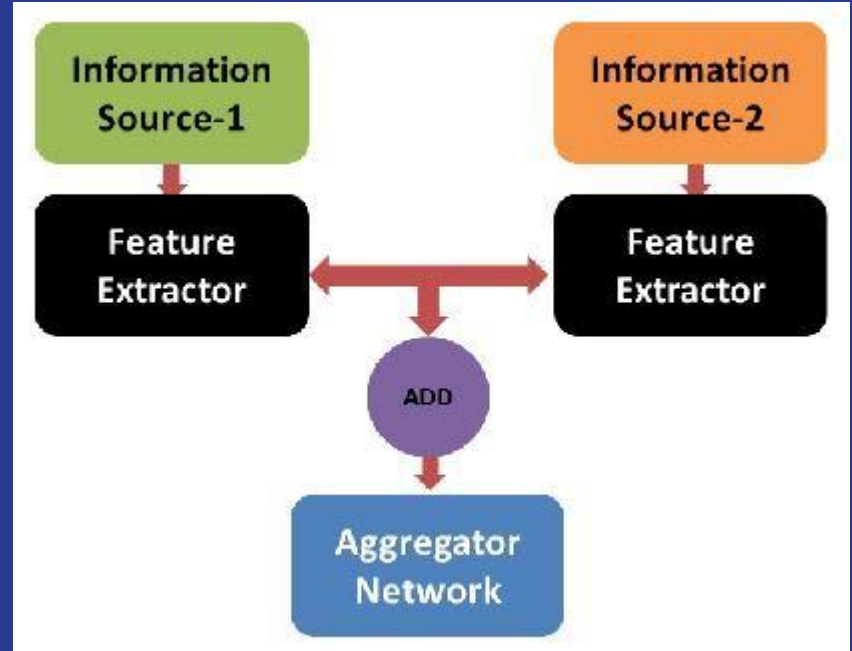
# Model Architecture



—

# Multimodal learning

- It suggests that when a number of our senses - visual, auditory, kinesthetic - are being engaged in the processing of information, we understand and remember more.
- In multimodal learning, we aim to do information fusion from different modalities to improve our network's predictive ability.



# Model Architecture



*Images/  
Movie  
Posters*

*CNN*

*Fully  
connected  
layer*

In the 1960s, Cambridge University student and future physicist Stephen Hawking (Eddie Redmayne) falls in love with fellow collegian Jane Wilde (Felicity Jones). At 21, Hawking learns that he has motor neuron disease. Despite this -- and with Jane at his side -- he begins an ambitious study of time, of which he has very little left, according to his doctor.

*Text/  
Movie  
Overview*

*LSTM*

*Fully  
connected  
layer*

*Concatenate  
images and text  
features*


*Fully  
connected  
layer*

*Final  
layers with  
18 outputs*

# Convolutional Neural Network (CNN)

The model comprises of two networks :

64 x 64 (RGB)



0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

Input Image

Convolution

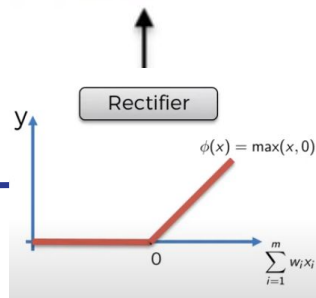
Pooling

Flattening

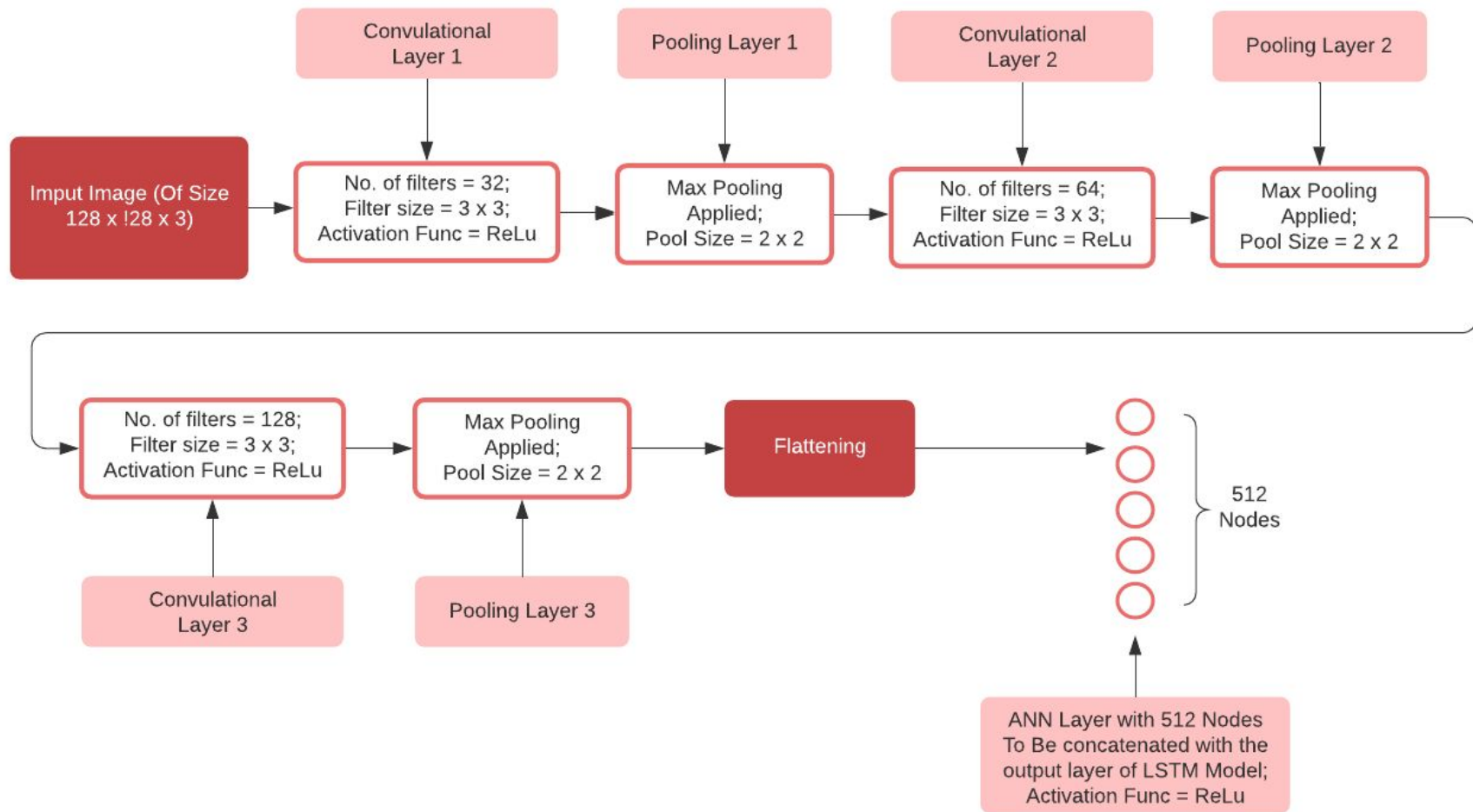
Convolutional Layer

Pooling Layer

ReLU activation function used on the feature maps in the convolution layer.



512 Nodes

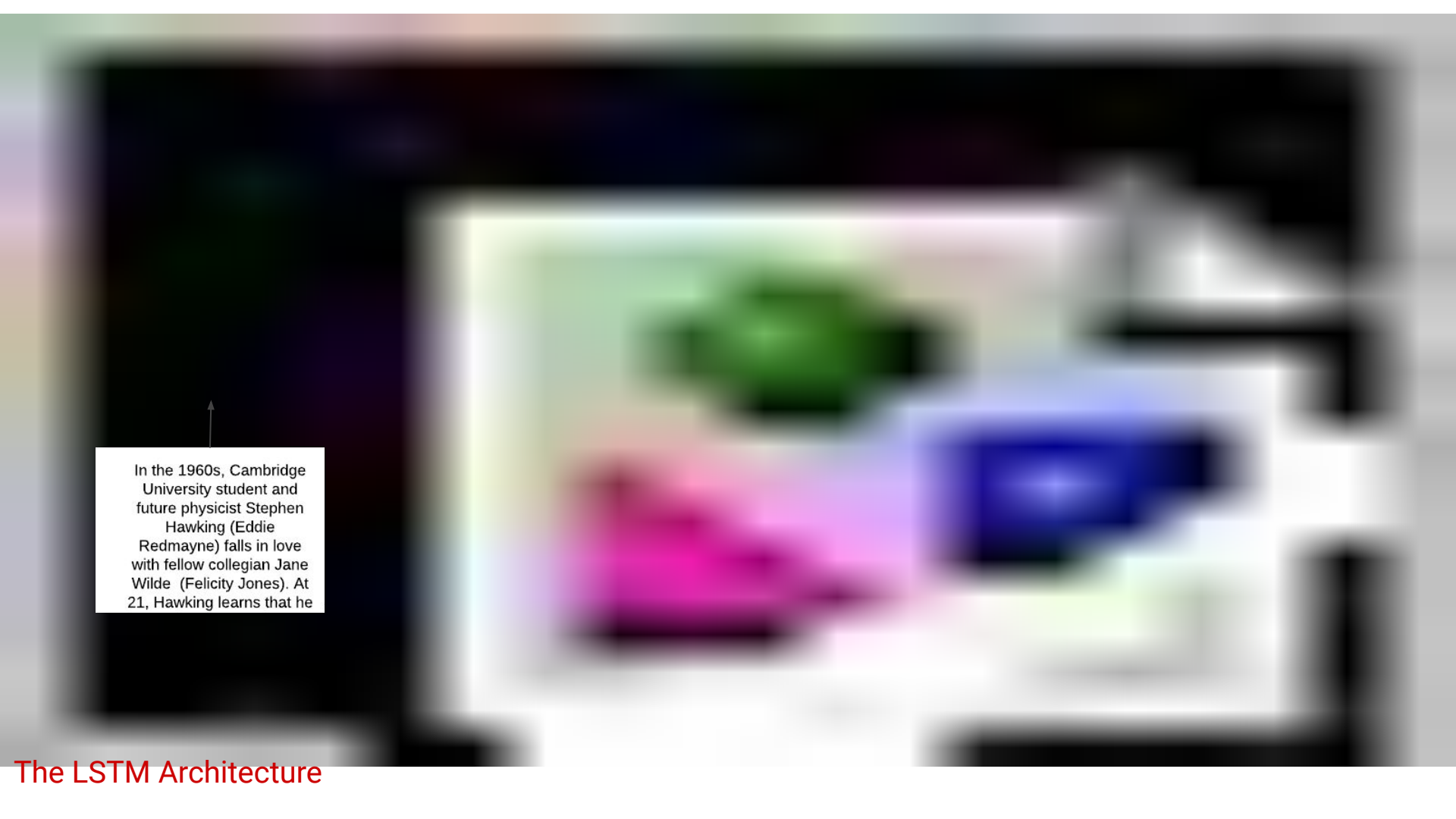


The CNN Architecture Explained

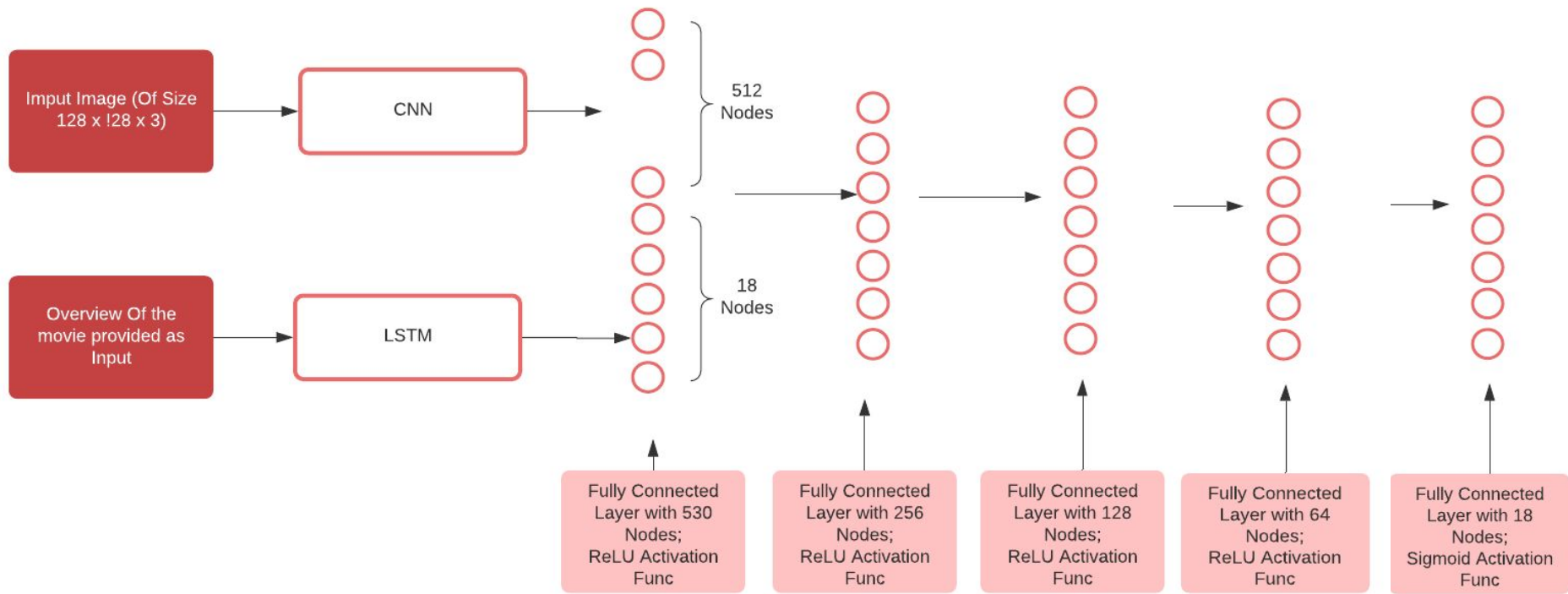


# Stacked LSTMs

The model comprises of two networks :



In the 1960s, Cambridge University student and future physicist Stephen Hawking (Eddie Redmayne) falls in love with fellow collegian Jane Wilde (Felicity Jones). At 21, Hawking learns that he



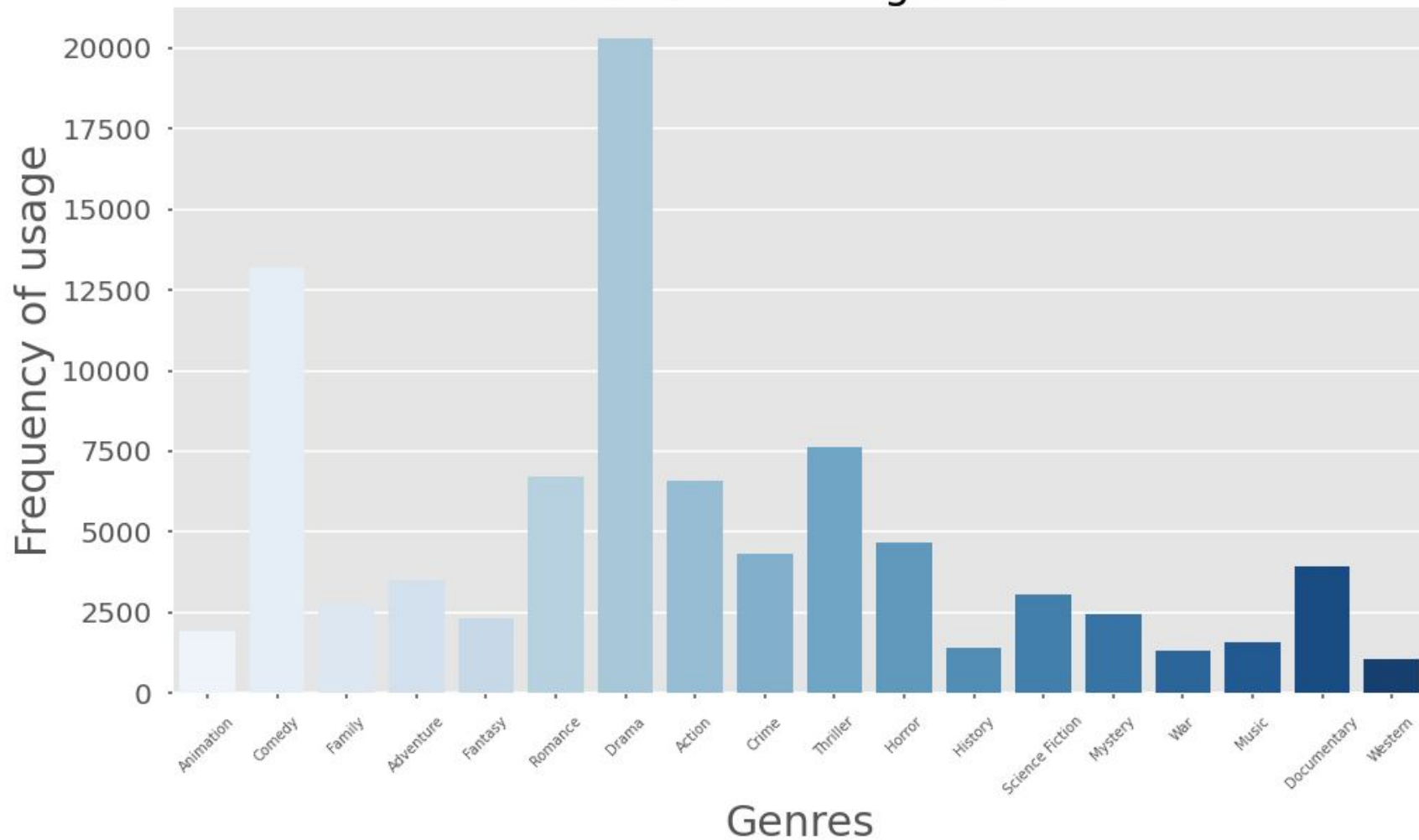
## The Complete Architecture Overview

The fully connected layers of the LSTM and CNN then concatenated and fed forward till the final fully connected layer. The last layer has an output size of 18, each correlating to one genre.



# Data Gathering and Preprocessing

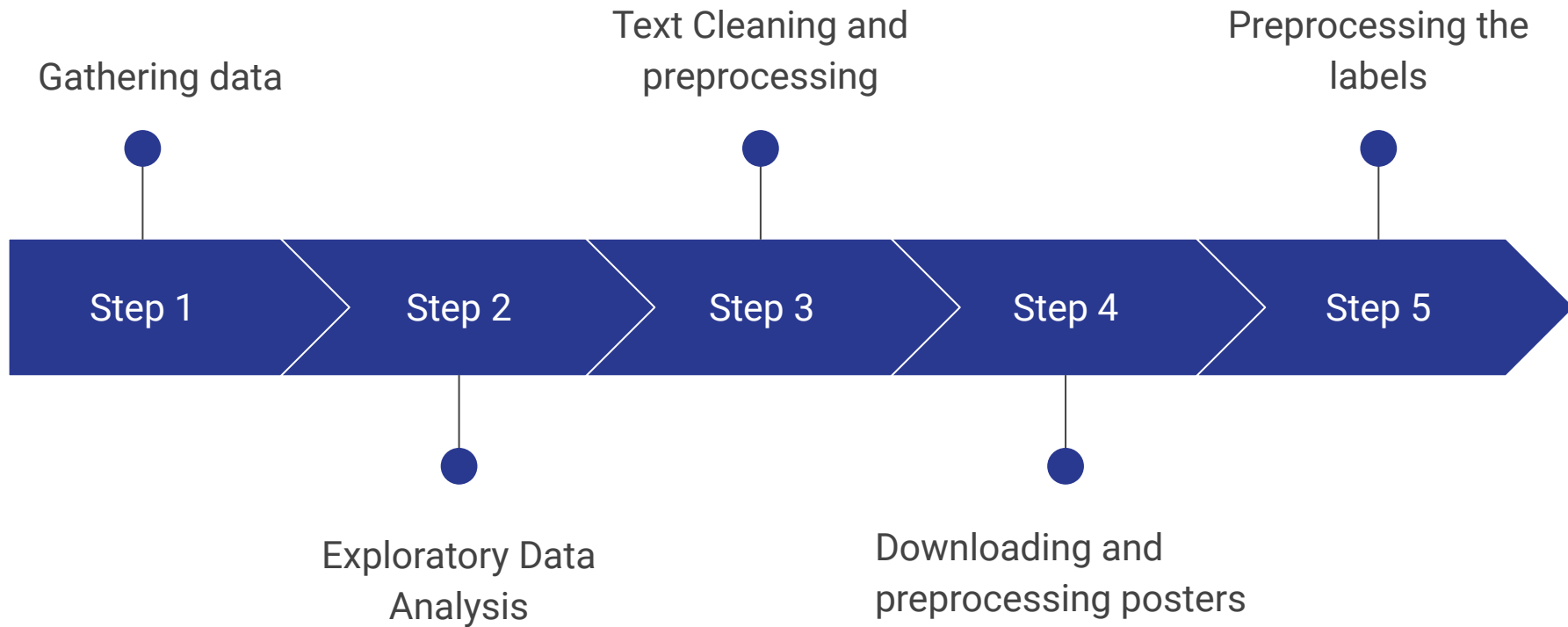
# Most common genres



# DATASET

- The dataset that we will be using is “[The Movies Dataset](#)” from kaggle.
- It consists of over 40000 tuples that includes:
  1. Movie title
  2. Overviews
  3. Poster URL link, and
  4. Genres





# Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1 score





# #Model 0

- **Sample size = 41,069 (train: test: val :: 29568: 8214: 3286)**
- **Model 1**
  - Number of LSTM layers = 2 each having 64 units
  - After LSTM, a fully connected layers with 18 neurons
- **Model 2**
  - Number of CNN layers = 2 each having 64, 128 filters resp.
  - In each CNN layer , the Kernel size was taken 3x3 ,pool size of 2x2 and activation function relu
  - After CNN, a fully connected layers with 512 neurons
- **Concatenated Layer**
  - 2 fully connected layers with 512 and 256 neurons and 2 dropout layers 30% each after each layer.
  - Final layer having 18 neurons corresponding to 18 movie genres

Optimizer : Adam , Threshold used for classification = 0.5, Batch Size = 128

# Final Results of Evaluation

- Accuracy= 0.8192975012049812
- Precision= 0.6106609321248502
- Recall= 0.6403235611810117
- f1= 0.62531638610571

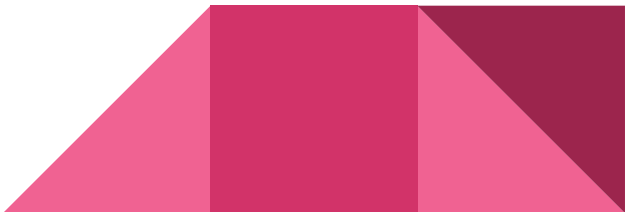


# #Model 1

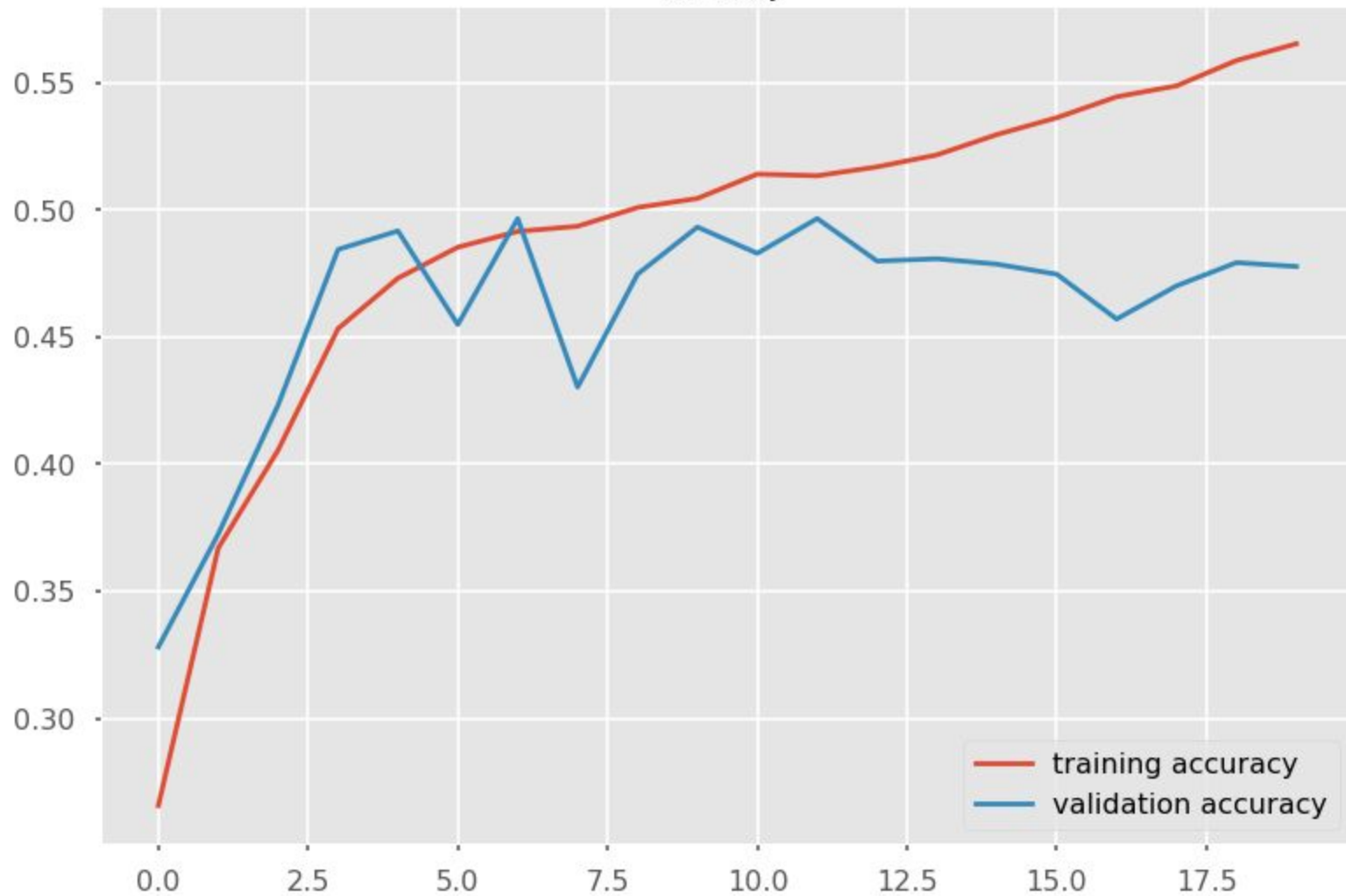
- **Sample size = 41,069 (train: test: val :: 29568: 8214: 3286)**
- **Model 1**
  - Number of LSTM layers = 2 each having 64 units and 2 dropout layers 30% each after each layer.
  - After LSTM, a fully connected layers with 18 neurons
- **Model 2**
  - Number of CNN layers = 4 each having 32, 64, 128,128 filters resp.
  - In each CNN layer , the Kernel size was taken 3x3 ,pool size of 2x2 and activation function relu
  - A dropout layer(30%) was added after the last convolution layer.
  - After CNN, a fully connected layers with 512 neurons
- **Concatenated Layer**
  - 2 fully connected layers with 512 and 256 neurons and 2 dropout layers 30% each after each layer.
  - Final layer having 18 neurons corresponding to 18 movie genres

Optimizer : Adam , Threshold used for classification = 0.5, Batch Size = 256

# Final Results of Evaluation

- Accuracy= 0.8955306658009362
  - Precision= 0.6683886877013445
  - Recall= 0.7197389648150766
  - f1= 0.693114039648153441
  - roc\_auc = 0.8734835137324075
- 

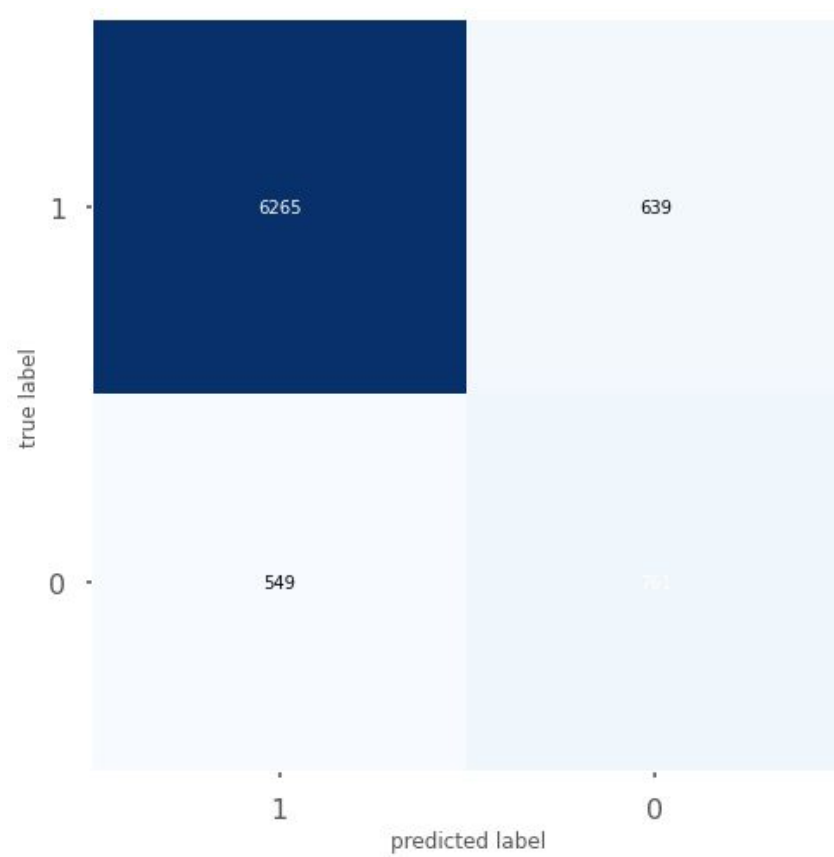
Accuracy





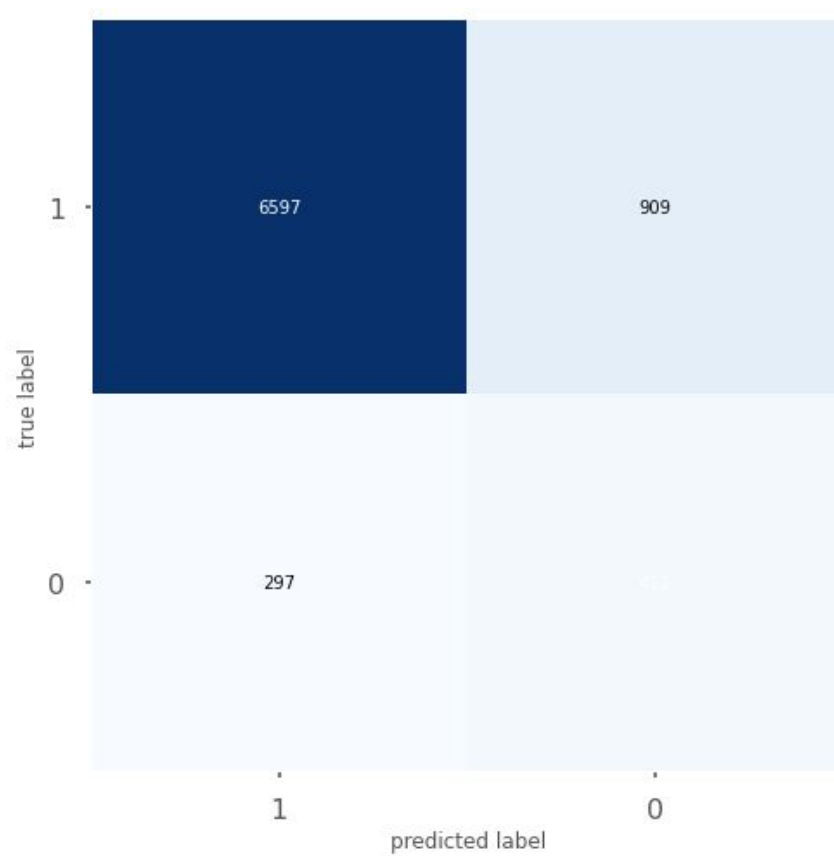
# Confusion Matrices

# Action

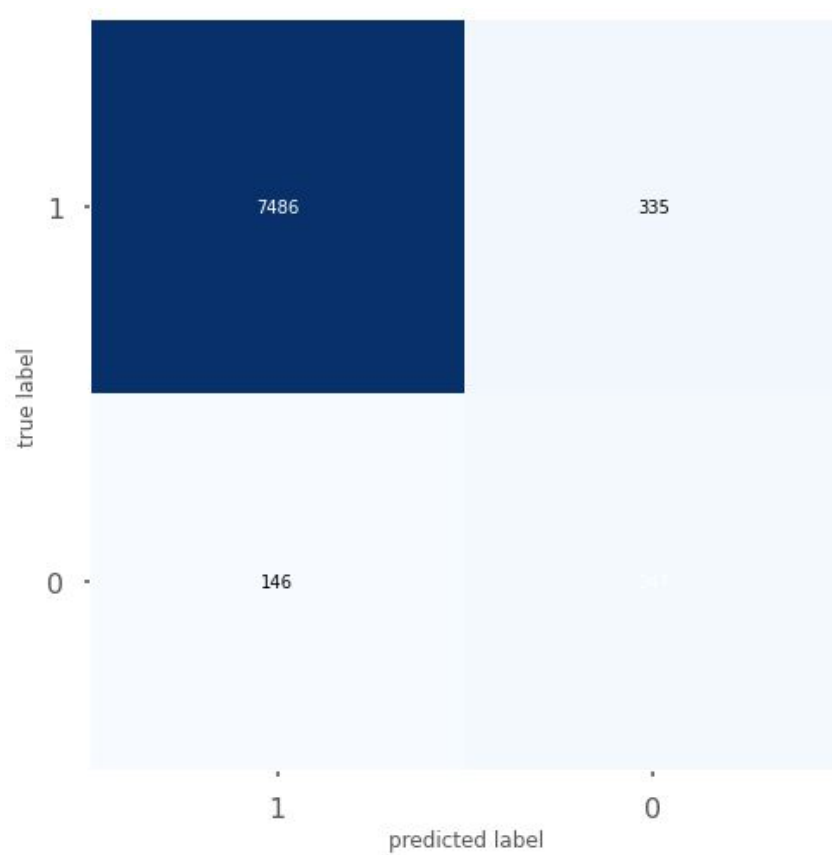




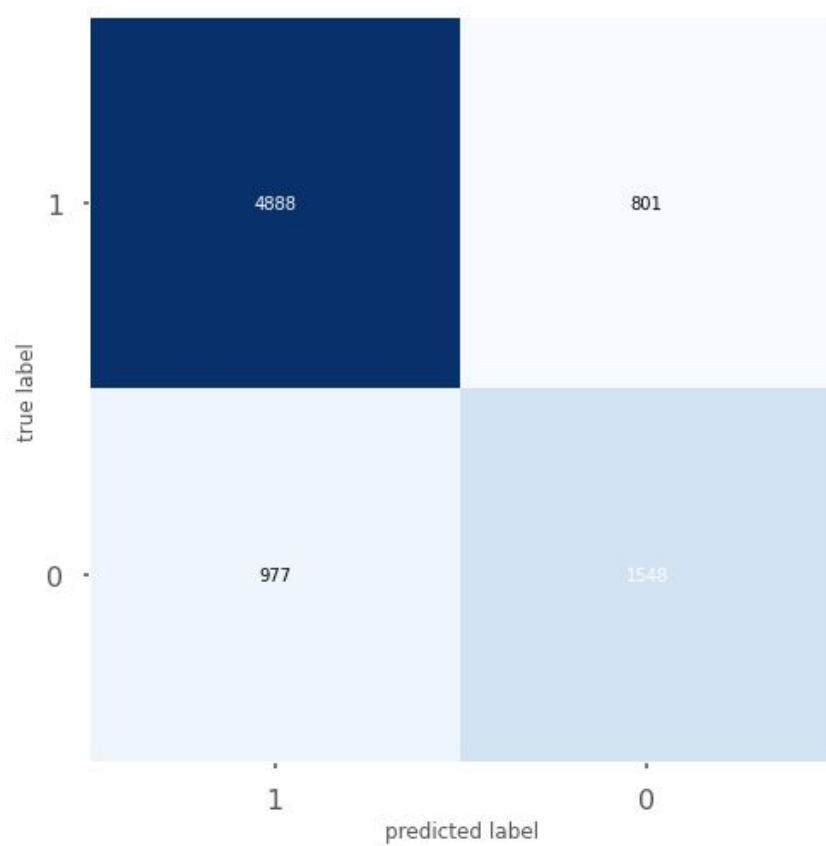
# Adventure



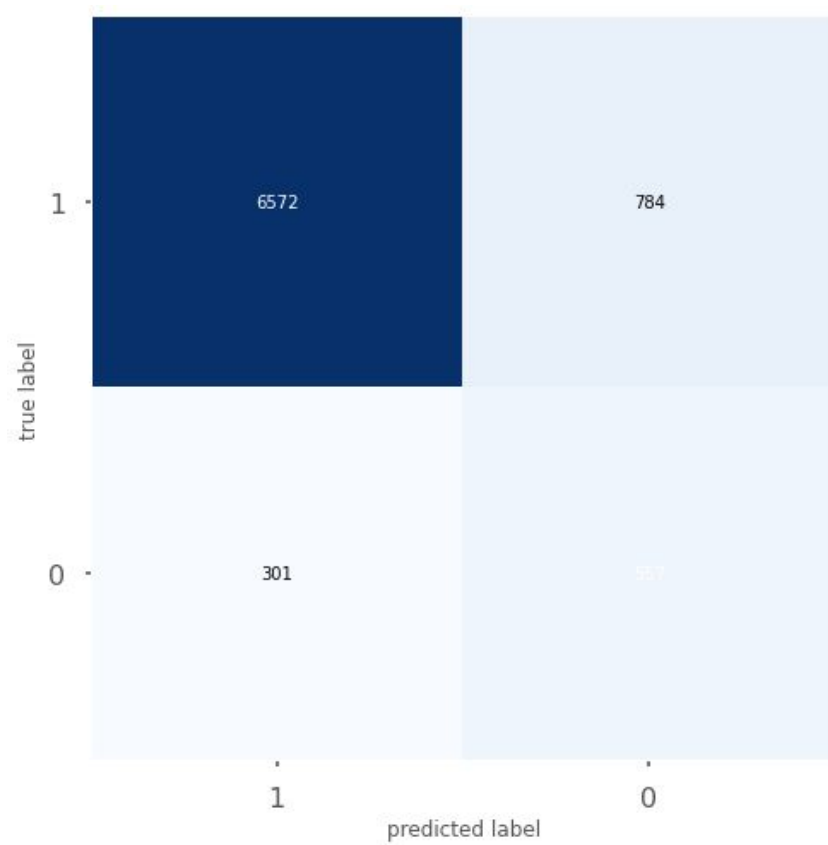
# Animation



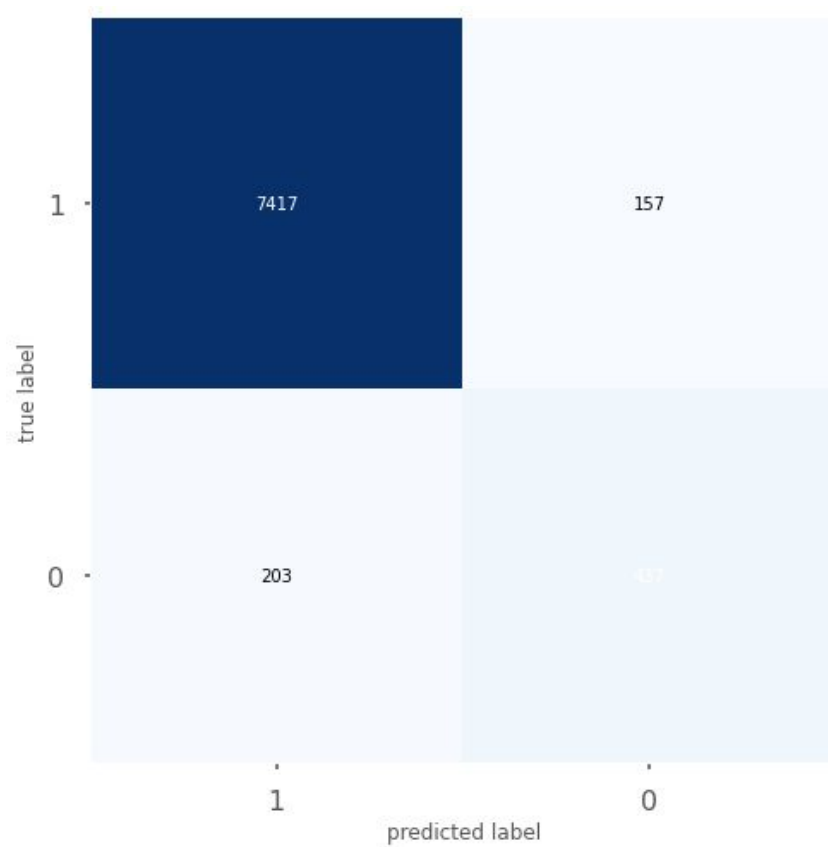
## Comedy



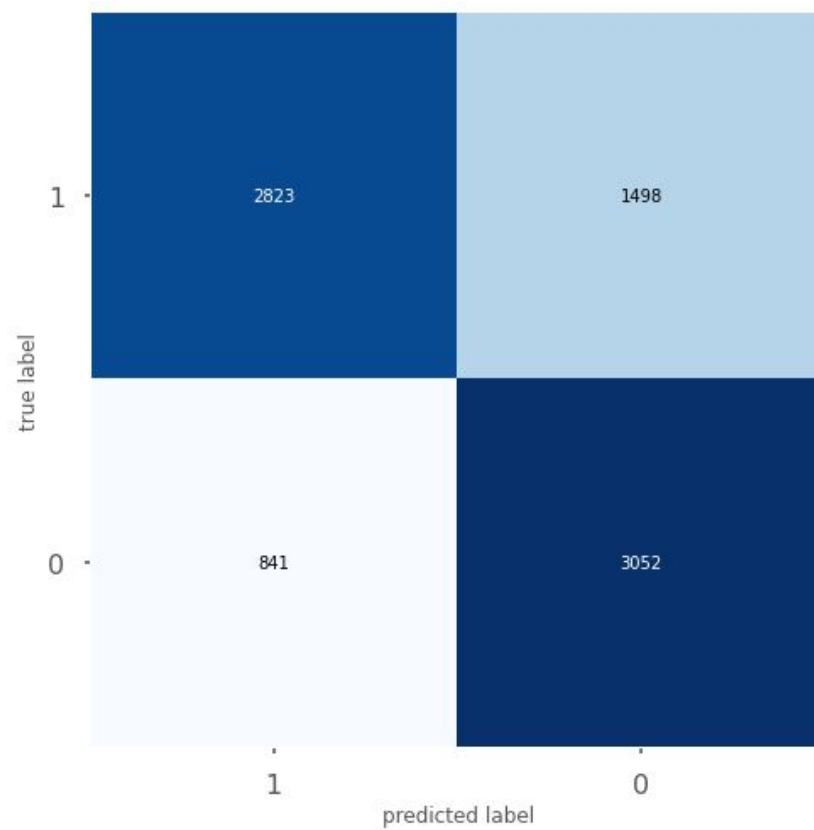
# Crime



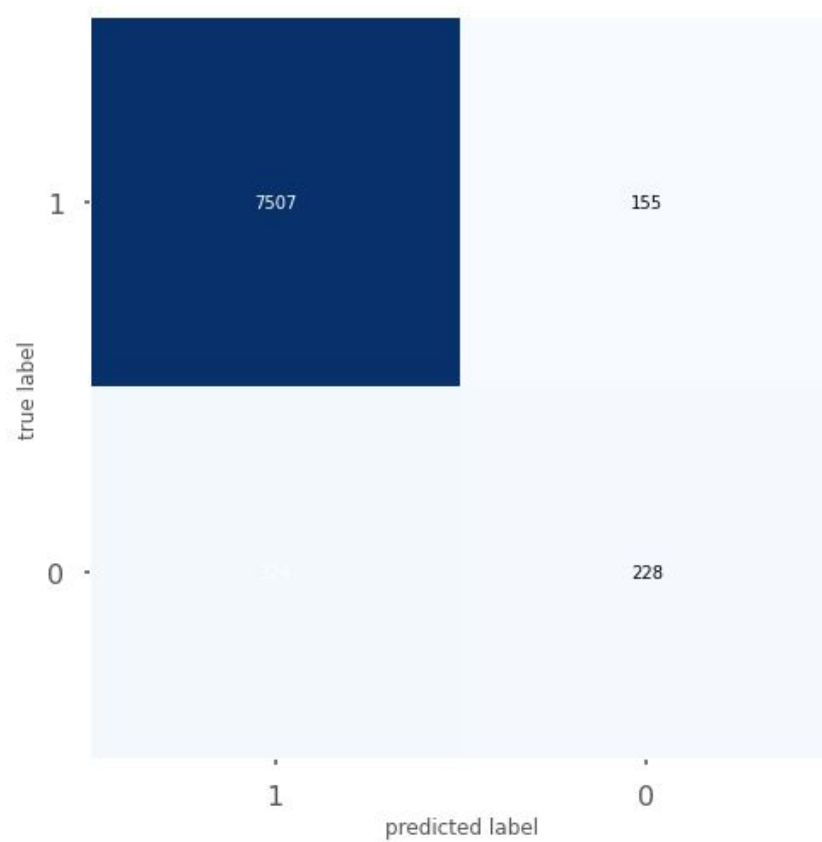
## Documentary



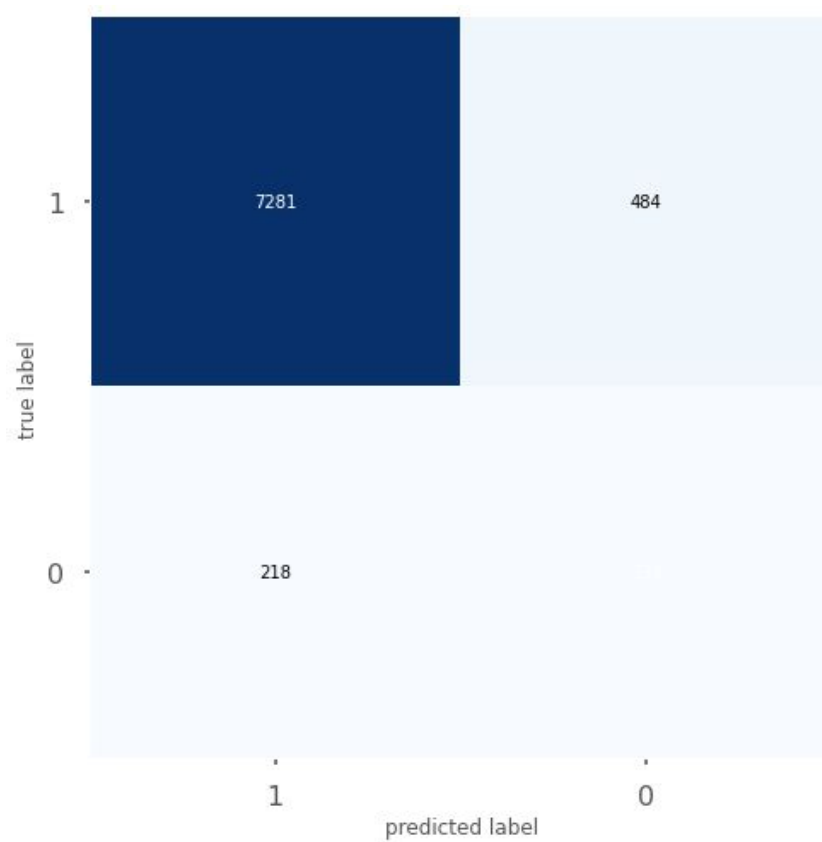
## Drama



# Family

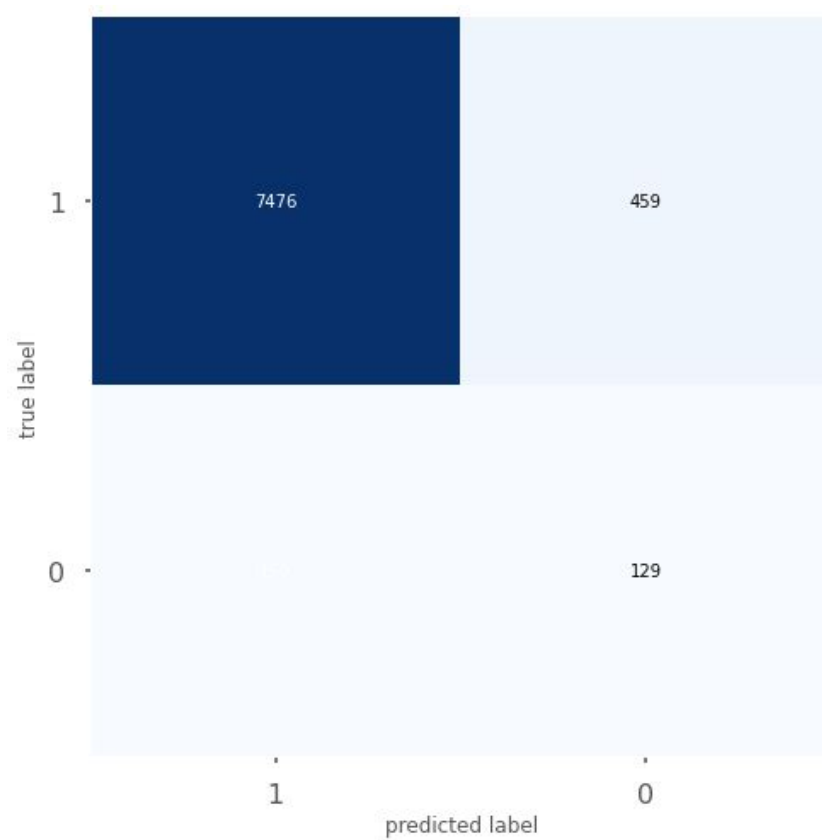


# Fantasy

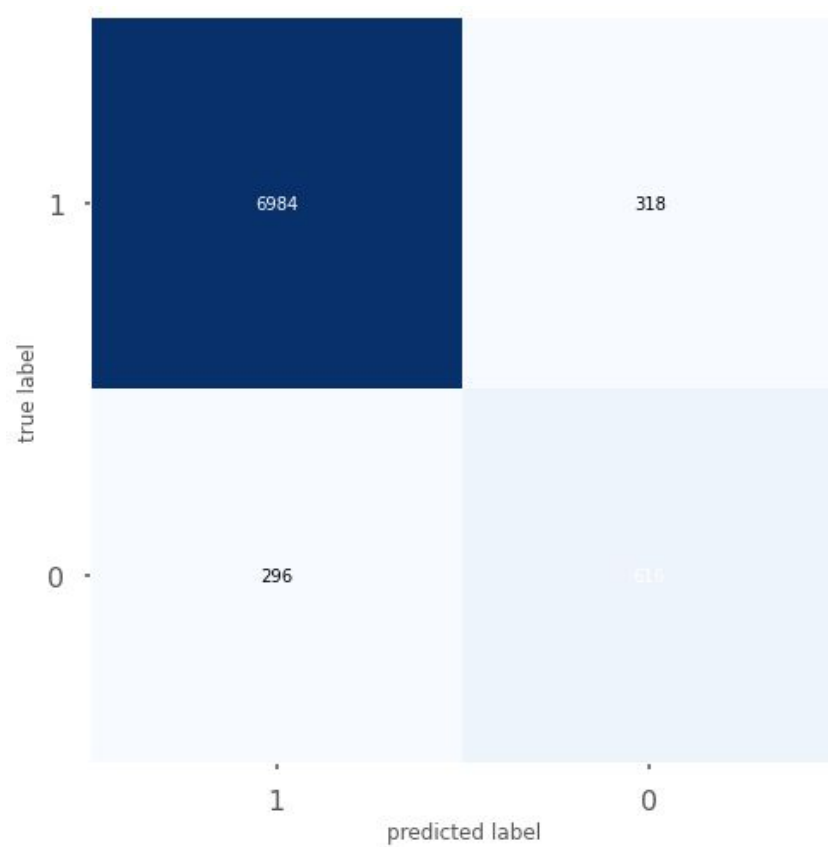




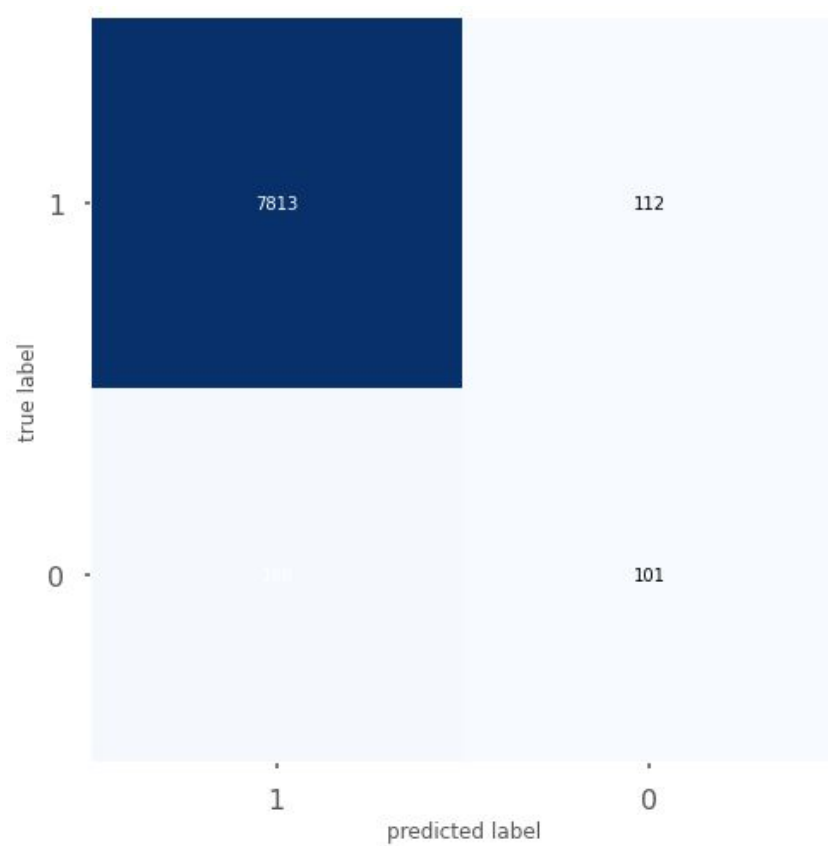
# History



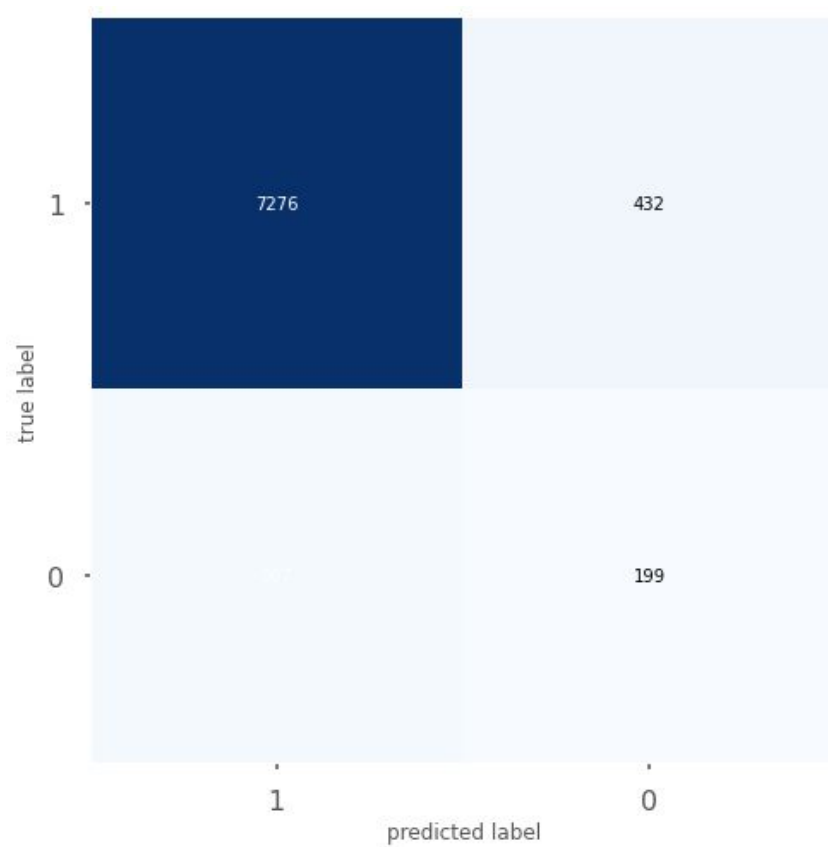
# Horror



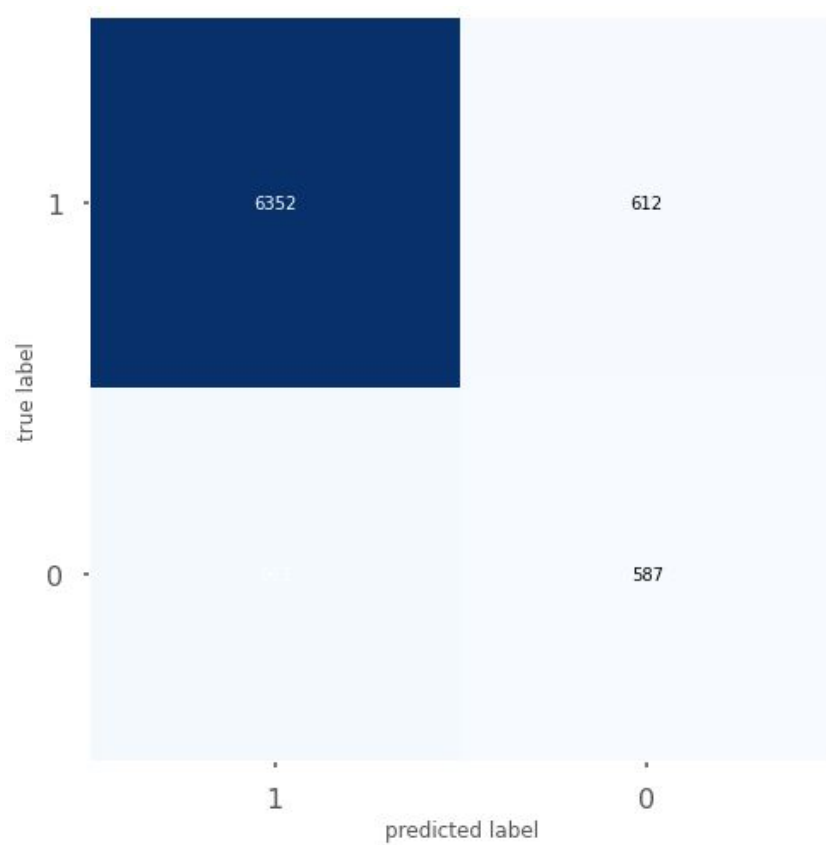
# Music



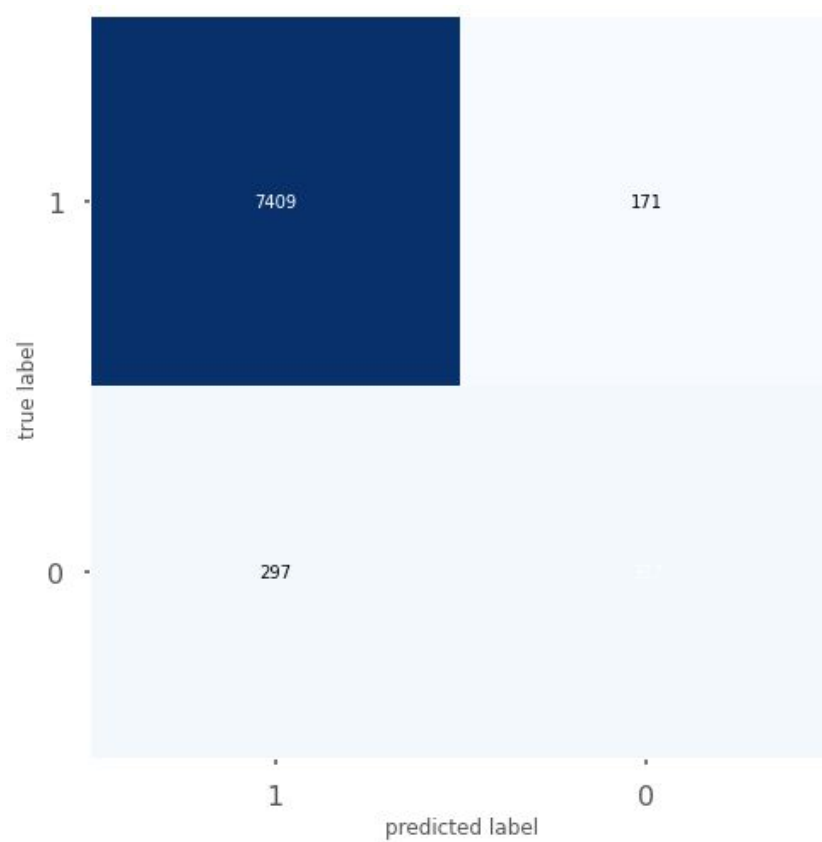
# Mystery



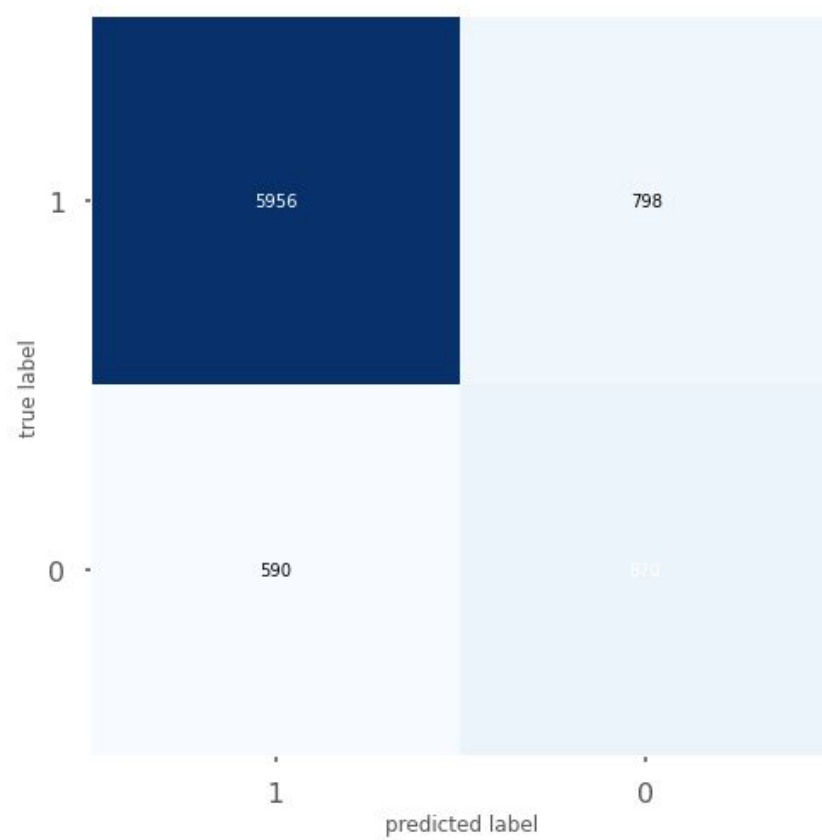
## Romance



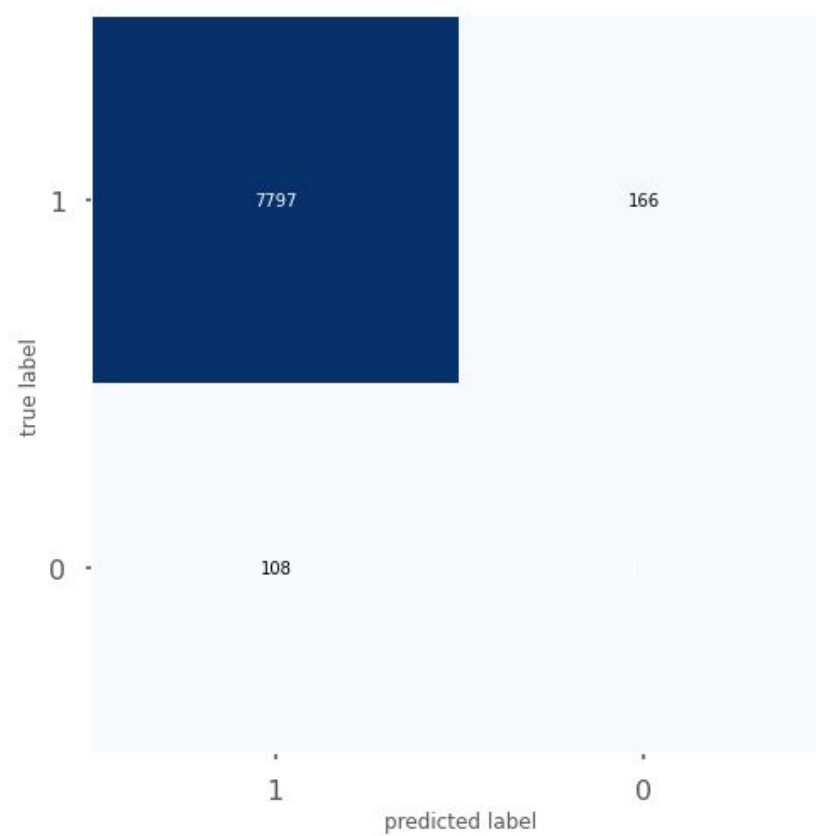
## Science Fiction



## Thriller

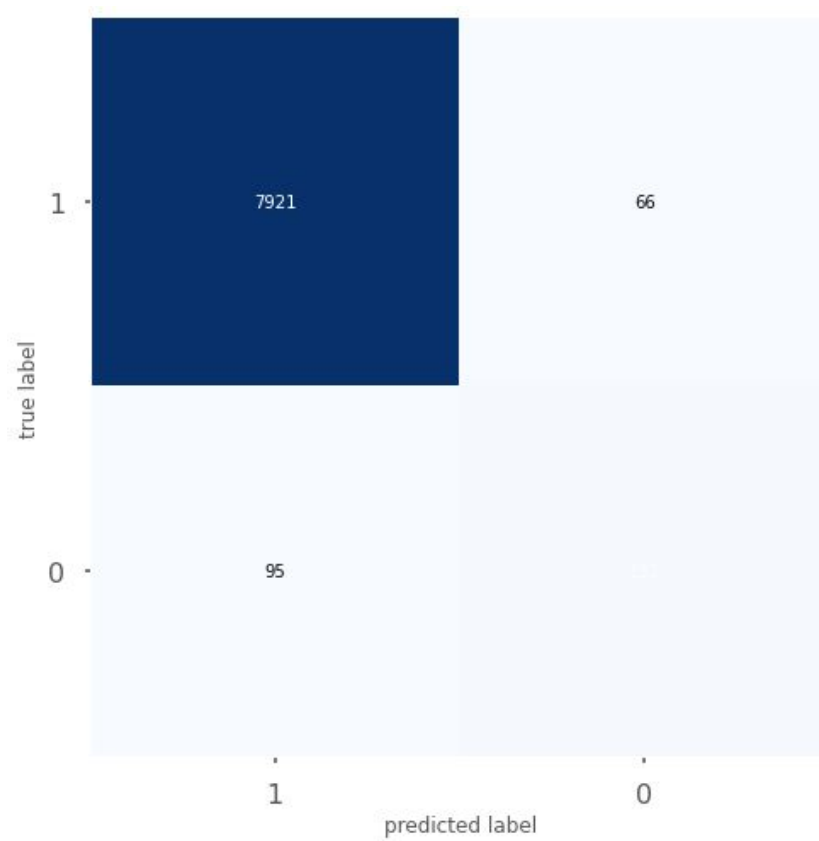


# War





## Western



# References

- <https://www.kaggle.com/rounakbanik/the-movies-dataset/kernels>
- [http://mlss.tuebingen.mpg.de/2015/slides/Fergus\\_1.pdf](http://mlss.tuebingen.mpg.de/2015/slides/Fergus_1.pdf)
- <https://www.themoviedb.org/movie/266856-the-theory-of-everything>
- <https://www.drivendata.co/blog/hateful-memes-benchmark/>
- <https://towardsdatascience.com/>
- <https://stats.stackexchange.com/questions/126238/>
- <https://www.superdatascience.com/>

