MACHINE LEARNING AND DEEP LEARNING (650.025) - TERM PROJECT PROPOSAL

Movie Genre Multi-label Classification

Mustafa Tahir BİNGÖL - 12231704
Mustafa Tayyip BAYRAM – 12237686

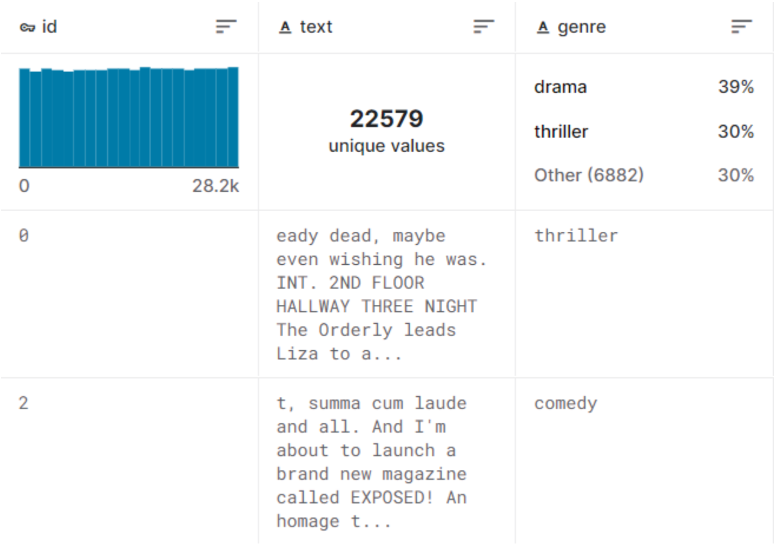
Table of Contents

Motivation

For movie viewers, the harmony of the specified genre of the movie and the content of the movie is extremely important. While the viewers want to watch a comedy movie, they never dream of encountering a drama movie. In this context, we aimed to solve this problem in our project by creating a model that works with high accuracy. The model we will obtain as a result of our work will determine the compatibility of the genre and content specified in the promotion of the film in a way that will satisfy all viewers.

Learning Task

The process of categorizing movies by their genre is a difficult classification challenge since genre is an intangible trait that cannot be immediately determined in any movie plot. Furthermore, movies might belong to numerous genres at the same time, making movie genre assignment a classic multi-label classification issue, which is substantially more difficult than ordinary single label classification.



id	text	genre
0 28.2k	22579 unique values	drama 39% thriller 30% Other (6882) 30%
0	eady dead, maybe even wishing he was. INT. 2ND FLOOR HALLWAY THREE NIGHT The Orderly leads Liza to a...	thriller
2	t, summa cum laude and all. And I'm about to launch a brand new magazine called EXPOSED! An homage t...	comedy

The data that found in Kaggle platform, is able to solve our problem. Besides, some data also can scrapped from IMBD. The data set contains just 1 column with the feature which describes the plot. The other columns corresponds to the possible 28 movie genres. These are basically our target columns. If the movie falls under a particular genre, then that particular column will indicate 1, else it will indicate 0. The data set only has one column with the characteristic that explains the plot. The remaining columns correspond to the potential all film categories. These are our main goal columns. If the film belongs to a specific genre, that column will show a 1, otherwise it will show a 0.

For example,

Plot Text = "In 1971, Carolyn and Roger Perron move their family into a dilapidated Rhode Island farm house and soon strange things start happening around it with escalating nightmarish terror. In desperation, Carolyn contacts the noted paranormal investigators, Ed and Lorraine Warren, to examine the house.”. Original label is Horror, but some words are also calling Thriller genre. Therefore, it is possible to predict it with Horror and Thriller.

Performance Measure

The data consists of highly imbalanced data. As it is multi-label classification, it is expected to get lower results than single label classification. Therefore, our metrics will be mainly F1-Score, Precision, Recall and Accuracy.

Precision is defined as the proportion of accurately predicted positive observations to all expected positive observations. A genre's precision is provided by

Precision (Genre=Action) = (Number of movies ‘correctly’ identified as Action Genre)/(Total number of movies that have been identified as Action Genre)

The ratio of accurately predicted positive observations to all observations in the actual class is defined as recall. A genre's recall is provided by

Recall(Genre=Action) = (Number of movies ‘correctly’ identified as Action Genre)/(Total number of Action Genre movies in the data set)

F1 Score is the Harmonic mean of Precision and Recall. F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Below is the method we use to come up with a single F1 score for our model performance

- Calculate Precision, Recall, and F1 Score for each genre.
- Calculate a weighted average of the F1 score based on the level of support (number of occurrences in the genre in our data set). This is our final metric.

Plan

First thing to do is EDA and Data Preprocessing after deciding which data is to be used. In Exploratory Data Analysis phase, data will be examined for understanding connections such as correlation between movie genres. For example, it is more usual to catch more connection between Horror-Thriller than Romance-Horror.

Data processing will consist of Removing non-proper keys (such as html tags, accented characters etc.), Lower Casing, Stemming or Lemmatization and Removing stop words. According to the situation, some other processes can be added.

Modelling phase will be followed after that, data splitting and model preparations (encoding) will be started, and it is also going to change the models that we selected. Encoding TF-IDF Vectorizer will be used. Logistic Regression and Linear Support Vector Machine Classifier will be following it as a beginning. According to our metric results, we will try other models and encodings.

Related Work

In literature review, we found many studies similar to ours. But most of the past work, unlike our work, was on image processing and making predictions from movie posters. The only work we think is similar to ours, “Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features” [x] was written in 2008 by Alex Blackstock and Matt Spitz.

In the related study, the targeted output coincides with the estimation of the genre of the films by taking the plots of the films, which is our targeted output. The main difference between this study and ours is that a single class estimate was obtained instead of multiple label estimates in the related study.

Risk Management

We tried to find more than one dataset against the problems that may be found in the dataset we use. At this point, we decided to work with 2 different datasets in order to minimize the risks arising from the dataset. In addition to these, we also applied the same method to solve problems that may arise from algorithms. We decided to work with 4 different machine learning algorithms and use the algorithm that gives optimum results.