International Journal for Modern Trends in Science and Technology, 6(11): 118-121, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061121
Available online at: http://www.ijmtst.com/vol6issue11.html

Movie Genre Prediction Based on Plot Synopsis

Jagjeet Singh[1] | Vibhor Sharma[2]

[1]B. Tech Student, Department of IT, Maharaja Agrasen Institute of Technology, Delhi, India
[2]Assistant Professor, Department of IT, Maharaja Agrasen Institute of Technology, Delhi, India

To Cite this Article
Jagjeet Singh and Vibhor Sharma, "Movie Genre Prediction Based on Plot Synopsis", *International Journal for Modern Trends in Science and Technology*, 6(11): 118-121, 2020.

Article Info
Received on 18-October-2020, Revised on 10-November-2020, Accepted on 19-November-2020, Published on 22-November-2020.

ABSTRACT

Movies have now become one of the main sources of entertainment for people. The extensive use of Internet has increased the creation and sharing of movie related data online. Movie plot summaries generally tell about the movie genres and many people read them before deciding to watch a movie. An automatic system can be applied to predict genres based on summaries. The objective dataset chosen by us consists of 14828 movies taken from Kaggle. We use different approaches such as TFIDF, Char gram, Skip gram etc to get better accuracy scores in predicting movie genre tags.

KEYWORDS: *Imdb, tfidf, char gram, skip gram, kaggle*

I. INTRODUCTION

Movie plot summaries usually reflects the different genres of movies such as horror, murder, romantic, comedy and people can know information about genre from the summaries of the movies. People normally read the summaries to get to know about the gist of the movie and its genres. That is why plot summaries are written in a way that it tells the reader about the genres of the movie.

The entertainment industries all around the world are making a lot of movies rapidly and attracting people of all ages. Nowadays we need an automated system to categorize the movie with the help of plot synopsis and movie title.

A genre prediction model could allow us to predict different genres of a movie that can be used by people to decide whether or not to watch a movie. To solve this problem, we have made a system that works on the IMDB movie data set taken from Kaggle and predicts the genres of a movie from its title and synopsis.

II. PREVIOUS WORK

Gabriel et al. [1] proposed a model which uses Convolutional Neural Network (CNN) for image processing in order to predict genre from movie trailers. They created a dataset containing trailers of movies and made it public and created a classification method using CNN architecture to classify movies on base of its genres.

Gabriel et al. [2] suggested to train a model that could learn different things about a movie poster and then predict the genre it represents. They used RESNET34 and a custom architecture in order to train the model. They models performed well in F-score metric and Top K Categorical Accuracy.

Quan [3] studied the problem of predicting genres from movie plot and used different methods like Recurrent Neural Networks and Word2Vec+XGBoost are used text classification, also K-binary transformation and probabilistic classification were employed to tackle the multi

label problem. He attained a high F-score of 0.56 along with a hit rate of 80%.

Haifeng et al.[4] suggested the use of a predictive model in order to find movie recommendation for users. They used a Gaussian kernel support vector machine(SVM) model along with a logistic regression model to extract features and compare them. They got an accuracy of 85% positive cases which increased to 93% with a smaller VC dimension and less over fitting.

You-Jin et al. [5] proposed an approach that was based on deep learning which utilized the ELMO embedding and sentiment scores of sentences in order to predict a movie's success only based on its plot.

Yin-Fu et al. [6] proposed a movie genre classification model based on audio and video features which used a meta-heuristic optimization algorithm called Self-Adaptive Harmony Search (i.e., SAHS) to select some features for corresponding genres. They reached an overall accuracy of 91.9%.

title	plot_synopsis	tags	clean_tags
I tre volti della paura	Note: this synopsis is for the orginal Italian...	cult, horror, gothic, murder, atmospheric	cult,horror,gothic,murder,atmospheric
Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...	violence	violence
The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...	romantic	romantic
Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...	inspiring, romantic, stupid, feel-good	inspiring,romantic,stupid,feel_good
Scarface	In May 1980, a Cuban man named Tony Montana (A...	cruelty, murder, dramatic, cult, violence, atm...	cruelty,murder,dramatic,cult,violence,atmosphe...

Figure 1: Dataset after cleaning tags

III. DATASET

The dataset has been taken from Kaggle which comprises of different information about movies from different sources. It contains the imdb id of the movie, title of the movie, plot synopsis, different genre tags and the synopsis source which is in this imdb for most of the movies. The imdb id is just a simple identifier for a movie. Title and plot synopsis being entirely textual contains the name and plot of the movie. Genres containing the different genres a particular movie belongs to. It is a multi-label classification problem as one movie can have more than one genre.

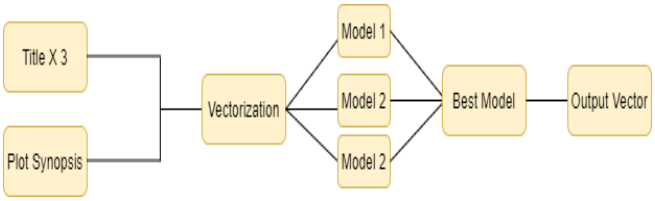

Figure 2: Complete Working

IV. EVALUATION METRICS

Most of the movies have more than one genre that is why the normal accuracy metric can't be used in this case. We would need to modify any metric we use slightly so that it can account for all the tags separately and not consider output as a fixed length sequence.

F1-Mean Score / F1- Score: F1 score is a measure of accuracy of test which is computed with the help of precision and recall of the test. It is basically the weighted average of the two. This can be extended by using weighted average over all classes in case of a multiclass setting. Its maximum value is 1 and minimum is 0.

F1 = 2 * (precision * recall) / (precision + recall)

F1-Micro: This scoring metric works for the multi label setting and also does well in case of class imbalance**.** It is the harmonic mean of micro-precision and micro-recall.

F1-Micro = 2 * (1 / (1/micro-precision + 1/micro-recall))

Hamming loss: It is the fraction of labels which are not correctly predicted. It can also be modified for multi label setting.

Precision: It is the ratio of true positives and sum of true and false positives.

Recall: It is the ratio of true positives and sum of true positives and false negatives.

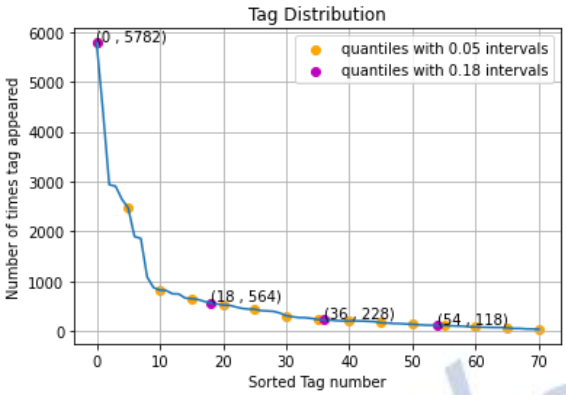

Figure 3: Tag number VS Tag frequency

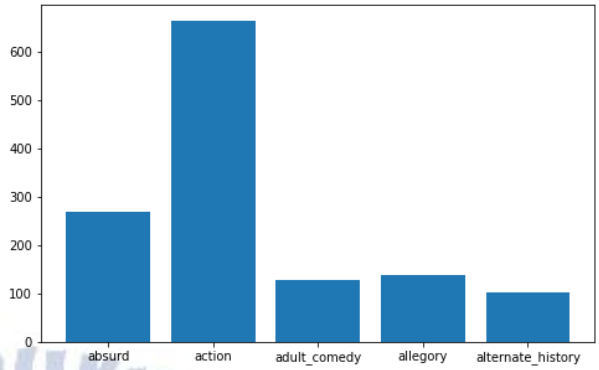

Figure 4: Tag names VS Tag Frequency for some tags

V. EXPERIMENTAL SETUP

The total number of distinct genres was found to be 71. So, the output vector would be of 71 dimensions. This means 71 models are needed to be trained.

It is obvious from figure 3, that the tags in first 18th percentile occurred more frequently, however, after that the frequency of others tags decreases exponentially. Therefore there are only some tags which are highly likely to appear. The dataset contains 14828 movies which needs to be divided into training and testing sets.

We have come up with a way to vectorize the information so it can be understood by the classification. We have used the bag of words model approach. In this we count the occurrence of different words in the plot of a movie and then represent it in form of a vector. Once the vectors are generated, different methods are applied on it in order to get the maximum f1 score for predicting genres of a movie.

Different methods used in training the models were Tfidf, Char grams, Skip grams, Topic modelling and different combination of the methods mentioned above.

VI. RESULTS

In this experiment to predict the genres associated with a particular movie, under different constraints, with respectable metric scores, the results are shown in the table below. It can be seen that the traditional models such as SGD (Logloss) and Logistic Regression showed pretty good results with Tfidf featurization giving the best micro f1 score of 0.3137 which is pretty decent for a multi class problem like this. Tfidf performed much better than the general Bag of words approach. Skipgrams method also gave a decent score but when we combined different methods giving the best scores, we didn't get a very good result.

S.no	Model	Feature	Test Micro f1	Train Micro f1
1	SGD(LogLoss)	Tfidf	0.312	0.828
2	Logistic Regression	Tfidf	0.311	0.976
3	SGD(Hinge Loss)	Tfidf	0.3117	0.894
4	SGD(LogLoss)	Tfidf(2 grams)	0.3137	0.7899
5	SGD(LogLoss)	Tfidf(2 grams)	0.312	0.671
6	SGD(LogLoss)	Skip grams	0.303	-
7	SGD(LogLoss)	Tfdif(1-2grams) + Char Grams (3,4) + Skip Grams (2)	0.27	-

VII. CONCLUSION

We have presented an automatic genre classification mechanism for movies based on its plot synopsis. We have discussed the implications of different techniques that are involved in building such system in order to accommodate a multi-label multi class output and also work in limited resources available to us. We have also discussed

why is it important to choose the correct metric system in order to get great results. For the future work we have thought of using deep learning approach in order to enhance our results and get the best output possible.

REFERENCES

[1] Gabriel S. Simoes, Jônatas Wehrmann, Rodrigo C. Barros, Duncan D. Ruiz. "Movie Genre Classification with Convolutional Neural Networks". In 2016 International Joint Conference on Neural Networks (IJCNN).

[2] Gabriel Barney and Kris Kaya ."Predicting Genre from Movie Posters".

[3] Quan Hoang ."Predicting Movie Genres Based on Plot Summaries".

[4] Haifeng Wang and Haili Zhang. "Movie Genre Preference Prediction Using Machine Learning for Customer-Based Information". In 2018 IEEE Annual Computing and Communication Workshop and Conference (CCWC).

[5] You-Jin Kim, Jung-Hoon Lee and Yun-Gyung Cheong. "Prediction of a Movie's Success From Plot Summaries Using Deep Learning Models".

[6] Yin-Fu Huang and Shih-Hao Wang. "Movie Genre Classification Using SVM with Audio and Video Features". In 8th International Conference, AMT 2012.