International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1344-1348 www.ijaem.net ISSN: 2395-5252

Movie Genre Prediction Using Deep Learning

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*

Submitted: 01-06-2021 Revised: 14-06-2021 Accepted: 16-06-2021

ABSTRACT : Movies are now one of the major sources of entertainment for people. The increase in usage of Internet has also increased the creation and sharing of data related to movies online. Movie synopsis usually gives us an idea about the genres of movie and people read them before deciding to watch the movie. An automatic system can be used to predict it. The dataset of IMDB is used in this project which contains about 14000 movies. Different techniques are used such s Bag of words model, Char gram, TFIDF, Deep Learning and Topic Modeling.

Keywords: Imdb, tfidf, char gram, skip gram, kaggle

I. INTRODUCTION

Movie synopsis generally tells us about the different genres that the movie corresponds to, such as romantic, comedy, murder, horror etc. People get to know the information about the movie by reading its synopsis. People read summaries to get the idea of what the movie is about and whether they should watch it or not. That is the reason that synopsis of the movies are written in a way such that it tells the reader about the genres of the movie.

Nowadays, the entertainment industry is making a lot of movies around the world so as to attract people of all ages. We would require an automated system that will be able to categorize movies based on its synopsis and title.

A genre prediction model will be able to predict different genres of movie which people can use in order to decide whether they want to watch the movie or not. To solve this problem, we have made a system that words on a dataset of IMDB taken from Kaggle and will be able to classify movie genres.

II. PREVIOUS WORK

Gabriel et al. [1] suggested a model that uses Convolutional Neural Network (CNN) to process image to predict the type from movie trailers. They have created a database containing movie trailers and made it public and created a classification method using the CNN structure to classify movies on the basis of their genres.

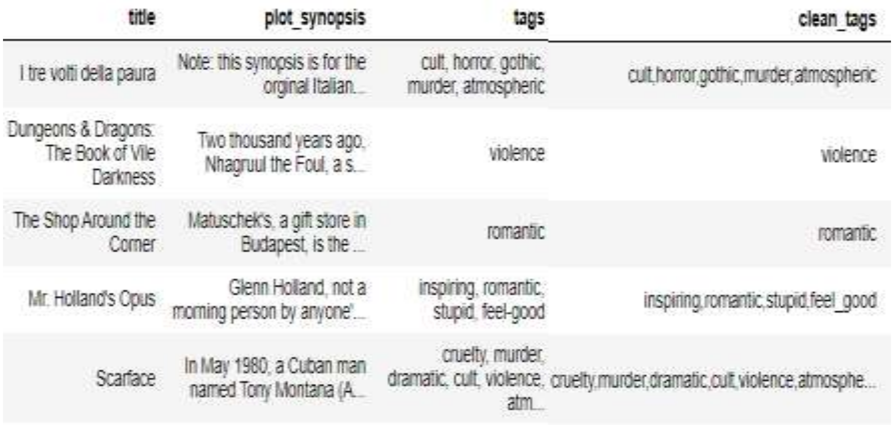
Gabriel et al. [2] suggested training a model that could learn different things about a movie poster and then predict the genre it represents. They used RESNET34 and a custom architecture in order to train the model. They models performed well in F-score metric and Top K Categorical Accuracy.

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] projected AN approach that was supported deep learning that used the ELMO embedding and sentiment millions of sentences so as to predict a movie’s success solely supported its plot.

Yin-Fu et al. [6] projected a pic genre classification model supported audio and video options that used a meta-heuristic optimisation algorithmic rule known as Self-Adaptive Harmony Search (i.e., SAHS) to pick out some options 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 used in this project is taken for Kaggle which contains different information about movies taken from different sources. It contains the title of the movie, plot synopsis, genre tags and the imdb id of movie. The ID is just a simple identifier for a movie. Title and plot summaries being textual information contains name and plot of movie. Genres tag tells us about the different genres of the movie. The problem of this project is a multi-label classification problem as one movie can have more than one genre.

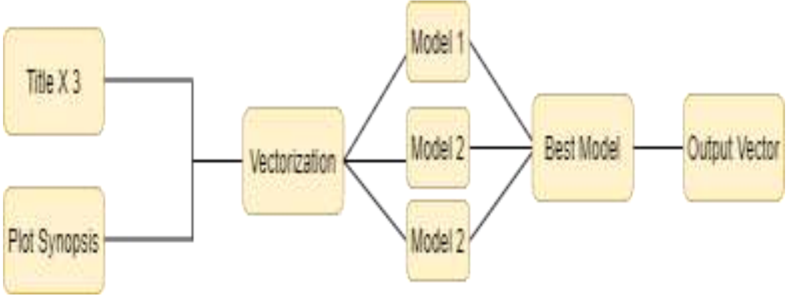


Figure 2: Complete Working.

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.

Most of the movies usually have multiple genres which make it harder to measure accuracy and is the main reason that why we can't use the normal accuracy metric in this case. We will have to modify the metric so as to account for all the different genre 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**.** Harmonic mean of micro-precision and micro-recall gives us the F1-Micro score.

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

Hamming loss: It is the fraction of labels that don't seem to be properly foretold. It also can be changed for multi label setting.

Precision: Ratio of true positives and sum of true and false positives gives us precision.

Recall: Ratio of true positives and sum of true and false negatives gives us recall.

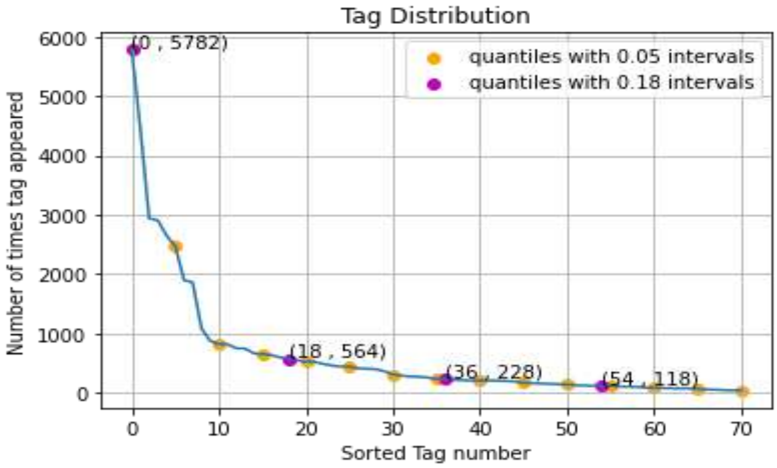


Figure 3: Tag Number vs Tag Appearance

V. EXPERIMENTAL SETUP

Some of the Techniques Used-

Topic Modeling - Topic modeling is an unsupervised machine learning technique which scans a set of documents and detects word and phrase patterns in them. Then it automatically forms different sets and clusters of words and similar expression that best characterizes a set of document.

Deep Learning- Deep Learning is a type of Machine Learning which is inspired by the structure of a human brain. This algorithm tries to draw similar conclusions as a human would by analyzing data with a logical structure. It uses a multi-layered structure of algorithms called neural network.

Experimental Information –

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.

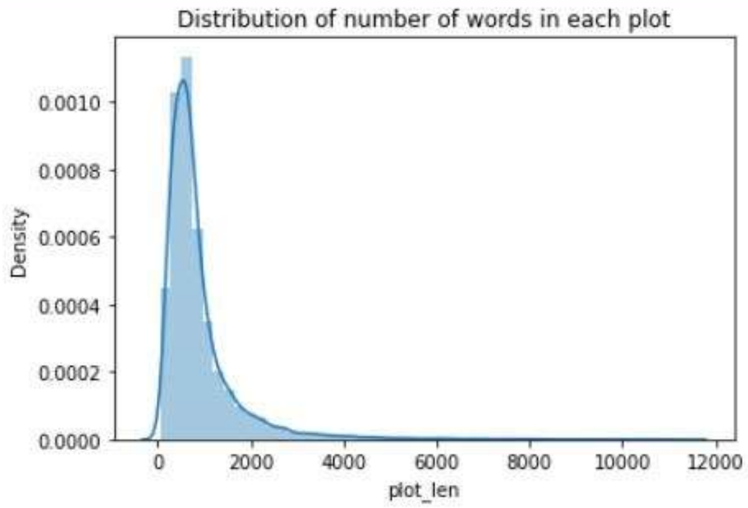


Figure 4: Distribution of no. of words in each plot

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.

Then we used Deep Learning model on Title of the movie and the summary of the movie and we got the highest score of 0.4545 which was a significant increase from the scores we got from previous techniques.

S.no	**Model**	**Featurization/Features**	**Test Mico 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(2grams)	0.3137	0.7899
5	SGD(Logloss)	Tfidf(2grams)	0.312	0.671
6	SGD(Logloss)	Skipgrams	0.303	-
7	SGD(Logloss)	Tfdif(1-2grams) + Char Grams (3,4) + Skip Grams (2)	0.27	-
8	Deep	Title and Synopsis	0.454	0.51

	Learning Model		5	
9	Bert Model	Title and Synopsis	0.31	-

VII. CONCLUSION

We have introduced a way to distinguish automatic genres of movies based on its subtitle. We have discussed the implications of the various strategies involved in building such a system in order to meet the release of multiple labels and to process the limited resources available to us. We also discussed why it is important to choose the right metrics to get the best results. We also implemented in-depth deep learning approach to improve our results and achieve a better outcome.

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.