Movie Genre Classification Using Movie Summaries

Joey He
August 2024

Abstract

This research paper explores different methodologies for movie genre classification using movie summaries. I use several machine learning techniques, including Neural Networks (NNs), Convolutional Neural Networks (CNNs), and Bidirectional Encoder Representations from Transformers (BERT) then employing the use of multi-label multiclass classifications to allow for classifications for multiple genres, along with plot summaries for model improvement. Results show that using a summarization + Bert approach significantly outperforms baseline neural networks. Findings from this research can be used in many other text classification applications for model improvement.

1.1 Introduction

This study aims to develop an effective deep-learning model that can accurately classify the genre of a movie based on its Wikipedia summary, typically around 600 words in length. Accurate genre classification is essential not only for the entertainment industry but also for information systems, where the correct categorization of content is crucial. Mislabeling of movie genres by journalists or distributors can lead to misleading audience expectations and impact the overall viewing experience. Additionally, subgenres, which often provide a more granular description of a film's content, are frequently overlooked. A robust multi-label classification model can address these challenges by correctly identifying both primary and sub-genres, thus enhancing the richness of metadata in film cataloging systems like IMDb.

The significance of this research extends beyond the realm of movie classification. The methodologies and techniques explored can be adapted to other domains where text classification is critical, such as news categorization, sentiment analysis, and content filtering. This research aims to contribute to the field of text classification by demonstrating the advantages of combining summarization techniques with advanced machine learning models.

1.2 Background

Movie genre classification poses a unique challenge due to the inherent ambiguity and overlapping characteristics of genres. Unlike traditional classification problems where categories are distinct and mutually exclusive, movies often encompass a blend of multiple genres.

I begin my movie classification by finding the most suitable model for single-class classification, using validation accuracy as my comparative metric. Upon discovering this model, I implement a multiclass classification method from a similar EU Legal Document Classification problem [2].

Approach - Single Label Classification

2.1 Data

For this study, I utilized two distinct datasets: one containing movies and their corresponding plot summaries, and another listing movies alongside their associated genres. By performing an inner join on the movie titles, I combined these datasets into a unified dataset, ensuring each record contained both plot and genre information. To streamline the initial model development and focus on identifying the best-performing model, I opted to prioritize the most accurate genre association by selecting the first genre listed for each movie. This approach simplifies the multi-genre complexity by creating a new column for this primary genre, thus treating it as a single-label classification problem for the time being.

The resulting dataset comprises approximately 20,000 records. I then split this dataset into a training set and a test set, with 16,000 records allocated for training and 4,000 reserved for testing. I tokenize the texts into embeddings and begin training

2.2 Baseline

For my baseline, I use a simple averaging network. The Averaging Network calculates the average of word embeddings from movie plot summaries, producing a vector for genre classification. In 20 epochs, it achieved a validation accuracy of 0.358.

2.3 Modeling

I explore some other modeling options

The Dan Network enhances the Averaging Network by using multiple fully connected layers after averaging word embeddings. This deeper architecture allows it to capture more complex patterns in the data, improving genre classification. In 20 epochs, it achieved a validation accuracy of 0.389.

The CNN Network applies Convolutional Neural Networks to detect local patterns in plot summaries. It uses convolutional layers to identify key phrases and pooling layers to distill these features, making it effective for capturing genre-specific nuances [5]. In 5 epochs it achieved a validation accuracy of 0.447.

BERT (Bidirectional Encoder Representations from Transformers) uses a transformer-based architecture to understand context in plot summaries [3]. It processes the entire text bi-directionally, allowing for highly accurate genre classification by capturing nuanced contextual information. In 3 epochs, it achieved a validation accuracy of 0.6296, aligning with research [3] that BERT is the most suitable for this type of classification problem.

2.4 Results

	precision	recall	f1-score	support
Action	0.71	0.75	0.73	1375
Adventure	0.47	0.30	0.37	408
Animation	0.67	0.46	0.54	113
Biography	0.79	0.50	0.61	155
Comedy	0.65	0.81	0.72	877
Crime	0.71	0.63	0.66	493
Drama	0.61	0.59	0.60	694
Family	0.00	0.00	0.00	32
Fantasy	0.50	0.03	0.05	36
Film-Noir	0.00	0.00	0.00	6
History	0.00	0.00	0.00	7
Horror	0.44	0.75	0.56	182
Music	0.00	0.00	0.00	4
Musical	0.00	0.00	0.00	15
Mystery	0.07	0.10	0.09	39
Romance	0.21	0.15	0.18	73
Sci-Fi	0.21	0.23	0.22	13
Thriller	0.11	0.12	0.12	40
War	0.00	0.00	0.00	3
accuracy			0.63	4565
macro avg	0.32	0.28	0.29	4565
weighted avg	0.62	0.63	0.61	4565

The model achieved varying levels of performance across different genres, with notable precision and recall in certain categories. For example, the **Action** genre showed a precision of 0.71 and a recall of 0.75, resulting in an F1 score of 0.73. **Comedy** also performed well, with a precision of 0.65 and a recall of 0.81, leading to a 0.72 F1 score. However, some genres, such as **Family**, **Fantasy**, **Film-Noir**, and **History**, had poor or zero scores due to insufficient data, reflected in the macro average F1-score of 0.29. The weighted average F1-score stood at 0.61, with an overall accuracy of 0.63 across the test set.

These results indicate that the model performs better with genres that have ample data, achieving higher precision and recall. It's important to note that movies often belong to multiple genres, so some films classified as incorrect may still fit into other appropriate categories, suggesting potential areas for refinement in multi-label classification handling.

Approach - Multi-label classification

3.1 Data

I used the same two datasets as before but excluded columns with low numbers of records, such as **Game-Show**, **Adult**, and **Reality-TV** genres, to focus on more substantial categories.

```
Drama          11818
Action          7013
Comedy          6590
Romance         5697
Crime           5291
Adventure       4051
Thriller        3200
Horror          2104
Mystery         1905
Family          1591
Fantasy         1317
Sci-Fi          1141
Biography        925
Music            795
Musical          771
War              733
History          704
Animation        658
Film-Noir        607
Sport            420
Western          301
Game-Show          1
Adult              1
Reality-TV         1
Name: count, dtype: int64
```

	Plot	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Family	Fantasy	...	Horror	Music	Musical	Mystery	Romance	Sci-Fi	Sport	Thriller	War	Western
0	In 2154, humans have depleted Earth's natural ...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Albus Dumbledore, Minerva McGonagall, and Rube...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	In the mid-21st century, crop blights and dust...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	In the mid-21st century, crop blights and dust...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	As punishment for a past rebellion, the 12 dis...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

[5 rows x 22 columns]

I created a data frame containing movie descriptions and implemented one-hot encoding [1] for the remaining 21 genres, enabling a multi-label classification framework. By the nature of the data, each movie is assigned exactly 3 genres. Additionally, I generated embeddings for the plot summaries, transforming textual data into numerical representations that the machine learning models could utilize for genre prediction.

3.2 Modeling

I trained my BERT model using one-hot encodings for the genres as the target variable (y). Instead of the traditional binary cross-entropy loss, I utilized hamming loss, which is more suitable for multi-label classification tasks [4].

Hamming loss is a metric used in multi-label classification to measure the fraction of incorrect labels. It calculates the number of times the predicted labels differ from the true labels, including both false positives and false negatives [4]. The formula for hamming loss is:

$$Precision(P) = \frac{1}{n} \sum_{i=1}^{n} \frac{\left|(y^{(i)} \wedge \hat{y}^{(i)}\right|}{\left|\hat{y}^{(i)}\right|}$$

Where,

$n \implies$ Number of training examples

$y^{(i)} \implies$ true labels for the ith training example

$\hat{y}^{(i)} \implies$ predicted labels for the ith training example

A lower hamming loss indicates better model performance

3.3 Results

This model resulted in a validation accuracy of 0.3067

As the plot summaries average around 350 words, it became apparent that it was possible that too much context would confuse the model. I switched to a different dataset with shorter plot descriptions, averaging around 60 words. This adjustment improved the model's efficiency and led to a slightly higher validation accuracy of 0.3389, indicating that concise summaries can still effectively capture essential information for genre classification.

The lower validation accuracy in multi-label classification compared to single-label classification can be attributed to several factors. Firstly, there are more categories to consider, increasing the complexity of the prediction task. Secondly, for a prediction to be considered correct in multi-label classification, the model must correctly identify all relevant genres for a given instance. This contrasts with single-label classification, where only one correct label is needed, making the task inherently simpler. These additional challenges contribute to the observed difference in accuracy between the two types of classification.

Results

In single-label classification, the BERT model outperformed other models significantly, demonstrating its superior capability in understanding context and semantics. For the multiclass task, I successfully developed a BERT model optimized using hamming loss, which helped in managing the complexities of multi-label predictions. By switching to a dataset with summarized plot descriptions, I achieved a slight increase in accuracy. However, the potential for further improvements remains evident. With access to a larger and more balanced dataset of movie genres, I believe that the accuracy could be further enhanced, providing more reliable genre classification.

References

[1] F. Zeidi, M. Amasyali, and Ç. Erol. LegalTurk Optimized BERT for Multi-Label Text Classification and NER, June 30 2024

[2] A. Fayad. Fine Tuning BERT for a Multi-Label Classification Problem on Colab, Nov 5, 2023

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[4] G. Wu, J. Zhu. Multi-label classification: do Hamming loss and subset accuracy really conflict with each other?,

[5] M. Amin, N. Nadeem. Convolutional Neural Network: Text Classification Model for Open Domain Question Answering System

[6] J. George. An Introduction to Multi-Label Text Classification, Nov 12 2020