Projeto Linguagem Natural

1 ID Grupo

Grupo 55

Filipe Resendes 96859

2 Modelo

Neste trabalho, utilizei um modelo de classificação de rótulo único baseado em Regressão Logística para prever o género de filmes a partir dos resumos das suas sinopses. A abordagem escolhida foi de One-vs-Rest (OvR), o que significa que o modelo treina um classificador para cada género e, no final, seleciona o género com a maior probabilidade para cada sinopse, garantindo que apenas um género seja atribuído a cada filme.

O modelo foi treinado com um conjunto de 8041 sinopses de filmes, cada uma associada a um único género. Após o treino, para cada sinopse no conjunto de teste, o modelo prevê as probabilidades de cada género e seleciona aquele com a maior probabilidade, assegurando que apenas um género é atribuído a cada filme.

Esta abordagem foi inspirada no artigo "Predicting Movie Genres Using NLP and Multi-label Classification", disponível em Analytics Vidhya, adaptando a técnica para um cenário de rótulo único.

Pré-processamento

1. Limpeza de Texto: As sinopses foram processadas para remover pontuação, números e caracteres especiais, utilizando expressões regulares. Apenas foram mantidas as letras, e todo o texto foi convertido para minúsculas para garantir a consistência.
2. Remoção de Stopwords: Foram eliminadas stopwords comuns em inglês (como "the", "is", "and") utilizando a biblioteca `nltk`. Este passo ajuda a reduzir palavras irrelevantes e a melhorar a relevância das características extraídas.
3. Tokenização e Normalização: Após a limpeza, o texto foi dividido em palavras individuais (tokens), e espaços em branco excessivos foram removidos.

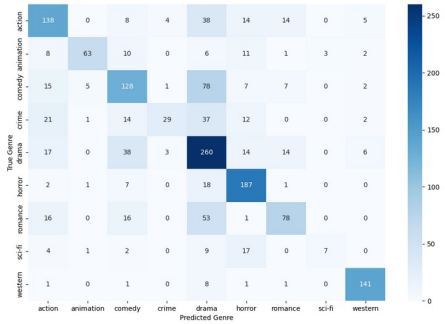
3 Configuração Experimental e Resultados

Para testar o meu modelo retirei amostras do conjunto de treino para criar conjuntos de teste personalizados, conforme solicitado. Os conjuntos de teste foram criados removendo 1%, 5%, 10% e 20% do conjunto de treino para avaliação, mantendo os restantes dados para treino. Dessa forma, foram usados quatro conjuntos de teste diferentes com 80, 400, 800 e 1600 resumos, respetivamente.

Os resultados obtidos em termos de acurácia geral e F1 score foram os seguintes:

Percentagem do Conjunto de Teste	Accuracy	F1 Score Macro	F1 Score Weighted
1%	59.26%	0.5161	0.5668
5%	65.26%	0.6023	0.6439
10%	62.90%	0.5823	0.6193
20%	64.08%	0.5999	0.6292

Como esperado, à medida que o tamanho do conjunto de teste aumentava, os resultados tornaram-se mais consistentes, com uma acurácia máxima de 65.26% obtida com 5% dos dados de teste.



Para uma visão mais detalhada dos erros de classificação, foi gerada uma matriz de confusão que mostra as previsões corretas e incorretas para cada género. A matriz de confusão ajuda a visualizar quais os géneros que são mais confundidos entre si, o que pode indicar oportunidades de melhoria no modelo ou na representação dos dados.

Genre	Precision	Recall	F1-Score	Support
Action	0.62	0.62	0.62	221
Animation	0.89	0.61	0.72	104
Comedy	0.57	0.53	0.55	243
Crime	0.78	0.25	0.38	116
Drama	0.51	0.74	0.61	352
Horror	0.71	0.87	0.78	216
Romance	0.67	0.48	0.56	164
Sci-Fi	0.70	0.17	0.28	40
Western	0.89	0.92	0.91	153

Os géneros mais difíceis de classificar corretamente foram a comédia e o drama. Isto sugere que o modelo tem dificuldade em reconhecer estes géneros, possivelmente devido á ambiguidade dos termos mais comuns nesses generos se sobreporem aos dos mais comuns.

4 Discussão

O modelo atingiu um desempenho global satisfatório, com um f1-score ponderado de 0.63 e precisão variada por género. No entanto, alguns erros comuns ocorreram devido a desafios típicos no processamento de linguagem natural. Estes erros incluíram:

- Ambiguidade semântica: Plots com descrições vagas ou que poderiam se enquadrar em diferentes géneros causaram erros de classificação. Por exemplo, um plot como "Escape by Night British" foi classificado como drama, quando o género correto era crime. A falta de contexto detalhado dificultou a previsão precisa.
- Sobreposição de géneros: Filmes que abordam múltiplos géneros simultaneamente, como comédia e drama, levaram a predições incorretas. Um exemplo foi um filme de tom mais leve “The Lemon Sisters", classificado apenas como comédia, embora tivesse elementos significativos de drama, o que levou a uma previsão parcial.
- Desequilíbrio nos géneros: Géneros com menos dados de treino, como sci-fi, tiveram um f1-score muito inferior (0.28), refletindo a dificuldade do modelo em capturar correctamente esses géneros devido à baixa representação no conjunto de treino. Estes filmes foram frequentemente rotulados como ação ou horror, que têm mais exemplos no dataset.

5 Trabalho Futuro

Se tivesse mais tempo, eu:

1. Modelos Avançados: Experimentaria modelos mais sofisticados, como Redes Neurais ou SVM, que podem capturar padrões complexos nas sinopses.
2. Aprimoramento de Recursos: Utilizaria TF-IDF ou embeddings de palavras para melhorar a representação semântica do texto.
3. Aumento de Dados: Aplicaria técnicas de aumento de dados, como substituição de sinônimos, para ampliar a variedade do conjunto de treino.

Modelo baseado: *https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/*