SINTRA
TECNOLOGIAS DIGITAIS
ECONOMIA E SOCIEDADE

Análise de Sentimentos nas Avaliações do IMDB

Lourenço Costa, 120048

Marianna Oliveira, 120054

Text Mining – 2024/2025

Afiliações: ISCTE-Sintra - escola de tecnologias digitais aplicadas.

Resumo

Este trabalho visa aplicar as ferramentas aprendidas no âmbito da UC Text Mining para análise de sentimentos em avaliações no site IMDb. Com um *dataset* que contém a avaliação em linguagem natural e o correspondente sentimento (positivo ou negativo), treinou-se modelos de aprendizagem automática para classificação do sentimento. E, de seguida, utilizou-se LDA para extrair tópicos latentes às avaliações. Assim, agrupando-os consoante os tópicos e observar o seu comportamento em termos da presença dos sentimentos associados às avaliações. Tendo este projeto a sua relevância no estudo das preferências dos consumidores de filmes.

Palavras-chave: Text Mining, Aprendizagem automática, Filmes, NLP, Análise, Sentimentos.

1. Introdução

Vivemos num mundo onde os a quantidade de dados a ser gerada é massiva. E com o trabalho correto é possível extrair conclusões valiosas e ajudar na tomada de decisões em diversos contextos, principalmente negócios. Neste projeto, o foco é na indústria cinematográfica. Com dados acerca nas avaliações dadas e os sentimentos associados (positivo ou negativo). Com essa informação, é possível tomar futuras decisões sobre filmes a serem feito, baseadas em dados, se o objetivo for maximizar o *feedback* positivo da audiência.

2. Métodos

2.1. Arquitetura do Estudo

A abordagem deste estudo divide-se nos passos a seguir descritos:

- Análise exploratória das avaliações.
- Pré-processamento e normalização das avaliações.
- Treino de modelos de aprendizagem automática supervisionada para classificação de sentimento.
- Extração de tópicos latentes.
- Análise dos dados, foco na relação entre tópicos e sentimentos.

2.2. Coleta dos Dados

Os dados foram adquiridos do repositório *Kaggle*, neste link: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-avaliações/data

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Imagem 1: Exemplo do dataset

2.3. Análise

A lingugagem utilizada para o trabalho foi Python. Para o pré-processamento do texto foi utilizada a biblioteca NLTK, scikit-learn para aprendizagem automática e LDA, matplotlib para visualização de gráficos e WordCloud para a criação de *wordclouds*.

3. Resultados

3.1. Análise exploratória dos dados

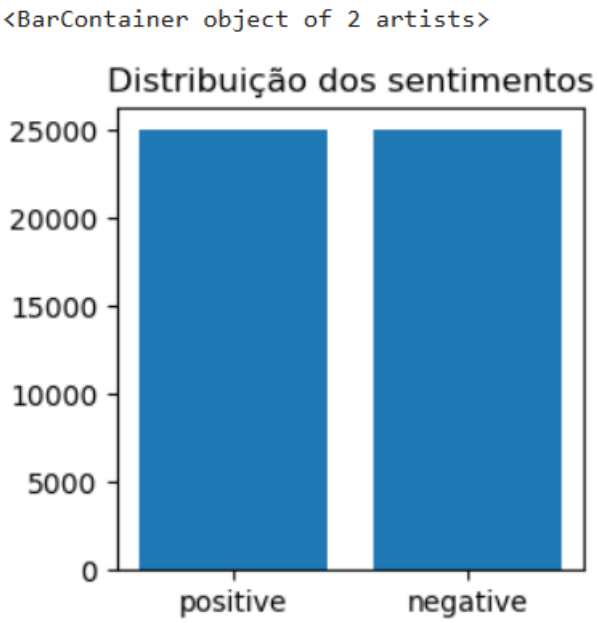

Imagem 2: Distribuição dos Sentimentos

```
# Média do comprimenta das avaliações
print('A média de palavras por review é de',int(len(str(df['review'].tolist()).split(' '))/len(df)),'palavras')
```

```
A média de palavras por review é de 231 palavras
```

Imagem 3: Média das palavras por avaliação

3.2. Aprendizagem Automática

```
Accuracy: 0.8871

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.88      0.89      5000
           1       0.88      0.90      0.89      5000

    accuracy                           0.89     10000
   macro avg       0.89      0.89      0.89     10000
weighted avg       0.89      0.89      0.89     10000
```

Imagem 4: Resultados com Logistic Regression

```
Accuracy: 0.8518

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.85      0.85      5000
           1       0.85      0.85      0.85      5000

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000
```

Imagem 5: Resultados com NaiveBayes

```
Accuracy: 0.84

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.86      0.84      5000
           1       0.85      0.81      0.83      5000

    accuracy                           0.84     10000
   macro avg       0.84      0.84      0.84     10000
weighted avg       0.84      0.84      0.84     10000
```

Imagem 6: Resultados com o Random Forest

3.3. Extração de tópicos com o LDA

```
Tópico 1
war film world character american series year human director documentary

Tópico 2
plot acting character actor script action performance cast better movie

Tópico 3
funny comedy watch movie think seen love laugh watching episode

Tópico 4
guy acting girl plot thing worst look watch minute woman

Tópico 5
love woman man young family performance father mother wife child

Tópico 6
horror effect thing movie look monster film zombie gore little

Tópico 7
action fight film jack king man cartoon western little hero

Tópico 8
think book say thing im read didnt seen watch movie

Tópico 9
music song musical film performance role rock dance michael star

Tópico 10
version original role cast james production john novel look star
```

Imagem 7: Resultados das palavras e tópicos latentes

3.4. Análise de Dados

Imagem 8: WordCloud – palavras mais utilizadas

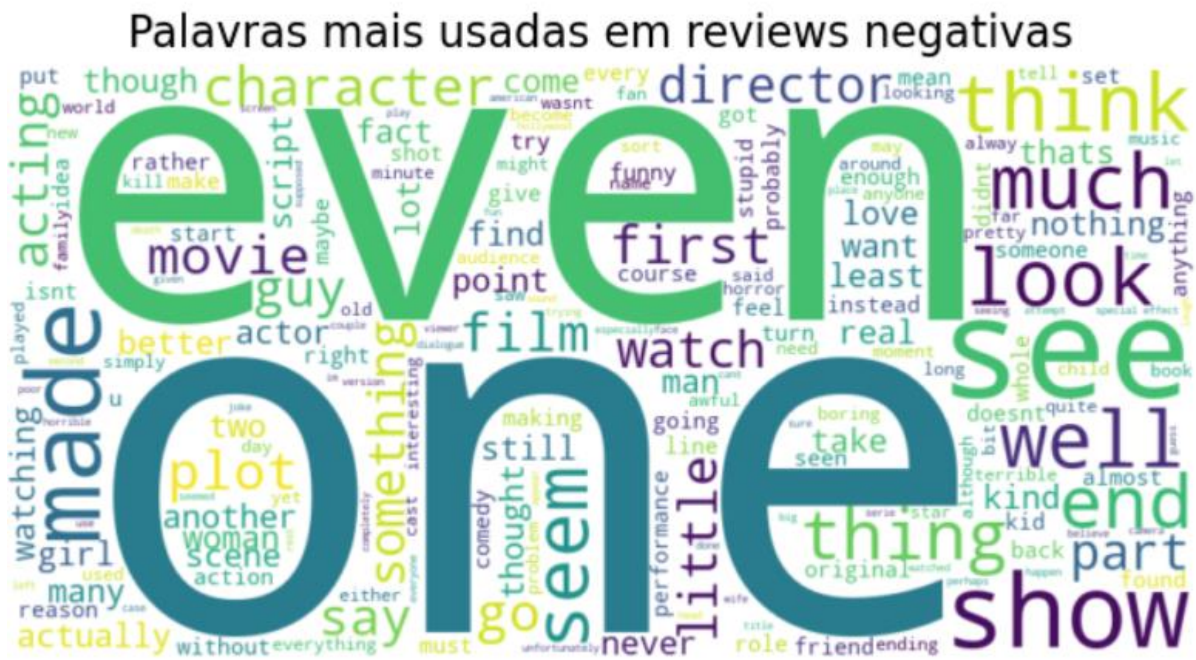

Imagem 9: WordCloud – palavras mais usadas nas avaliações negativas

Imagem 10: WordCloud – palavras mais utilizadas nas avaliações positivas

```
Palavras únicas no top 100 palavras mais usadas em comentários negativos:  'set', 'family', 'quite', 'bit', 'time', 'excellent', 'long', 'may', 'role',
'without', 'comedy', 'always', 'day', 'series', 'star', 'young', 'music', 'world', 'must', 'performance', 'saw', 'fun', 'right', 'fan', 'friend'

Palavras únicas no top 100 palavras mais usadas em comentários positivos:  'thats', 'there', 'ive', 'anything', 'reason', 'script', 'horror', 'shot', 'le
ast', 'enough', 'might', 'seems', 'pretty', 'minute', 'cant', 'effect', 'original', 'line', 'worst', 'isnt', 'try', 'whole', 'interesting', 'nothing', 'f
ar'
```

Imagem 11: Palavras únicas ao top 100 palavras mais utilizadas em cada sentimento

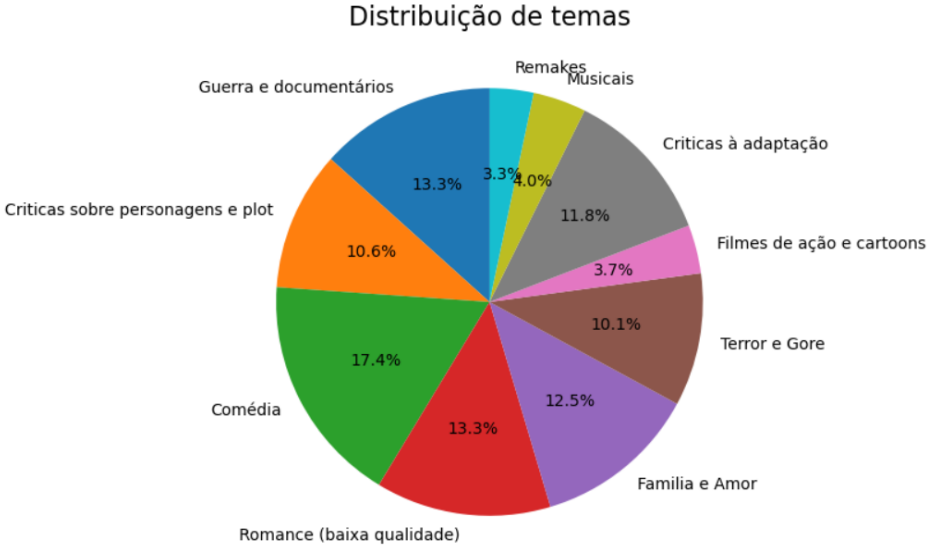

Imagem 12: Distribuição dos tópicos latentes pelas avaliações

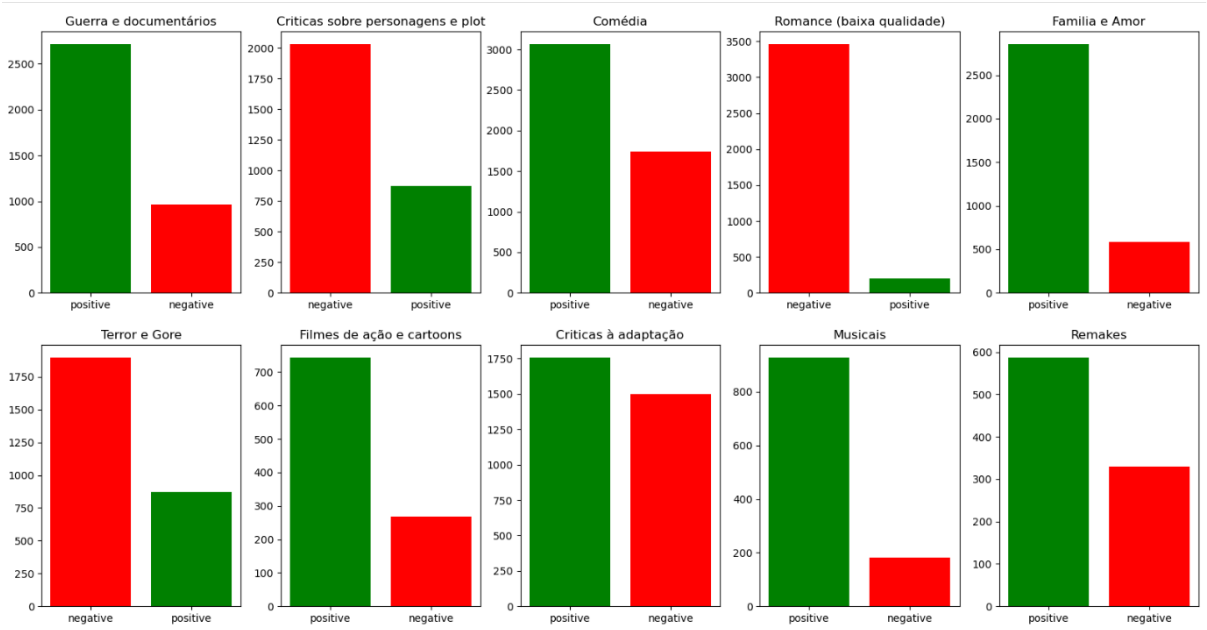

Imagem 13: Distribuição do sentimento por tópico

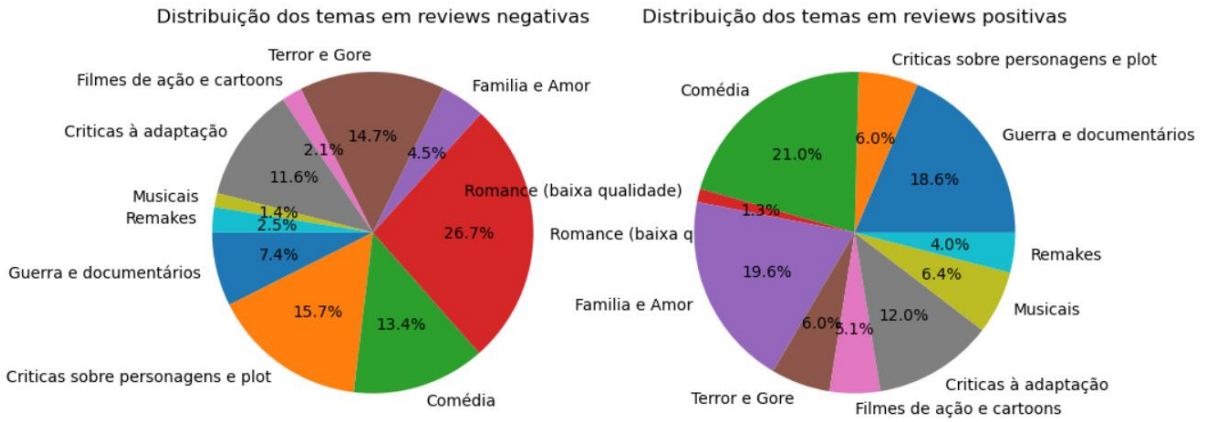

Imagem 14: Distribuição dos temas por sentimento

4. Discussão

4.1. Interpretação dos Resultados

Na análise exploratória, observamos que temos um *dataset* balanceado, e que a média das palavras por avaliação é de 231 palavras. O que nos leva a concluir que isto trata-se de avaliações extensas.

Na secção de aprendizagem automática, foram obtidos resultados satisfatórios, obtendo-se uma precisão acima de 82% nos modelos treinados.

Na extração de tópicos com o LDA, e depois de algumas mudanças nos parâmetros e *stopwords*, foi possível identificar 10 tópicos bem definidos tendo em conta o agrupamento de palavras. Os tópicos que melhor se adequam a cada grupo foram os seguintes:

1. **Guerras e documentários**: “war film world character american series year human director documentary “
2. **Criticas sobre personagens e plot:** “plot acting character actor script action performance cast better movie”
3. **Comédia:** “funny comedy watch movie think seen love laugh watching episode”
4. **Romance (baixa qualidade):** “guy acting girl plot thing worst look watch minute woman”
5. **Família e amor: “**love woman man young family performance father mother wife child”
6. **Terror e gore:** “horror effect thing movie look monster film zombie gore little”
7. **Filmes de ação e cartoon: “**action fight film jack king man cartoon western little hero”
8. **Criticas à adaptação: “**think book say thing im read didnt seen watch movie”
9. **Musicais:** “music song musical film performance role rock dance michael star”
10. **Remakes:** “version original role cast james production john novel look star”

Na secção da análise dos dados é aonde relacionamos os tópicos sugeridos e os sentimentos associados. Como podemos ver na imagem 12, como os temas estão distribuídos. E na Imagem 14, a sua distribuição por sentimento. Podemos observar que as avaliações de romance com tendência a baixa qualidade, críticas sobre personagens/plot e terror e gore têm uma presença muito forte nas avaliações negativas, ocupando 55.8% desse espaço. E a sua presença nas avaliações positivas é muito menor, ocupando 13%. Esse fato é ainda mais destacado na Imagem 13, aonde estes 3 tópicos são os únicos que tem mais avaliações negativas do que positivas com uma notória margem de diferença.

Críticas à adaptação de um filme e Remake são polarizantes na sua distribuição por sentimento. Comédia mesmo tendo uma presença nas avaliações negativas ainda continua a ter maioritária presença positiva. Filmes de ação/cartoon e guerra e documentário são solidamente positivas, com menor presença negativa que os tópicos anteriores. Por fim, os tópicos Musicais e Família/Amor são os que tem maior associação positiva.

4.2. Hipóteses

Existe algumas hipóteses para explicar alguns dos comportamentos observado no ponto anterior.

O tópico Romance (baixa qualidade) não ser tão aclamado pode ser devido a, e o mais óbvio, ter a tendência a ser de baixa qualidade. Segundo, ser um gênero de bastantes clichês e pouca originalidade, que pode levar a um *feedback* menos positivo. Para suportar isso, podemos ver na Imagem 11 que palavras como “original” e “interesting” existem unicamente em avaliações positivas. Logo, se algo for original e fora de clichês e arquétipos recorrentes tem a tendência a um impacto positivo na audiência. A presença negativa no tópico Terror e Gore pode ser explicada pelo facto de haver maior tendência a usar uma estética chocante e grotesca, que por natureza não é massivamente apelativo. Principalmente ao sacrifício de uma história bem escrita. O que se relaciona com as Críticas de Personagem/Plot serem altamente negativas, e também elementos como personagens e plot têm a

tendência de ser mais mencionadas quando há alguma dissatisfação com estes.

Guerra e documentários pode ter mais associação positiva pela sua natureza mais factual e de aprendizagem.

Críticas à adaptação e Remakes podem ser bastante polarizados. Críticas negativas nesses tópicos vem de uma audiência que, na maioria das vezes, é avidamente fã de uma franquia ou obra já existente e tem altas expectativas para o seu *remake* ou adaptação, que podem tão facilmente correspondidas como podem não ser. O que pode explicar a palavra “fan” ser única nas top 100 palavras mais presentes nas avaliações negativas.

Os tópicos Família/Amor, Musicais e Filmes de Ação/Cartoons podem ter o sentimento positivo fortemente associado pela sua natureza inofensiva.

4.3. Limitações e futuras melhorias

As conclusões a ser retiradas podem ser limitadas na sua veracidade devido a uma distribuição desigual dos tópicos presentes, podendo haver um *bias* na amostra. E também o LDA só cobriu cerca de metade do *dataset*, com um *threshold* de 0.5 que consideraria a *avaliação* parte do correspondente tópico. Havendo metade dos dados sem utilidade, porém foi por preferência de manter os tópicos bem definidos.

5. Conclusão

Em suma, com o uso de ferramentas de *text mining*, este projeto foi possível retirar relações entre os tópicos das *avaliações* e o seu sentimento, e as diversas razões e fatores dessas relações existirem em primeiro lugar.