**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: Rodrigo Franciozi Rodrigues da Silva | | **Data início curso**: 01/10/2022 |
| **Orientador**: Adâmara Felício | | **Defesa em:**  07/2024 |
| **Curso**: MBA Data Science e Analytics | **Modalidade**: Distância | Turma: 222 |

1. **Título do projeto:** Avaliando a concordância entre sentimentos expressos e notas atribuídas a avaliações de filmes.
2. **Introdução**

Nos últimos anos a indústria audiovisual cresceu consideravelmente e junto a esse crescimento advém o interesse de agradar o público para que obras de sucesso sejam criadas. Uma maneira de se verificar tal sucesso é por meio da checagem geral de notas e pela leitura de comentários sobre críticos especializados e o público geral que já assistiram determinado conteúdo. (Yassine, 2019)

Quando observado o conjunto dos comentários produzidos sobre o assunto, identificamos plataformas online que trabalham no registro das críticas e avaliações por meio do texto escrito como o Rotten Tomatoes, IMDb e Metacritic. Esses sites contém um grande número de registros que funcionam como bases de dados, para que pesquisadores consigam avaliar o engajamento e potencial sucesso de produções cinematográficas. (Harsh et al., 2023)

De acordo com Yassine (2019), até pouco tempo essas análises de comentários eram feitas de forma bastante manual, mas com o advento de técnicas de aprendizado de máquina como a análise de sentimentos, essa realidade acabou sendo modificada

A análise de sentimentos, conhecida também como mineração de opinião, consiste em uma técnica que avaliar a polaridade dos sentimentos, emoções e atitudes de determinado indivíduo de acordo com um objeto de interesse podendo ele ser um filme, produto, serviço e organização Azilawati et al. (2019). Diante das ideais previamente expostas, apresenta-se o problema de pesquisa e a hipótese a ser discutida e avaliada.

Problema da pesquisa: É possível classificar os textos de acordo com sua polaridade e avaliar se existe uma relação entre comentários e notas médias de filmes?

Hipótese: Entendemos que os textos escritos pela maior parte do público que já assistiu alguma obra cinematográfica, refletem a nota média geral atribuída para aquela obra em específico.

1. **Objetivo**

Utilizar modelos de aprendizados de máquina capazes de classificar sentimentos expressos nas avaliações de filmes, com o principal propósito de avaliar se existe coerência entre os comentários e a nota geral atribuída à determinada obra cinematográfica. Esse estudo contribuirá para uma compreensão maior entre percepções individuais e avaliações gerais, de forma a despertar ou não o interesse dos telespectadores na tomada de decisão com relação a investir tempo em determinada seleção de filmes para consumo.

A classificação dos comentários será alcançada com base no emprego da metodologia de análise de sentimentos a partir de uma coleta abrangente de avaliações de filmes, realizando os devidos procedimentos de pré-processamento e limpeza dos dados, assim como, na validação do modelo resultante para que se possa garantir uma eficácia apropriada.

1. **Material e Métodos**

A base de dados será obtida por meio da técnica de raspagem de dados (web scrapping) dentro da plataforma do Internet Movie Database[[1]](#footnote-1). Web Scrapping consiste em extrair dados da internet programaticamente e transformá-los em uma base de dados estruturada, permitindo uma coleta de grandes quantidades de dados de maneira automatizada, minimizando potenciais erros. (Mine Dogucu e Mine Çetinkaya-Rundel, 2021).

Apenas comentários em inglês serão considerados, devido ao site ser americano e a natureza desses comentários estar majoritariamente na língua nativa. O período escolhido será entre 2022 e 2023 para que se possa analisar avaliações mais recentes, a quantidade de filmes será decidida com base na obtenção de uma quantidade significativa para treinamento e teste dos modelos. A linguagem utilizada em todo o processo de análise e coleta de dados será desenvolvida em Python.

Logo após a estruturação dessa base de dados, técnicas de limpeza e transformação como a remoção de espaços em branco, emojis, números e caracteres especiais, assim como a padronização de palavras para letras minúsculas e remoção de stop-words serão aplicadas.

A utilização de dicionários léxicos auxiliará de forma que o processo de marcação dos comentários seja feita de uma maneira não manual. Essa técnica consiste em avaliar a polaridade do texto de acordo com a presença de algumas palavras que já possuem um determinado sentimento atrelado, seja ele positivo ou negativo. (Hota et al., 2021)

Por fim, dois modelos foram selecionados para que se possa realizar a comparação entre seus resultados. O primeiro será o Naive Bayes que é um tipo de modelo derivado do teorema de Bayes, onde se calcula a probabilidade de um evento acontecer com base em probabilidades condicionais, podendo ser aplicado no contexto de análise de sentimentos que leva em consideração a probabilidade de um texto pertencer a determinada classificação Murtadha et al. (2022). O segundo será o BERT, que é um modelo de rede neural desenvolvido pela Google em 2018 com o intuito de compreender o contexto de frases, auxiliando máquinas no processo de linguagem natural com o emprego da arquitetura de transformadores e transferência de aprendizado. BERT é o primeiro modelo a atingir o estado da arte com relação a performance para atividades em nível de sentença e token Jacob et al. (2018). Todas as comparações serão feitas com base nas métricas Acurácia, Sensibilidade, Precisão e Teste F1.

1. **Resultados Esperados**

Baseando-se em uma tabela de frequências gerada a partir das classificações (positiva, negativa, neutra) feitas pelo modelo treinado, levando em consideração cada um dos comentários, espera-se que filmes contendo uma nota média alta tenham um volume de comentários classificados como positivos maior, quando comparados com negativos e neutros. Seguindo a mesma premissa, quando a nota média for baixa, é esperado um maior volume de comentários classificados como negativos em comparação com positivos ou neutros.

1. **Cronograma de Atividades**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **9** | **10** | **11** | **12** | **1** | **2** | **3** | **4** | **5** | **6** |
| Revisão bibliográfica | **X** | X |  |  |  |  |  |  |  |  |
| Escrita e revisão do projeto de pesquisa | **X** | **X** |  |  |  |  |  |  |  |  |
| Entrega do projeto de pesquisa |  | **X** |  |  |  |  |  |  |  |  |
| Coleta e entendimento dos dados |  | **X** | **X** |  |  |  |  |  |  |  |
| Análise exploratória |  |  | **X** |  |  |  |  |  |  |  |
| Tratamento e limpeza dos dados |  |  | **X** | **X** | **X** |  |  |  |  |  |
| Aplicação dos modelos escolhidos |  |  | **X** | **X** | **X** |  |  |  |  |  |
| Teste e avaliação dos modelos treinados |  |  | **X** | **X** | **X** |  |  |  |  |  |
| Entrega dos resultados preliminares |  |  |  |  | **X** |  |  |  |  |  |
| Escrita do TCC |  | **X** | **X** | **X** | **X** | **X** | **X** |  |  |  |
| Entrega do TCC |  |  |  |  |  |  |  | **X** |  |  |
| Apresentação do TCC |  |  |  |  |  |  |  |  | **X** |  |
| Evento da defesa do TCC |  |  |  |  |  |  |  |  |  | **X** |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

1. **Referências Bibliográficas**

Harsh Sharma, Satyajit Pangaonkar, Reena Gunjan, Prakash Rokade. Sentimental Analysis of Movie Reviews Using Machine Learning. ITM Web of Conferences, 2023

Azilawati Azizan, Nurul Najwa SK Abdul Jamal, Mohammad Nasir Abdullah, Masurah Mohamad, and Nurkhairizan Khairudin. Lexicon-based sentiment analysis for movie review tweets. 1st International Conference on Artificial Intelligence and Data Sciences, 2019.

Mine Dogucu & Mine Çetinkaya-Rundel. Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities, Journal of Statistics and Data Science Education, 2021

Rodani, Yassine. Movie Sentiment Analysis: A Multinomial Naıve Bayes-Based Approach for Assessing User and Critic Opinions, 2019

Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. Sensors,2023

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018

Hota HS, Sharma DK, Verma N. Lexicon-based sentiment analysis using Twitter data: a case of COVID-19 outbreak in India and abroad. Data Science for COVID-19, 2021

Murtadha B. Ressan, Rehab F. Hassan. Department of Computer science, University of Technology, Baghdad, Iraq, 2022

1. www.imdb.com [↑](#footnote-ref-1)