Avaliação de Técnicas de Pré-Processamento e Algoritmos de Classificação para Analise de Sentimento em Comentários de Filmes do Letterboxd

Emanuel Flávio Dos Santos Silva[1]

Universidade Federal do Vale do São Francisco

Resumo O artigo aborda a relevância e a aplicação da análise de sentimento nas redes sociais, com foco no Letterboxd, uma plataforma voltada para entusiastas de cinema. O estudo investiga a eficácia de diferentes técnicas de pré-processamento na análise de sentimento de resenhas de filmes, utilizando quatro algoritmos de classificação: Naive Bayes, Random Forest, Regressão Logística e Support Vector Machine (SVM).

Keywords: Letterboxd · pré-processamento · classificação

1 Introdução

Nos dias atuais, é muito fácil para qualquer pessoa expor sua opinião na internet. As opiniões postadas nas redes sociais e outras plataformas online têm um impacto significativo na comunidade como um todo. Com a popularização dessas redes sociais, elas se tornaram uma fonte cada vez maior de material para análises, disponibilizando uma vasta quantidade de dados textuais. Esses dados podem ser extremamente úteis para diversos fins, incluindo a análise de sentimento, que busca entender as emoções expressas pelos usuários em relação a produtos, serviços ou outros tópicos.

Uma das plataformas que exemplifica bem essa tendência é o Letterboxd, uma rede social voltada para entusiastas de cinema. O Letterboxd permite que os usuários registrem os filmes que assistiram, classifiquem-nos, escrevam resenhas, criem listas personalizadas e interajam com outros usuários[1]. A plataforma se tornou um repositório rico de opiniões e sentimentos sobre filmes, oferecendo uma base de dados textual robusta para a realização de análises de sentimento. A análise de sentimento, neste contexto, permite classificar as resenhas dos usuários como positivas, neutras ou negativas[2][3], proporcionando insights valiosos sobre a recepção dos filmes.

A análise de sentimento tem se tornado cada vez mais relevante em diversas áreas, incluindo marketing, atendimento ao cliente e desenvolvimento de produtos[2]. Entender como os usuários se sentem em relação a algo pode fornecer informações cruciais para melhorar produtos e serviços, responder a feedbacks de maneira mais eficaz e adaptar estratégias de comunicação. Com o uso crescente

e intensivo das redes sociais, a análise de sentimento passou a ser um campo de estudo importante, focando em técnicas que possam lidar com grandes volumes de dados textuais e extrair sentimentos com precisão[3].

Este trabalho tem como objetivo analisar a eficácia de diferentes técnicas de pré-processamento na análise de sentimento de comentários de filmes. A análise foi aplicada utilizando quatro algoritmos de classificação: Naive Bayes, Random Forest, Regressão Logística e Support Vector Machine (SVM). Para isso, foi criada uma base de dados a partir das resenhas de uma seleção de 12 filmes no Letterboxd. Inicialmente, foram extraídas 24 mil resenhas, juntamente com suas respectivas classificações, que variam de 0.5 a 5 estrelas. Esta base de dados é composta por resenhas em diversos idiomas, embora a maioria seja em inglês.

2 Trabalhos Relacionados

Em Andrade e Rebs (2022), a pesquisa explora como os usuários do aplicativo Letterboxd, uma plataforma de catalogação e avaliação de filmes com mais de 3 milhões de usuários, utilizam as ferramentas disponíveis para avaliar obras audiovisuais. O estudo destaca a formação de capital social dentro das redes sociais online, especificamente nesta plataforma voltada para o consumo de conteúdo audiovisual. O artigo analisa os dados coletados na plataforma para entender como as redes sociais presentes no Letterboxd influenciam o consumo de filmes. Ao examinar as críticas feitas no aplicativo, a pesquisa identifica os valores que as redes utilizam ao avaliar as obras. Portanto, o estudo oferece uma visão sobre a interação entre as redes sociais e o consumo de conteúdo audiovisual.

Em Paes (2022), os autores utilizaram a mineração de texto do Twitter para entender como os usuários brasileiros opinam e dialogam sobre o desmatamento da Floresta Amazônica. Os resultados revelaram que os usuários brasileiros tendem a reagir a acontecimentos relacionados ao desmatamento da floresta Amazônica no Twitter e, que em sua maioria, os usuários apresentam sentimento negativo sobre o tema, alcançando picos de aproximadamente 60% dos tweets em determinado momento.

Segundo Britto e Pacífico (2019), a análise de sentimento tem sido objeto de interesse de pesquisadores devido às suas diversas aplicações, como análise de desempenho de produtos e detecção de distúrbios como depressão. Um dos principais desafios é a classificação de polaridade, que categoriza o sentimento expresso em um texto. No contexto do português, a maioria dos estudos se concentra nessa classificação, utilizando algoritmos de aprendizado de máquina para analisar sentimentos em redes sociais como o Twitter. No entanto, o português enfrenta uma escassez de recursos para pesquisa em análise de sentimento, como bases de dados públicas. Para contornar essa limitação, pesquisadores desenvolvem Web Corpus, uma coleção de documentos da web. Foi desenvolvido um Web Corpus de comentários de aplicativos móveis em português brasileiro e compara diferentes classificadores utilizados na literatura de análise de sentimento. O estudo demonstra a qualidade do corpus desenvolvido e sugere sua utilidade

para futuras pesquisas em análise de sentimentos e classificação de textos em português.

3 Materiais e Metodos

Nessa sessão, vai ser discutido os meios que esse estudo foi feito, partindo da descrição da base de dados e como ela foi contruida, como foi feita a rotulação, sobre as tecnicas de pré-processamento, algoritmos de classificação que foram utilizados, e as metricas de avaliação.

3.1 Base de Dados

Nesse estudo foi utilizada uma base de dados contendo comentários e notas de usuarios da plataforma de filmes Letterboxd. Inicialmente, a base continha o total de 23.972 comentários, coletados por meio de um crowler utilizando a linguagem de programação Python e a biblioteca Selenium. A extração foi feita nas avaliações de 12 filmes especificos, todos do gênero drama, com a tematica central sendo: a exploração do erro e redenção de um indivíduo. Os comentários coletados estavam em diversos idiomas, mas majoritáriamente no inglês. Para garantir uma uniformidade foi utilizado somente os comentários do idioma inglês, para isso, foi separado em duas bases, uma contendo somente as do inglês e outra com os demais, para fazer isso foi utilizada a biblioteca langdetect do python para fazer essa filtragem. Cada comentario vem acompanhado de uma nota que é atribuida pelo usuário, variando de 0.5 a 5, também pode fazer um comentario sem nota, nesse caso, foi atribuida nota 0 durante a extração, mas posteriormente foram removidos. Depois desses tratamentos, a base ficou com 19.025 comentários.

Como dito, os comentários da base são acompanhados por notas atribuidas pelo usuário, variando de 0.5 a 5. Para fazer a atribuição dos sentimentos de cada comentario por meio da nota foi usado:

- **Positivo:** Comentários com nota igual ou acima de 3.5
- **Neutro:** Comentários com nota igual a 3
- **Negativo:** Comentários com nota abaixo de 3

Essa rotulação reflete uma suposição comum na análise de sentimentos, onde notas mais altas indicam sentimentos positivos, notas intermediárias indicam sentimentos neutros, e notas baixas indicam sentimentos negativos. A rotulação permite a facilitação da distinção entre os sentimentos, sendo mais preciso que apenas as notas.

3.2 Técnicas de pré-processamento

Para se fazer a construção de modelos de classificação envolvendo texto, adicionar tecnicas de pré-processamento é uma etapa fundamental, logo que estudos

apontam que com a adição dessas tecnicas os resultados obtidos são melhores[4].Dito isso, foi utilizada algumas tecnicas de pré-processamento, sendo elas: Remoção de caracteres especiais e emojis, remoção das stopwords, Lemmatization e Stemming. Durante essa etapa é importante apontartambém foi utilizado uma etapa de pré-processamento chamada lowercase, que foi a de deixar todas em minusculas, que foi aplicada em todos.

Caracteres especiais são simbolos que não são letras alfabeticas ou digito numerico, sendo pontuações(!.,?), simbolos(@"(/), entre outros. Já emojis são representações gráficas de emoções, lugares, objetos entre outros. A remoção de caracteres especiais resulta em um texto mais limpo removendo elementos que não vão possuir um valor semantico para a análise, fazendo com que a classificação seja mais eficiente. A tecnica foi implementada usando expressões regulares (regex) da biblioteca re em Python, que identifica e substitui as strings de forma eficiente.

Stopwords são palavras comuns que geralmente não adicionam muito significado ao texto, como artigos, preposições e etc. E podem ser removidas para focar nas palavras mais relevantes[8]. A remoção foi feita utilizando a biblioteca NLKT, que oferece uma lista pré-definida de stopwords em vários idiomas. Com isso, podemos perceber que remover stopwords deixa o texto mais conciso e relevante já que remove palavras que não agregam muito aos algoritmos de classificação.

Lemmatization é o processo de transformar palavras em suas formas base, levando em consideração o contexto semantico e a gramatica da palavra[7]. Como a palavra “correu” e “corrida” nessa tecnica são reduzidas a forma base “correr”. Isso reduz a variação linguistica ao reduzir a quantidade de palavras que seriam classificadas de forma diferente para seu termo base, deixando a análise mais fácil. A implementação foi feita utilizando a biblioteca NLTK.

Stemming é a tecnica que reduz as palavras as suas raizes ou radicais, removendo sufixos e prefixos, direferente do lemmatization não considera o contexto semantico das palavras[6]. Como no exemplo anterior, as palavras “correu” e “corrida” serão reduzidas ao radical “corr”. Igualmente a lemmatization, reduz a variação linguistica tornando uma palavra que tem diversas formas em um único radical, simplificando o texto tornando os algoritmos mais rapidos e eficientes. A implementação foi feita utilizando a biblioteca NLTK.

3.3 Modelos utilizados

Para a construção desse estudo foram escolhidos quatro modelos de classificação supervisionada, sendo elas: NB (Naive Bayes), RF (Random Forest), LR (Logistic Regression) e SVM (Support Vector Machine), e sendo extraido as métricas de avaliação precisão (P), recall(R) e F1-Score(F1). Foi utilizada a biblioteca de aprendizado de máquina Scikit-Learn em Python. Antes dos dados serem passados para os classificadores foi utilizada uma técnica chamada vetorização que converte os dados para um formato numerico[5], para esse foi utilizado o CountVectorizer da biblioteca Sklearn.

3.4 Métricas de avaliação

Para o estudo as metricas que foram utilizadas como já citado anteriormente, foi a precisão (P), recall(R) e F1-Score(F1). A precisão calcula o número de instâncias que são pertinentes, é a proporção de verdadeiros positivos que vai ser representado como (TP), em relação ao número total de exemplos que foram classificados como positivos, ou seja, qualquer tipo de positivo, sendo verdadeiro positivo(TP) ou falso positivo(FP), (TP + FP)[8].

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{1}$$

O recall calcula o número de instâncias pertinentes que são recuperadas, sendo a proporção de verdadeiros positivos(TP), em relação ao numero total de exemplos que são realmente positivos, ou seja os verdadeiros positivos(TP) e os falsos negativos(FN), (TP + FN)[8].

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{2}$$

O F1- Score é a média harmonica da precisao e recall, só vai ser alto se tanto a precisão quanto o recall forem altos, já que é um valor para equilibrar, sendo a multiplicação do recall e precisão multiplicado por 2 (2.(P.R)), em relação a soma do recall e precisão, (P + R)[8].

$$F_1 = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3}$$

4 Resultados

Seguindo o que foi dito, foram construidos modelos e utilizadas tecnicas de pré-processamento em cada modelo, para a facilitação e melhor apresentação das tabelas, vai ser utilizada abreviações tanto das tecnicas de pré-processamento quanto dos classificadores, segue as tabelas 1 e 2.

Abreviação	**Classificador**
SVM	Support Vector Machine
RF	Random Forest
LR	Logistic Regression
NB	Naive Bayes

Tabela 1. Descrição das abreviações dos classificadores

Abreviação	**Método de Pré-processamento**
LEM	Lemmatization
STEM	Stemming
STOP	Remoção de Stopwords
RCE	Remoção de Caracteres Especiais e Emojis
NON	Sem Pré-processamento

Tabela 2. Descrição das abreviações dos métodos de pré-processamento

Com isso, a tabela 3 mostra os resultados obtidos, sendo a precisão, recall e F1-Score, dos classifcadores junto com uma técnica de pré-processamento.

Precisão e Recall

Modelo e Técnica	**Precisão**	**Recall**	**F1- score**
NON + NB	0.82	0.69	0.74
NON + RF	0.78	0.85	0.80
NON + LR	0.80	0.77	0.79
NON + SVM	0.82	0.67	0.73
RCE + NB	0.82	0.69	0.74
RCE + RF	0.78	0.85	0.80
RCE + LR	0.80	0.77	0.78
RCE + SVM	0.82	0.68	0.73
STOP + NB	0.81	0.69	0.74
STOP + RF	0.77	0.84	0.79
STOP + LR	0.80	0.76	0.78
STOP + SVM	0.82	0.68	0.73
LEM + NB	0.82	0.68	0.74
LEM + RF	0.77	0.85	0.79
LEM + LR	0.81	0.77	0.78
LEM + SVM	0.82	0.67	0.73
STEM + NB	0.82	0.67	0.73
STEM + RF	0.78	0.85	0.80
STEM + LR	0.80	0.75	0.77
STEM + SVM	0.82	0.65	0.71

Tabela 3. Precisão, Recall e F1-Score

A precisão, que mede a proporção de verdadeiros positivos entre todas as previsões positivas, atingiu seu valor máximo de 0.82 em várias combinações de técnicas e modelos. As combinações que alcançaram este valor máximo foram NON + NB, NON + SVM, RCE + NB, RCE + SVM, STOP + SVM, LEM + NB, LEM + SVM, STEM + NB e STEM + SVM. Este resultado indica que esses modelos são altamente eficazes em minimizar falsos positivos, garantindo que uma alta proporção das previsões positivas feitas sejam corretas.

O recall, que mede a capacidade do modelo de identificar verdadeiros positivos entre todas as instâncias reais positivas, podemos ver que algumas combinações tambem chegaram ao valor máximo, que foi de 0.85, nas combinações RCE + RF, LEM + RF, e STEM + RF, pode-se notar que o modelo Random Forest(RF) se destaca nessa metrica tendo os maiores valores, junto com as tecnicas de pré-processamento Remoção de caracteres e emojis(RCE), Lemmatization(LEM) E Streaming(STEM).

O F1-score, que é a média harmônica da precisão e do recall e fornece uma medida equilibrada da performance do modelo, o valor máximo obtido foi de 0.80, nas combinações NON + RF, RCE + RF, e STEM + RF, como pode ser observado o classificador Random Forest, foi o que apresentou melhores valores de F1-Score mostrando que tem um bom equilibrio entre a precisão e o recall.

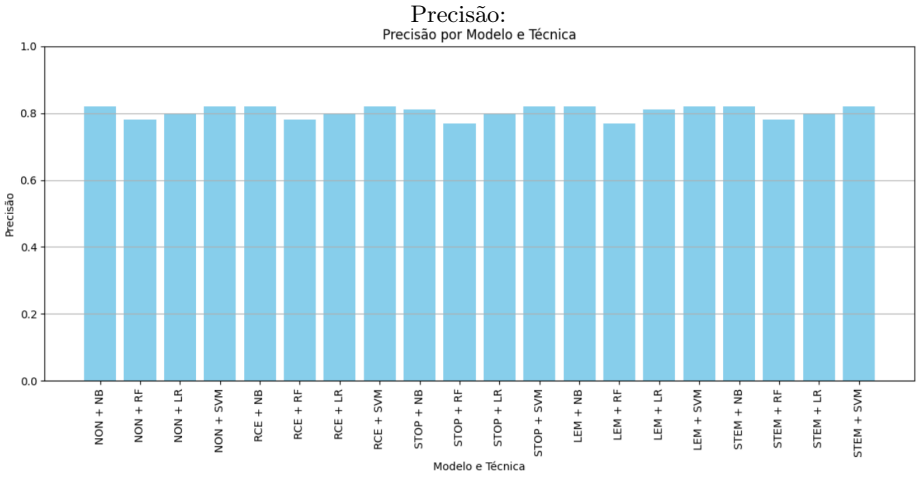

Figura 1. Precisão dos modelos.

Recall:

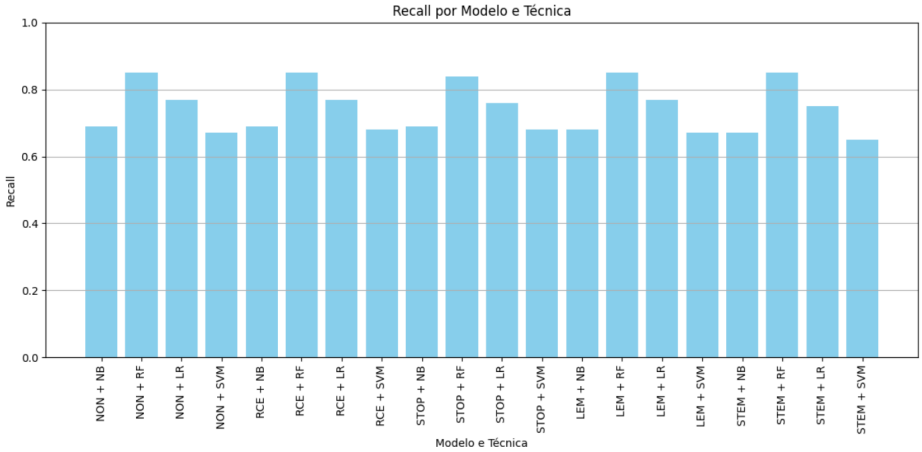

Figura 2. F1-Score dos modelos.

F1-Score:

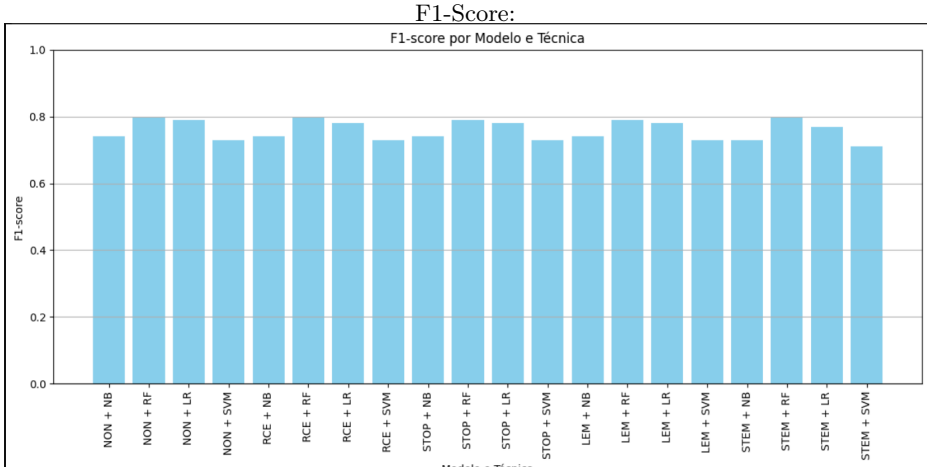

Figura 3. Recall dos modelos.

Os resultados mostram que para análise de sentimento, com essa base de dados, o modelo Random Forrest se mostrou superior aos demais, utilizado principalmente em conjunto com as tecnicas de Remoção de caracteres e emojis e o steming, pode proporcionar um desempenho robusto quando se tratar de análise de semtimento, mesmo que quando se tratava de precisão o Random forest tenha apresentados valores menores, mas o equilibrio geral fornecido pelo F1-Score favorece o uso desse modelo.

5 Conclusão

Neste estudo, foi realizada uma análise detalhada para verificar quais são as melhores combinações entre técnicas de pré-processamento e algoritmos de classificação em uma análise de sentimento de comentários de filmes da plataforma Letterboxd. Através de uma avaliação cuidadosa utilizando métricas como precisão, recall e F1-score, foi possível obter resultados significativos e identificar combinações superiores.

Os resultados mostraram que o modelo Random Forest (RF), quando combinado com as técnicas de pré-processamento stemming (STEM) e remoção de caracteres especiais e emojis (RCE), apresentou um desempenho superior em comparação com outras combinações. Este modelo se destacou particularmente nas métricas de recall e F1-score, indicando sua eficácia em identificar verdadeiros positivos e manter um bom equilíbrio entre precisão e recall.

A importância das métricas de avaliação não pode ser subestimada na escolha do modelo mais adequado para uma análise específica. A precisão, recall e F1-score oferecem diferentes perspectivas sobre o desempenho do modelo, cada uma delas destacando aspectos essenciais da classificação. A precisão é crucial para entender a proporção de previsões corretas entre as positivas, enquanto o recall foca na capacidade de capturar a maioria dos verdadeiros positivos. O F1-score, sendo a média harmônica dessas duas métricas, fornece uma visão equilibrada, combinando os benefícios de ambos.

Referências Bibliográficas

[1] ANDRADE, Juliana de; REBS, Rebeca Recuero. A estética do consumo de filmes na plataforma Letterboxd. 2022.

[2] PAES, Vinícius J. et al. Análise de sentimento em tweets relacionados ao desmatamento da floresta amazônica. In: Anais do XI Brazilian Workshop on Social Network Analysis and Mining. SBC, 2022.

[3] BRITTO, Larissa; PACÍFICO, Luciano. Análise de sentimentos para revisoes de aplicativos mobile em português brasileiro. In: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2019.

[4] DE ALMEIDA NETO, José A.; DE MELO, Tiago. Exploring Supervised Learning Models for Multi-Label Text Classification in Brazilian Restaurant Reviews. In: Anais do XX Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2023.

[5] DA SILVA, Marcos VJ et al. Preprocessing Applied to Legal Text Mining: analysis and evaluation of the main techniques used. In: Anais do XX Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2023.

[6] SOUZA, Ellen et al. Assessing the impact of stemming algorithms applied to brazilian legislative documents retrieval. In: Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC, 2021.

[7] ALMANSA, Luciana; RUBIO, Gabriel; MACEDO, Alessandra. A Question Answering System over Chronic Diseases and Epigenetics Knowledge. In: Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde. SBC, 2020.

[8] GARCIA, Marina Pinho; GARCIA, Giovana Pinho; DA SILVA, Nádia Félix Felipe. Humor Detection using Support Vector Machine. In: Anais da IX Escola Regional de Informática de Goiás. SBC, 2021.