Fresco ou Podre: uma análise dos *reviews* do *Rotten Tomatoes*

AUTORA

Bárbara Gonçalves Oliveira, aluna de pós-graduação da PUC Minas.

1

DEFINIÇÃO DO PROBLEMA

Examinar *reviews* de filmes a fim de fazer uma classificação e tirar uma conclusão final sobre se o filme é bom ou ruim em termos gerais.

2

RESULTADOS E PREDIÇÕES

Classificador: Possibilitar a classificação correta dos *reviews* de críticos entre "*Rotten*" e "*Fresh*".
Espera-se um modelo capaz de aferir corretamente a classificação fornecida pelo crítico dado o teor do texto.

Objetivo: Obter a capacidade de classificar um filme entre *Rotten* e *Fresh* dado um texto de revisão qualquer.

3

AQUISIÇÃO DOS DADOS

Dados obtidos do *Rotten Tomatoes* a partir de duas fontes distintas:

Dados dos filmes e dos *reviews* dos críticos coletados da plataforma *Kaggle.*

Dados dos *reviews* de usuários coletados através de uma raspagem de dados do próprio site do *Rotten Tomatoes.*

4

MODELAGEM

Algoritmos de aprendizado de máquina supervisionado.

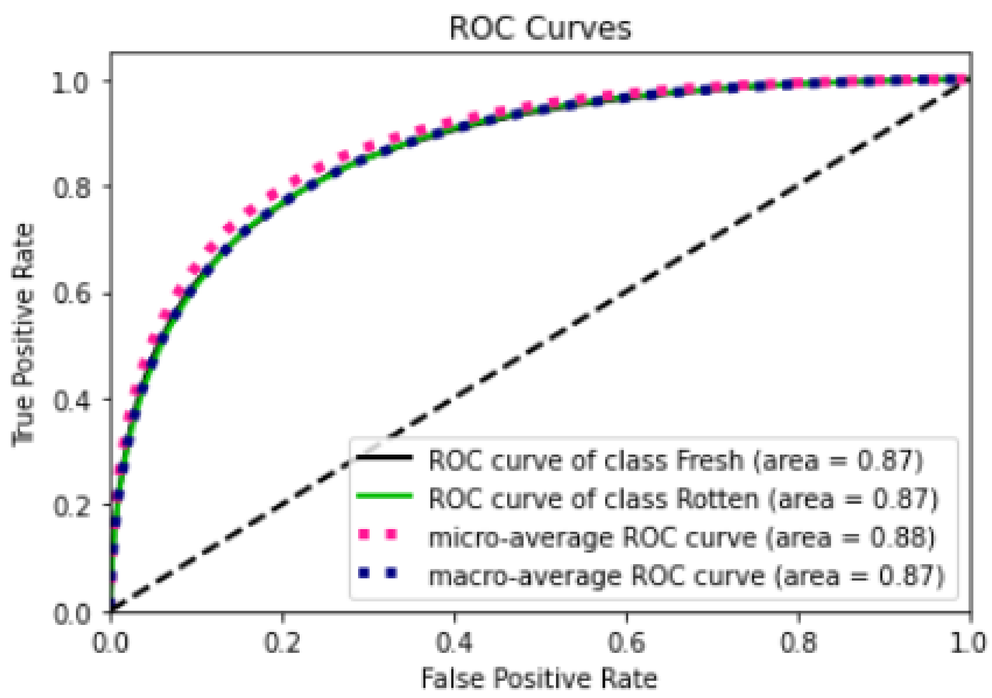
Foram feitas análises utilizando os algoritmos clássicos da biblioteca sklearn:

Regressão Logística, Gradient Boosting, Linear SVC e KNN.

5

AVALIAÇÃO DO MODELO

Relatórios de Classificação, olhando para a acurácia geral, os índices de previsão e revocação das classes eparadamente.
A área sob a curva ROC também foi uma das métricas importantes para a definição do modelo.



6

PREPARAÇÃO DOS DADOS

Remoção de dados nulos, verificação do idioma dos textos e limpeza dos não-ingleses, redução das palavras dos textos a seus termos comuns utilizando a técnica de lematização.

ATIVAÇÃO (VIDE REPOSITÓRIO DO GITHUB)

https://github.com/BarbaraOlive/rotten_tomatoes_classifier