SEMINÁRIO FINAL

GRUPO

- Enzo Dezem Alves RA: 801743
- Lucca Couto Barberato RA: 800257
- Luís Fernando do Carmo Lourenço RA: 800210
- Matheus Menecucci RA: 800310
- Rodrigo Takizawa Yamauchi RA: 800226
- Thales Winther RA: 802499

Tarefa escolhida

Contexto e Motivação

- **Problema:** prever quantas estrelas (1 a 5) um usuário deu a um filme com base apenas no texto da avaliação;
- **Aplicação prática:** sistemas como IMDb, Rotten Tomatoes e Letterboxd;
- **Desafio:** linguagem subjetiva, ambígua e altamente variável entre usuários;
- **Relevância em PLN:** combina análise de sentimentos com classificação multiclasse.

Tarefa escolhida

Características da Tarefa

- Classificação supervisionada com múltiplas classes (1 a 5);
- Difere de tarefas clássicas de análise de sentimentos binária (positivo vs. negativo);
- Avaliações podem ser contraditórias (ex: críticas negativas com nota alta);
- Uso de feedback textual como sinal indireto de nota numérica.

Dataset Escolhido - UTLCorpus

- **Escolha :**
  - Escolhemos para o nosso trabalho o dataset recomendado no slide 23 da Aula 8.
  - https://github.com/RogerFig/UTLCorpus/tree/master

- **Domínios:**
  - Movies: opiniões/avaliações de filmes.

- **Campos:** domain, object, author, text, likes, unlikes, stars, date, collect_date, replies, favorite, want_see, recommend, see, internal_id

Dataset Escolhido - UTLCorpus

Amostra

corpus.txt

C: > Users > enzod > Desktop > corpus.txt

```
,domain,object,author,text,likes,unlikes,stars,date,collect_date,replies,favorite,want_see,recommend,see,internal_id
0,movies,007-cassino-royale-t23,0,Um dos melhores do 007,0,0,3.5,8 anos atrás,20-04-2019 06:20,0,0,0,,1,0
1,movies,007-cassino-royale-t23,1,"assisti só pela metade..do meio ao fim, o filme parece ser mt bom ! espero poder assistir o filme inteiro logo logo!",0,0,4,8 anos
2,movies,007-cassino-royale-t23,2,foi um dos filmes mais violentos q já vi. mas foi booom demaais! *-*,0,0,4,8 anos atrás,20-04-2019 06:20,0,0,0,,1,2
3,movies,007-cassino-royale-t23,3,Adoroo esse filme!,0,0,0,8 anos atrás,20-04-2019 06:20,0,1,0,,1,3
4,movies,007-cassino-royale-t23,4,O Daniel Crieg tem carisma zero como 007...,1,0,4,8 anos atrás,20-04-2019 06:20,0,0,0,,1,4
5,movies,007-cassino-royale-t23,5,um dos melhores de 007,0,0,4,8 anos atrás,20-04-2019 06:20,0,0,0,,1,5
6,movies,007-cassino-royale-t23,6,"Nunca pensei que diria isso em um filme 007, mas esse é bastante ""meigo""",0,0,2.5,8 anos atrás,20-04-2019 06:20,0,0,0,,1,6
7,movies,007-cassino-royale-t23,7,preciso rever,0,0,5,8 anos atrás,20-04-2019 06:20,0,0,0,,1,7
8,movies,007-cassino-royale-t23,8,Adorei... Nossa esse Daniel Craig é UM HOMEM...,2,0,0,8 anos atrás,20-04-2019 06:20,0,0,0,,1,8
9,movies,007-cassino-royale-t23,9,"Divertido, mas achei no inicio um pouco entediante, com a Vesper que melhora. a parte do veneno na bebida é a mais emocionante do f
10,movies,007-cassino-royale-t23,10,"Absolutamente surpreendente! Temia que fosse detestar o roteiro do presunçoso Paul Haggis e que acharia a escolha de Daniel Craig
11,movies,007-cassino-royale-t23,11,"Nota 8,0. Ao contrario de muitos Daniel Craig está muito bem como James Bond. E este filme tem todos os elementos de um bom filme
12,movies,007-cassino-royale-t23,12,"Na moral,um dos melhores filmes que assisti na minha vida",3,0,0,8 anos atrás,20-04-2019 06:20,0,0,0,,1,12
13,movies,007-cassino-royale-t23,13,mto melhor 007 q o pierce ..#prontofalei,2,0,5,8 anos atrás,20-04-2019 06:20,0,0,0,,1,13
```

- **Campos principais para o projeto:**
  - **text**: texto livre da resenha (não estruturado, em PT-BR, pode conter gírias, emojis, pontuações informais e HTML residual).
  - **stars**: nota decimal (0.0–5.0) atribuída pelo autor. Será convertido para classe inteira (1–5).

Dataset Refinado - UTLCorpus

Amostra

test_movies.csv

C: > Users > enzod > Desktop > PLN > test_movies.csv

```
object,text,stars,discrete_stars
meninas-malvadas-t2902,A D O R OOOOO!,0.0,1
star-wars-episodio-i-a-ameaca-fantasma-t1065,Emocionei *ooooooooooooooo*,0.0,1
a-partida-t4792,Filme de uma delicadeza incrível. Direção de arte maravilhosa.,0.0,1
o-livro-de-eli-t8427,"Afinal de contas, ele era ou não era cego?!",0.0,1
esqueceram-de-mim-t6572,O conselho tutelar deveria se envolver na historia. e rever isso aí,0.0,1
cidadao-kane-t4647,"Se fizessem um filme nacional baseado em ""Cidadão Kane"" com certeza o título seria ""CIDADÃO MARINHO"".",0.0
a-vida-e-bela-t2115,lindo filme :),0.0,1
cinquenta-tons-de-cinza-t59132,Deusolivre que filme ruim.Nada salva. Nem o Jamie que está maravilhoso em The Fall.,0.5,1
sim-senhor-t3535,"Super bacana, leve e com o carey neh cara!sempre bom!",0.0,1
copias-de-volta-a-vida-t109621,Vale uma estrela pela cara dura do Keanu hehehehehe,1.0,1
a-ultima-musica-t11844,"A história é legal em si, mas a atuação da Miley, por favor! ¬¬",0.0,1
noites-de-cabiria-t5490,"Cabííííria, cabííííria!",0.0,1
te-amarei-para-sempre-t8418,"erick bana lindooooooo!!Rachel McAdams eh perfeita,adoroo o filme eh lindooo recomendoo..chorei!",0.0
```

- **Campos usados nos modelos:**
  - object - Identificador do filme avaliado.
  - text - Texto livre contendo a opinião do usuário sobre o filme.
  - stars - Nota original em ponto flutuante (0.0–5.0) atribuída pelo usuário.
  - discrete_stars - Nota convertida para classes inteiras de 1 a 5.

Balanceamento do Dataset

- **Motivação**
  - Desequilíbrio de classes (classe 2 tinha ~60 000 amostras e classe 5 1.305.863)
  - Risco de viés em modelos supervisionados
- **Estratégia de Balanceamento**
  - Amostragem de 10 000 registros da classe 2
  - Seleção de 10 000 registros para cada classe remanescente (1, 3, 4 e 5)
  - Concatenação em um dataset final de 50 000 instâncias

ESTRATÉGIA 1

MODELO BASEADO EM REGRAS LÉXICAS

Estratégia 1 - Pré-processamento

- **Normalização**: e-mails → <EMAIL>, números → <NUM>
- **Limpeza**: remoção de tags HTML, múltiplos espaços
- **Tokenização & Stopwords**: NLTK (português) - pórem preservando negadores e intensificadores usados no código.
- Remoção de espaços extras e números.
- **Stemming**: RSLP (ex.: “assistir” → “assist”) - não realizado pois atrapalha análise e acurácia.

Estratégia 1 - Classificação Léxica

- **Fonte léxica: LaCAfe** (AffectPT-br v1.0)

- Recurso lexical para o português brasileiro desenvolvido por [Silva et al., 2012],

- Códigos “31” → posemo (positivo)
- Códigos “32” → negemo (negativo)

```
3 30  affect (Affect)
4 31  posemo (Positive Emotions)
5 32  negemo (Negative Emotions)
6 33  anx (Anx)
7 34  anger (Anger)
8 35  sad (Sad)
9
10 %
11 (:  30  31
12 ):  30  32
13 :(  30  32
14 :)  30  31
15 abaix*  30  32  35
16 abandon*  30  32  35
17 abatid* 30  32  35
18 abenço* 30  31
19 aberração 30  32
20 abertura  30  31
21 aborrec*  30  32
22 abraço  30  31
23 abraços 30  31
24 abus* 30  32  34
```

Fonte: GitHub

Estratégia 1 - Regra de Pontuação

- Para cada review, calculamos um escore emocional bruto (lex_score)
  - +1 para cada termo positivo
  - –1 para cada termo negativo

```
# Caso normal
if any(tl.startswith(stem) for stem in LEX_POS_STEMS):
    score += 1
if any(tl.startswith(stem) for stem in LEX_NEG_STEMS):
    score -= 1
```

Estratégia 1 - Tratamento de Negações

- Palavras de negação invertendo polaridade do próximo token

- **Conjunto de negadores:**
  - não
  - nem
  - nunca
  - jamais

- **Exemplo**:
  - “Não gostei” → token “gostei” positivo invertido → score -1

Estratégia 1 - Adição de Intensificadores

- Intensificadores duplicam o peso emocional do termo seguinte

- **Conjunto de intensificadores:**
  - muito, super, extremamente, bem, tão, totalmente, mega, bastante

- **Exemplos**:
  - “muito bom” → score +2
  - “super ruim” → score –2

```
# Se for intensificador e tiver palavras em janela de uma palavra
if tl in intensificadores and i+1 < len(tokens):
    nxt = tokens[i+1].lower()
    if any(nxt.startswith(stem) for stem in LEX_POS_STEMS):
        score += 2
        i += 2
        continue
    if any(nxt.startswith(stem) for stem in LEX_NEG_STEMS):
        score -= 2
        i += 2
        continue
```

Estratégia 1 - Normalização da Pontuação

- Convertemos o lex_score bruto para uma escala discreta de 1 a 5 estrelas
- Mapeamento por “faixas” de escore:
- Garante notas comparáveis a avaliações humanas

```
4
5 # calcula percentis
6 def map_score_faixa(s):
7     if s <= -1: return 1
8     elif s <= 0: return 2
9     elif s <= 1: return 3
0     elif s <= 2: return 4
1     else:        return 5
```

Avaliação Quantitativa - Modelo Baseado em Léxico

Distribuição de classes

Classe (Estrelas)	Distribuição Real (discrete_stars)	Previsão Léxica (lex_stars_q)	Diferença
1 ★	1929	1591	-338
2 ★	1818	2807	989
3 ★	1865	2395	530
4 ★	1839	1233	-606
5 ★	1961	1386	-575

Avaliação Quantitativa - Modelo Baseado em Léxico

Métricas por Classe

Classe	Precision	Recall	F1-score	Support
1 ★	20,1	16,5	18,1	1929
2 ★	21,8	33,6	26,4	1818
3 ★	19,1	24,6	21,5	1865
4 ★	23,0	15,4	18,4	1839
5 ★	25,0	17,6	20,7	1961

Avaliação Quantitativa - Modelo Baseado em Léxico

Matriz de Confusão

Real \ Previsto	1 ★	2 ★	3 ★	4 ★	5 ★
Real 1 ★	319	654	573	228	155
Real 2 ★	506	611	347	177	177
Real 3 ★	313	527	458	254	313
Real 4 ★	220	481	460	283	395
Real 5 ★	233	534	557	291	346

Avaliação Quantitativa - Modelo Baseado em Léxico

Considerações finais

- **Baixa Acurácia e Centralização das Previsões**: o modelo obteve apenas 21,4% de acurácia, com tendência a prever notas intermediárias (2 e 3 estrelas), ignorando os extremos emocionais.

- **Léxico Limitado e Pontuação Neutra**: a ausência de termos no léxico ou o equilíbrio entre palavras positivas e negativas resultou em scores próximos de zero e previsões imprecisas.

Avaliação Qualitativa - Modelo Baseado em Léxico

Amostras de 1 estrela

Texto (resumo)	Estrelas reais	lex_score	Estrelas previstas
“A D O R OOOOO!”	1⭐	0	2⭐
“Emocionei *ooooooooooooo*”	1⭐	0	2⭐
“Filme de uma delicadeza incrível. Direção de arte maravilhosa.”	1⭐	3	5⭐
“lindo filme :)”	1⭐	1	3⭐

Avaliação Qualitativa - Modelo Baseado em Léxico

Amostras de 2 estrelas

Texto (resumo)	Estrelas reais	lex_score	Estrelas previstas
muito lenta a história	2⭐	0	2⭐
Desleixado e fora de contexto. Deveriam ter esperado um tempo maior e desenvolvido o projeto com mais calma e inteligência.	2⭐	3	5⭐

Avaliação Qualitativa - Modelo Baseado em Léxico

Amostras de 3 estrelas

Texto (resumo)	Estrelas reais	lex_score	Estrelas previstas
Tinha q ser mais dinâmico, daria até pra colocar o desfecho de cenas q ficaram pra subentender no final.Digo q serve	3⭐	0	2⭐
Achei muito parecido e considerei uma cópia de As Duas Vidas de Audrey Rose	3⭐	0	2⭐

Avaliação Qualitativa - Modelo Baseado em Léxico

Amostras de 4 estrelas

Texto (resumo)	Estrelas reais	lex_score	Estrelas previstas
Muito bom as cenas Pós créditos. O filme é bom.A cruzada de pernas me tirou risadas	4⭐	3	5⭐
Trilha sonora linda demais e o Gerard de Fantasma está lindo, mesmo deformado!	4⭐	2	4⭐

Avaliação Qualitativa - Modelo Baseado em Léxico

Amostras de 5 estrelas

Texto (resumo)	Estrelas reais	lex_score	Estrelas previstas
Profundo e sincero. Stephen Daldry é, como sempre, impecável.	5⭐	2	4⭐
É uma necessidade ver esse filme em todas as páscoas! No mínimo! <3	5⭐	0	2⭐

Avaliação Qualitativa - Modelo Baseado em Léxico

Considerações finais

- **Fraco em Ironia, Emojis e Contexto**: o modelo falha ao lidar com ironias, símbolos emocionais e textos subjetivos, gerando previsões incoerentes com o conteúdo real.

- **Melhor em Textos Curtos e Diretos**: desempenho aceitável em avaiações curtas e objetivas com termos reconhecidos.

ESTRATÉGIA 2

MODELO BASEADO EM APRENDIZADO DE MÁQUINA

Estratégia 2 - Análise Exploratória

Análise dos tokens que mais aparecem:

Token	Quantidade
que	43.917
o	42.717
de	40.822
e	39.099
a	35.031
filme	30.041
é	25.962
um	21.998
não	21.185
do	19.445

Stop-words no top100 (sem pontuação): 81/100 (81.00%)

Tokens do top100 que não estão na lista de stop-words: 'filme', 'pra', 'história', 'melhor', 'filmes', 'achei', 'gostei', 'roteiro', 'cenas', 'personagens', 'assistir', 'cena', 'vida', 'personagem', 'acho', 'cinema', 'vi', 'trilha', 'atuação'

Muitas palavras do domínio e algumas palavras que não devem ser removidas pois indicam sentimento

Estratégia 2 - Análise Exploratória

Palavras do domínio que aparecem no top100

Token	Posição
história	38
filmes	44
roteiro	57
cenas	61
personagens	67
assistir	69
cena	77
personagem	81
cinema	88
trilha	93
atuação	98

Adição desse conjunto de palavras específicas do domínio dos dados no conjunto de stopwords, visto que não trazem sinal de sentimento e podem atrapalhar a vetorização

Estratégia 2 - Pré-processamento

Diferente da estratégia 1. Foco em produzir representações numéricas relevantes

- **Normalização**: Letras minúsculas e substituição de padrões por placeholders como <NUM>, <EXCL> e <QST>
- **Remoção de ruídos estruturais**: Eliminação de elementos como URLs, tags HTML e pontuação irrelevante
- **Conversão de emojis**: Substitui emojis por shortcodes no formato <EMOJI_nome>
- **Stopwords**: Remoção de stopwords padrão da lista do spaCy e inclusão de palavras específicas do domínio
- **Tokenização e lematização:** Redução das palavras à sua forma base (lematização), descartando stopwords e reduzindo a esparsidade dos dados.
- **Vetorização TF-IDF**: geração de representações numéricas como entrada para os modelos

Estratégia 2 - Modelos de Classificação

Modelos selecionados: k-NN, SVM, Naive Bayes, Regressão Logística

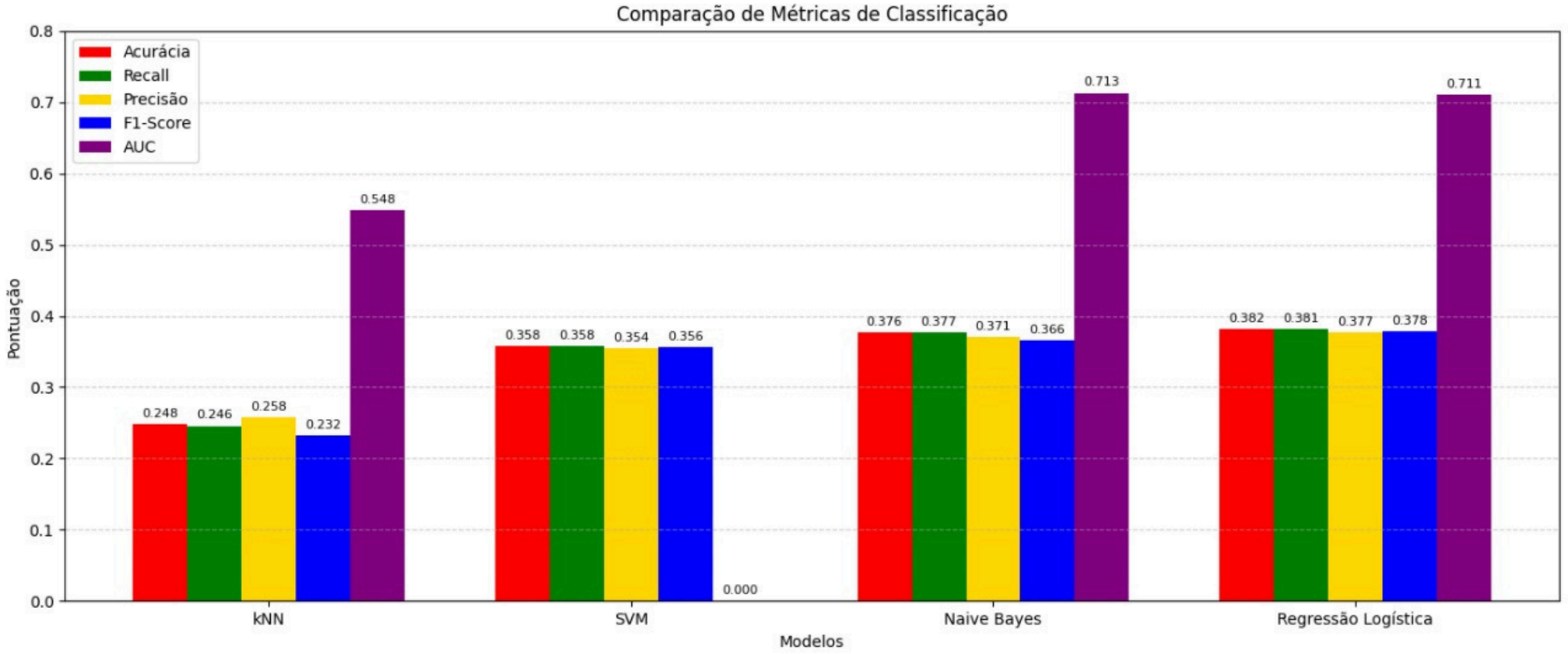
Entrada: Vetor numérico a partir do TF-IDF - representação numérica

Saída: Predição da classe (1 a 5) e probabilidades* (para cálculo da AUC)

Somente o SVM não contém as probabilidades devido à implementação do LinearSVC do scikit-learn

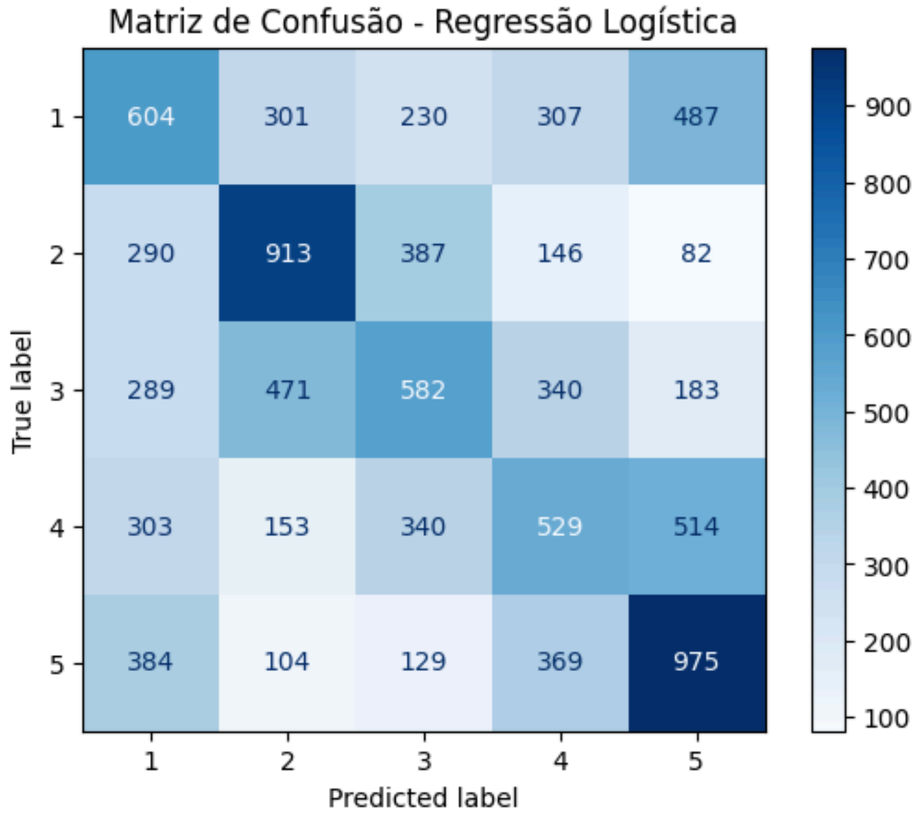
Estratégia 2 - Avaliação Quantitativa

- Desempenho do Modelo:
  - **Acurácia: 38%** (acima da linha de base aleatória de 20%, mas indica dificuldade)
  - **AUC: 0.71** (sugere que o modelo aprendeu a distinguir sentimentos, mesmo errando a nota exata)

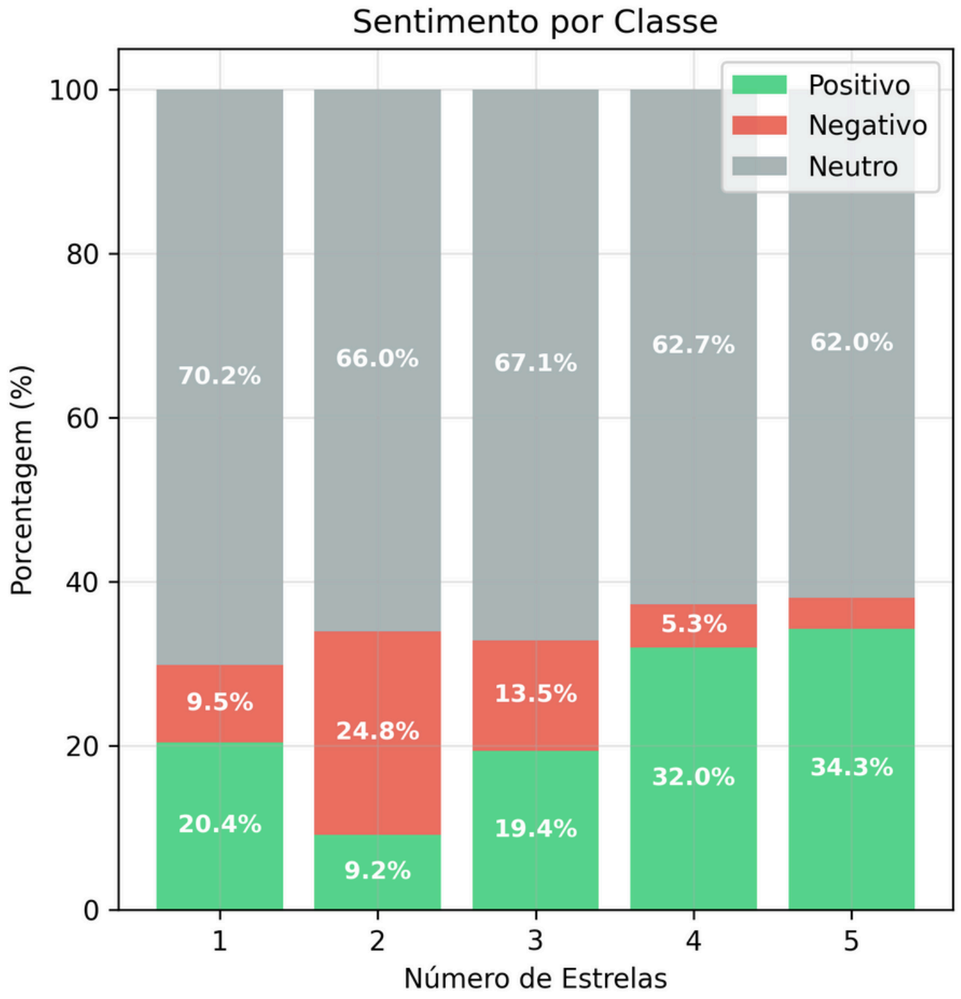


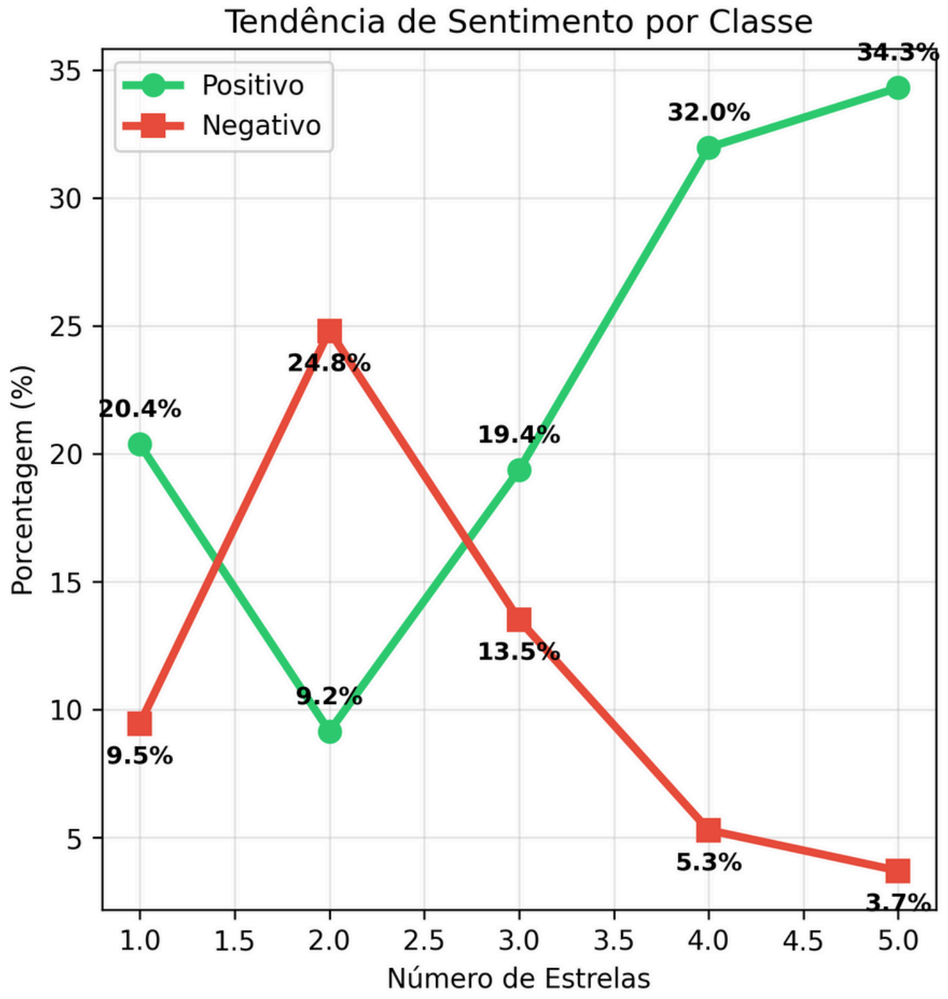
Estratégia 2 - Avaliação Quantitativa

- Principal Desafio:
  - Inúmeras avaliações com **texto positivo** ("amei o filme"), mas **com nota 1**.
  - O modelo recebeu "sinais" conflitantes, aprendendo uma **associação errada entre palavras positivas e notas baixas**.



Estratégia 2 - Avaliação Qualitativa





Avaliação Qualitativa - Modelo Baseado em Aprendizado de Máquina

Amostras de 1 estrela

Texto (resumo)	Estrelas reais	Estrelas previstas
“A D O R OOOOO!”	1⭐	5⭐
“Emocionei ***oooooooooooooo***”	1⭐	5⭐
“Filme de uma delicadeza incrível. Direção de arte maravilhosa.”	1⭐	5⭐
“lindo filme :)”	1⭐	1⭐

Avaliação Qualitativa - Modelo Baseado em Aprendizado de Máquina

Amostras de 2 estrelas

Texto (resumo)	Estrelas reais	Estrelas previstas
muito lenta a história	2★	2★
Desleixado e fora de contexto. Deveriam ter esperado um tempo maior e desenvolvido o projeto com mais calma e inteligência.	2★	3★

Avaliação Qualitativa - Modelo Baseado em Aprendizado de Máquina

Amostras de 3 estrelas

Texto (resumo)	Estrelas reais	Estrelas previstas
Tinha q ser mais dinâmico, daria até pra colocar o desfecho de cenas q ficaram pra subentender no final.Digo q serve	3⭐	4⭐
Achei muito parecido e considerei uma cópia de As Duas Vidas de Audrey Rose	3⭐	4⭐

Avaliação Qualitativa - Modelo Baseado em Aprendizado de Máquina

Amostras de 4 estrelas

Texto (resumo)	Estrelas reais	Estrelas previstas
Muito bom as cenas Pós créditos. O filme é bom.A cruzada de pernas me tirou risadas	4⭐	4⭐
Trilha sonora linda demais e o Gerard de Fantasma está lindo, mesmo deformado!	4⭐	4⭐

Avaliação Qualitativa - Modelo Baseado em Aprendizado de Máquina

Amostras de 5 estrelas

Texto (resumo)	Estrelas reais	Estrelas previstas
Profundo e sincero. Stephen Daldry é, como sempre, impecável.	5⭐	4⭐
É uma necessidade ver esse filme em todas as páscoas! No mínimo! <3	5⭐	2⭐