UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Larissa Vitória Vasconcelos Sousa

1. Introdução

O Plenty of Fish (POF) é um popular aplicativo de namoro online, amplamente utilizado em países como Canadá, Reino Unido, Irlanda, Austrália, Nova Zelândia, Espanha, Brasil e Estados Unidos. Como uma das plataformas de namoro mais acessadas, é fundamental para a empresa manter um ambiente positivo e funcional para seus usuários, sendo as avaliações dos usuários um indicador essencial para medir o sucesso dessas iniciativas.

Sendo assim, este estudo tem como objetivo comparar os resultados obtidos na AV1 com os da AV2, utilizando os mesmos dados e aplicando um pipeline de LLM (*Large Language Model*) chamado "*distilbert-base-uncased*" para classificar as avaliações. Este modelo foi escolhido por ser uma versão mais compacta e rápida do BERT, baseado na arquitetura de transformadores.

A análise se concentra em duas variáveis principais, o conteúdo das avaliações e a pontuação atribuída. O *dataset* contém mais de 435 mil registros de avaliações do aplicativo disponíveis na *Google Play Store*, no entanto, devido a limitações de processamento, apenas 10 mil linhas foram utilizadas nesta análise. A Tabela 1 apresenta todas as colunas presentes no *dataset*.

Tabela 1 - Base de dados

Nome da coluna	Descrição
Nome do usuário	Nome do usuário que fez a avaliação
Imagem do usuário	Imagem de perfil associada ao usuário
Conteúdo	Comentário deixado pelo usuário sobre o aplicativo
Pontuação	Avaliação do aplicativo, em uma escala de 1 a 5 estrelas
Contagem de polegares para cima	Número de vezes que o comentário foi considerado útil por outros usuários
Versão da revisão	Versão do aplicativo no momento em que a avaliação foi feita

Data de criação da avaliação	Data e hora em que o comentário foi postado
Resposta ao comentário	Resposta da empresa ao comentário do usuário
Data da resposta	Data e hora em que a empresa respondeu ao comentário
ID da avaliação	Identificador único da avaliação

2. Observações

Durante o desenvolvimento do trabalho, alguns desafios foram enfrentados, como a necessidade de refatorar a função de classificação, que não utilizava um objeto *Dataset* personalizado, complicou o tratamento dos dados e aumentou a propensão a erros. A implementação da nova classe *TextDataset* trouxe dificuldades na tokenização, especialmente em relação ao tamanho máximo de 512 *tokens*. Embora a modularização tenha melhorado a clareza do código, o treinamento inicial dos modelos revelou que a falta de estrutura nos *datasets* resultou em métricas de desempenho insatisfatórias, com a acurácia não ultrapassando 50%. No entanto, esses desafios foram solucionados ao longo do processo.

3. Resultados e discussão

3.1 Melhorias na AV1

Inicialmente, foi adotada uma abordagem na coluna "*score*", que originalmente variava de 1 a 5 estrelas, conforme a figura 1. Nessa transformação, os valores 1 e 2 foram mapeados para "ruim", os valores 3 foram desconsiderados, e os valores 4 e 5 foram mapeados para "bom", resultando em uma classificação binária (0 para "ruim" e 1 para "bom"), como mostra a figura 2. Também foi realizada a busca de melhores parâmetros na AV1, utilizando *Grid Search.* Essas modificações na AV1 simplificaram a tarefa de classificação, o que resultou em melhorias nos resultados da AV1, veja na tabela 2 e figura 4 . Somente após essa etapa, utilizei *LLM* para otimizar ainda mais a classificação.

Figura 1 - Distribuição inicial das notas das avaliações

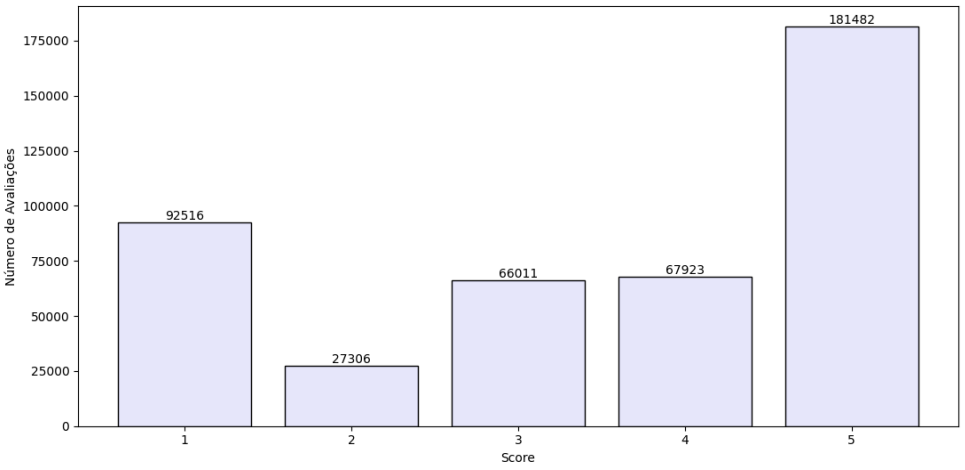
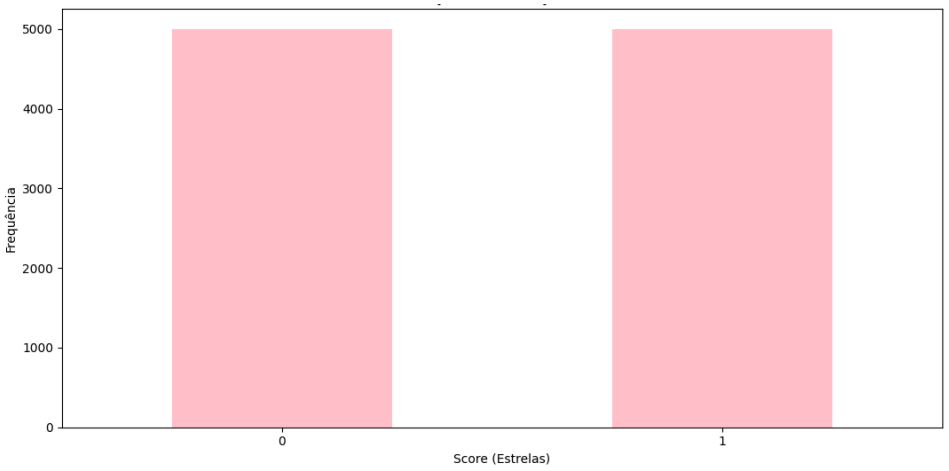


Figura 2 - Distribuição das notas após a transformação e balanceamento



As Tabelas 1 e 2 apresentam os resultados da AV1 antes e depois das modificações implementadas, incluindo a aplicação do *Grid Search*. Observa-se um avanço nas métricas de desempenho dos modelos após as melhorias, por exemplo, a acurácia dos modelos *TF-IDF* e *CountVectorizer* com pré-processamento aumentou de 72% e 50%, respectivamente, para 74% e 75%. Além disso, o modelo *TF-IDF* com lemmatização alcançou uma acurácia de 89%, demonstrando a eficácia das otimizações realizadas. As melhorias na precisão e no recall, tanto para as avaliações "Ruim" (0) quanto "Bom" (1), refletem um bom progresso na qualidade da classificação.

Tabela 1 - Métricas dos modelos antes do Grid Search

Modelos	Acurácia	Precisão		Recall	
		Ruim (0)	Bom (1)	Ruim (0)	Bom (1)
TF-IDF com pré-processamento	72%	66%	84%	90%	55%
TF-IDF sem pré-processamento	50%	50%	51%	79%	21%
CountVectorizer com pré-processamento	50%	50%	51%	65%	36%

Figura 3 - Matrizes de correlação (antes do Grid Search)

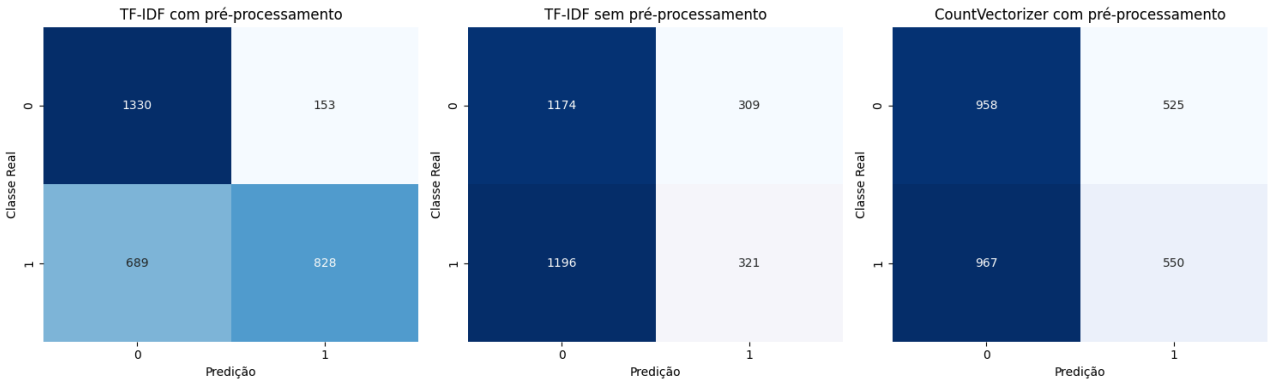
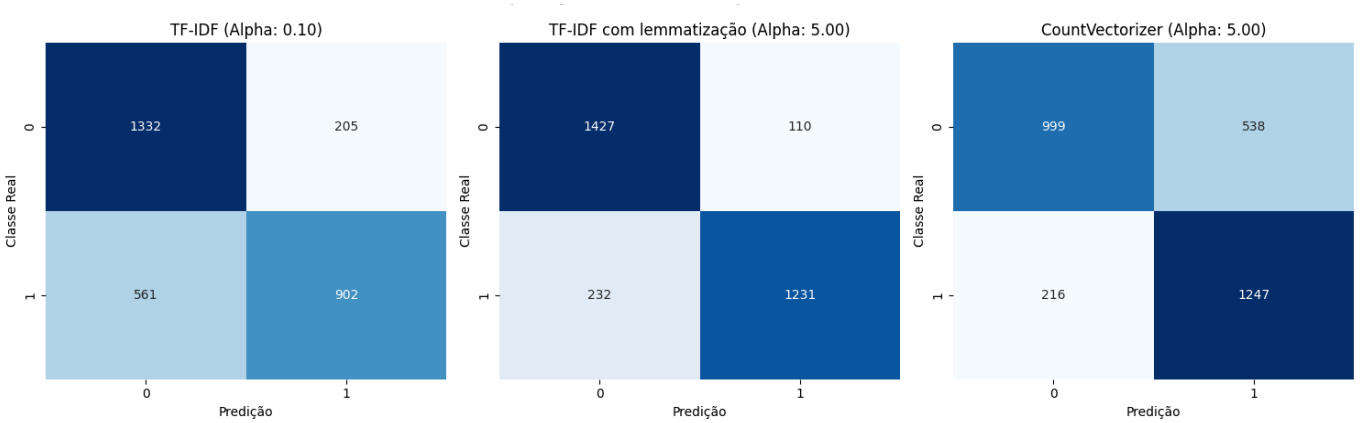


Tabela 2 - Métricas dos modelos após o Grid Search

Modelos	Acurácia	Precisão		Recall	
		Ruim (0)	Bom (1)	Ruim (0)	Bom (1)
TF-IDF com pré-processamento (Alpha: 0.10)	74%	70%	81%	87%	62%
CountVectorizer com pré-processamento (Alpha: 5.00)	75%	82%	70%	65%	85%
TF-IDF com lemmatização (Alpha: 5.00)	89%	86%	92%	93%	84%

Figura 4 - Matrizes de correlação (após Grid Search)



3.2 Resultados AV2

Os resultados, utilizando LLM, demonstraram uma melhoria notável em comparação com a AV1. A Tabela 3 revela que a acurácia dos modelos alcançou impressionantes 98% tanto com quanto sem pré-processamento. As métricas de precisão e *recall* também se destacaram, com a precisão chegando a 99% nas avaliações "Bom" (1) quando pré-processadas, e mantendo valores elevados para as avaliações "Ruim" (0). Essas melhorias evidenciam a eficácia do modelo LLM na captura de nuances nas avaliações, resultando em uma classificação muito mais precisa e consistente em relação aos métodos anteriores utilizados na AV1.

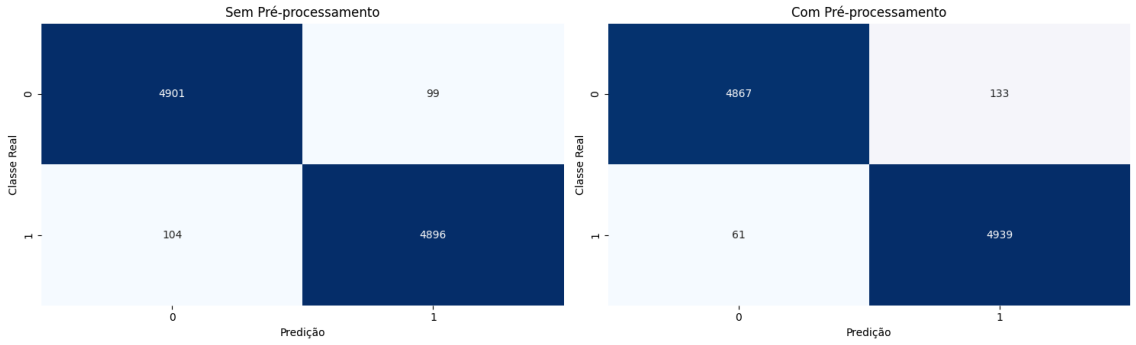
Tabela 3 - Métricas dos modelos após uso de LLM

Modelos	**Acurácia**	**Precisão**		**Recall**	
		Ruim (0)	**Bom (1)**	**Ruim (0)**	**Bom (1)**
Sem Pré-processamento	98%	98%	98%	98%	98%
Com Pré-processamento	98%	99%	97%	97%	99%

A figura 5 mostra as matrizes de confusão da AV2, que sugerem que não há indícios de *overfitting* nos modelos, uma vez que as classificações estão acertando de maneira consistente para ambas as classes. As matrizes mostram que, sem

pré-processamento, o modelo classificou corretamente 4.901 avaliações "Ruim" (0) e 4.896 avaliações "Bom" (1), com apenas 99 e 104 erros, respectivamente. Com pré-processamento, os resultados foram igualmente incríveis, com 4.867 classificações corretas para "Ruim" (0) e 4.939 para "Bom" (1), resultando em apenas 133 e 61 erros. Essa precisão nas previsões para ambas as classes reforça a confiança de que o modelo está generalizando bem, sem se ajustar excessivamente aos dados de treinamento.

Figura 5 - Matrizes de correlação (AV2)



4. Conclusões

Os resultados obtidos neste trabalho superaram as expectativas iniciais, demonstrando avanços na classificação das avaliações do aplicativo *Plenty of Fish*. Ao comparar a AV1 com a AV2, evidenciou-se uma melhora considerável nas métricas de desempenho após a implementação de técnicas de otimização, como o *Grid Search* e o uso de LLM. Na AV1, a acurácia dos modelos estava abaixo de 50%, mas após as melhorias, os modelos atingiram uma acurácia de até 89%. Com a introdução do modelo "*distilbert-base-uncased*" na AV2, as acurácias chegaram a impressionantes 98%, tanto com quanto sem pré-processamento. Além disso, as altas taxas de precisão e *recall* sugerem que o modelo é eficaz na distinção entre avaliações "Ruim" e "Bom". As matrizes de confusão reforçam a confiança na generalização do modelo, indicando que não há indícios de *overfitting*.

5. Próximos passos

Para próximos passos, sugiro que seja considerada a expansão do *dataset*. Com apenas 10 mil registros utilizados nesta análise, a inclusão de mais dados poderia melhorar a generalização e a robustez do modelo. No entanto, é importante ressaltar

que essa expansão exigirá uma capacidade de memória substancial e um ambiente de processamento adequado, uma vez que um maior volume de dados pode impactar a performance e a eficiência do treinamento. A ampliação do conjunto de dados pode proporcionar uma representação mais rica e diversificada das avaliações, resultando em um desempenho ainda mais preciso e confiável nas classificações.