CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

Faculdade de Tecnologia Baixada Santista Rubens Lara

Curso Superior de Tecnologia em Ciência de Dados

Thiago Fernandes Vilela

Álgebra Linear – Algoritmo de similaridade de cosseno

Santos 2025

Análise de Similaridade de Reviews de Battlefield 6 utilizando TF-IDF

1. Descrição do Dataset e Tema

O dataset utilizado contém **avaliações de usuários brasileiros sobre o jogo *Battlefield 6* (BF6)**, coletadas da plataforma Steam. Cada linha representa uma review escrita por um jogador, contendo texto livre e a respectiva classificação (positiva ou negativa).

O tema escolhido foi a **análise semântica de sentimentos negativos**, com foco em identificar **padrões linguísticos e possíveis ocorrências de ironia** em comentários considerados negativos.

O objetivo principal foi verificar se, a partir de uma **review modelo negativa**, seria possível encontrar **outras avaliações com conteúdo similar**, mesmo que expressassem insatisfação de forma implícita ou irônica.

2. Metodologia

O processo de análise seguiu as seguintes etapas:

1. **Leitura e pré-processamento do dataset**:
    a. Conversão de todos os textos para minúsculas;
    b. Remoção de pontuação, números e caracteres especiais;
    c. Exclusão de *stopwords* da língua portuguesa.
2. **Vetorização com TF-IDF (Term Frequency–Inverse Document Frequency)**:
   Cada review foi transformada em um vetor numérico com base na frequência e relevância das palavras. Termos comuns a várias reviews receberam peso menor, enquanto palavras específicas (como *lag*, *rollback*, *bug*) tiveram maior relevância.
3. **Cálculo da Similaridade do Cosseno**:
   A similaridade do cosseno foi utilizada para comparar o vetor da **review negativa modelo** com todas as outras do dataset, gerando um valor entre 0 e 1 (sendo 1 a equivalência total).
4. **Análise dos ângulos**:
   Como o cosseno reflete o ângulo entre os vetores, valores mais próximos de 1 indicam ângulos pequenos (alta similaridade semântica), enquanto valores menores indicam ângulos maiores (menor semelhança).

3. Resultados

O comentário modelo selecionado foi o seguinte:

“Pedi devolução pois enfrentei diversos problemas de performance no modo multiplayer, com muito lag, rollback e input delay, mesmo tendo especificações acima do mínimo e com drivers atualizados... problemas que não rolaram no beta. Ainda vejo potencial no game, mas vou dar um tempo e esperar corrigirem esses problemas (que eu sei que nem todos estão enfrentando).”

A similaridade calculada com as demais avaliações gerou os seguintes resultados (sem duplicatas):

Similaridade	Comentário resumido	Observação
1.000	O próprio comentário modelo	—
0.126	“O jogo é bom, mas tem vários problemas de rede, bugs e latência.”	Compartilha vocabulário técnico de reclamação
0.123	“Ainda não é o momento de comprar, há bugs e falhas no matchmaking.”	Tom crítico construtivo, próximo semanticamente
0.073 a 0.059	Comentários positivos, destacando diversão ou gráficos	Vocabulário diferente, ângulo maior
≤ 0.05	Reviews elogiando o jogo, sem menção a defeitos	Sem similaridade textual significativa

4. Análise dos Ângulos

Os valores de similaridade obtidos correspondem aos seguintes ângulos aproximados (em graus):

Similaridade	Ângulo (°)
1.000	0°
0.126	82.8°
0.123	82.9°
0.073	85.8°
0.065	86.3°
0.059	86.6°
0.051	87.1°
0.040	87.7°

Esses valores mostram que apenas dois textos apresentam ângulos relativamente menores (em torno de 83°), indicando **maior proximidade semântica** com o comentário modelo. As demais possuem ângulos mais abertos (acima de 85°), o que reflete **diferença significativa de contexto e vocabulário**.

5. Discussão

Apesar da review modelo apresentar **críticas diretas à performance e rede**, o modelo TF-IDF conseguiu identificar outros textos que **expressam descontentamento técnico**, ainda que de forma mais genérica.

Por outro lado, não foram encontradas **reviews ironicamente negativas**, pois o método TF-IDF analisa apenas a frequência e relevância de palavras, **sem compreender o tom emocional ou o sarcasmo**. Assim, expressões irônicas com vocabulário positivo (ex.: “o jogo está incrível, pena que não funciona”) não seriam identificadas como similares.

6. Conclusão

O experimento demonstrou que o **TF-IDF combinado com a Similaridade do Cosseno** é eficaz para encontrar **reviews com queixas semelhantes em termos técnicos e semânticos**, mas apresenta limitações ao lidar com **contextos irônicos ou ambíguos**.

Para capturar esse tipo de sutileza, seria necessário empregar **modelos de linguagem contextual**, como BERT ou embeddings semânticos mais modernos.

Ainda assim, o método alcançou o objetivo proposto de **identificar padrões de insatisfação** em textos com estrutura e vocabulário próximos à review negativa inicial.