Relatório de Análise de NLP Híbrida em Comentários do Instagram

Vinícius de Paula R Carvalho

30 de setembro de 2025

Sumário

1 Introdução

Este relatório apresenta uma análise detalhada dos resultados obtidos a partir de um pipeline de Processamento de Linguagem Natural (NLP) aplicado a um corpus de comentários extraídos de publicações no Instagram (Reels). A abordagem utilizada é híbrida, combinando duas técnicas principais:

1. **Análise de Sentimentos:** Para classificar a polaridade emocional de cada comentário (positivo, negativo ou neutro).
2. **Modelagem de Tópicos:** Para identificar e descrever os temas subjacentes discutidos nos comentários.

O grande diferencial deste projeto reside na metodologia de modelagem de tópicos, que integra o robusto framework **BERTopic**, baseado em Transformers, com o poder de sumarização e representação do modelo generativo **Gemini API**. Essa sinergia permite não apenas agrupar comentários por similaridade semântica, mas também gerar rótulos descritivos e contextuais para cada tópico, superando as tradicionais listas de palavras-chave.

O objetivo deste documento é detalhar a metodologia empregada, analisar os resultados quantitativos e qualitativos e interpretar as visualizações geradas para extrair insights valiosos sobre o conteúdo dos comentários.

2 Metodologia

O pipeline de análise foi estruturado em três etapas principais: pré-processamento dos dados, análise de sentimentos e a modelagem de tópicos híbrida.

2.1 Dados e Pré-processamento

Os dados brutos consistem em um DataFrame contendo comentários de Reels do Instagram. Antes da modelagem, foram aplicadas as seguintes etapas de pré-processamento no texto dos comentários:

- **Remoção de Stop Words:** Palavras comuns da língua portuguesa (ex: "de", "para", "com") que não carregam significado semântico relevante foram removidas para reduzir o ruído nos dados.
- **Demojização (Demojize):** Emojis foram convertidos para sua representação textual correspondente (ex: ‘‘ se torna ‘:rosto$_c horando_d e_r ir$: ‘).$Essa etapa é crucial para que o modelo de$

2.2 Análise de Sentimentos

Para a classificação de sentimentos, foi utilizado um modelo pré-treinado robusto, o `cardiffnlp/twitter-xlm-roberta-base-sentiment`, através da biblioteca Transformers. Este modelo é especializado em textos de redes sociais e classifica cada comentário em uma de três categorias: `positive`, `neutral` ou `negative`. Além do rótulo, o modelo fornece um **score de confiança** (de 0 a 1), que indica o grau de certeza da classificação.

2.3 Modelagem de Tópicos Híbrida com BERTopic e Gemini

A modelagem de tópicos foi a etapa central da análise, utilizando o BERTopic, que funciona da seguinte forma:

1. **Embeddings de Documentos:** Cada comentário foi transformado em um vetor numérico (embedding) usando o modelo `rufimelo/bert-large-portuguese-cased-sts`, que captura o contexto semântico do texto.

2. **Redução de Dimensionalidade:** A dimensionalidade dos embeddings foi reduzida com UMAP, preservando a estrutura semântica em um espaço de menor dimensão.

3. **Clusterização:** Os documentos foram agrupados em clusters (tópicos) com base em sua proximidade no espaço reduzido, utilizando o algoritmo HDBSCAN.

Após todo este processo, finalmente, o número de tópicos foi reduzido para 50 para garantir uma análise mais focada e com maior interpretabilidade.

O ponto de inovação foi a criação de uma classe personalizada, `GeminiDocsRefiner`, para a etapa de representação de tópicos. Em vez de usar TF-IDF ou outras técnicas baseadas em palavras-chave, esta classe envia uma amostra de comentários de cada tópico para a API do Gemini com um prompt específico para gerar uma descrição rica e em parágrafo.

```python
class GeminiDocsRefiner(BaseRepresentation):
    def __init__(self, api_key: str, model: str = "gemini-1.0-pro", ...):
        # ... inicialização ...
        self.prompt_template = (
            "Escreva uma descrição de um parágrafo que descreva "
            "detalhadamente o que os comentários do instagram "
            "presentes neste tópico tem em comum: {documents}"
        )

    def extract_topics(self, topic_model, documents, c_tf_idf, topics):
        # ... lógica para amostrar documentos e chamar a API do Gemini ...
        response = genai.GenerativeModel(self.model).generate_content(prompt)
        label = response.text.strip()
        # ... lógica para atualizar os rótulos do tópico ...
```

A abordagem com Gemini permitiu gerar descrições ricas para cada cluster. Por exemplo, um tópico foi descrito como: *"Os comentários compartilham uma forte demanda e frustração em relação ao concurso da FUNSAÚDE, especificamente quanto à reposição de vagas... e uma cobrança direcionada a figuras políticas como Elmano de Freitas e Camilo Santana."*. Isso demonstra a clareza e o poder interpretativo alcançados com este modelo híbrido.

3 Resultados e Análise

3.1 Resultados da Análise de Sentimentos

A análise de sentimentos revelou uma predominância massiva de comentários positivos. A distribuição quantitativa foi a seguinte:

- **Positivos:** 4788 comentários
- **Neutros:** 955 comentários
- **Negativos:** 918 comentários

Essa distribuição é visualmente confirmada no Gráfico de Barras da Figura 1. O Histograma de Scores de Confiança mostra que o modelo realizou a maioria de suas classificações com alta confiança (scores próximos de 1.0), indicando robustez. O Boxplot detalha essa confiança por categoria, mostrando que as classificações `positive` e `negative` possuem scores consistentemente mais altos do que a categoria `neutral`, o que é esperado, já que a neutralidade é semanticamente mais ambígua.

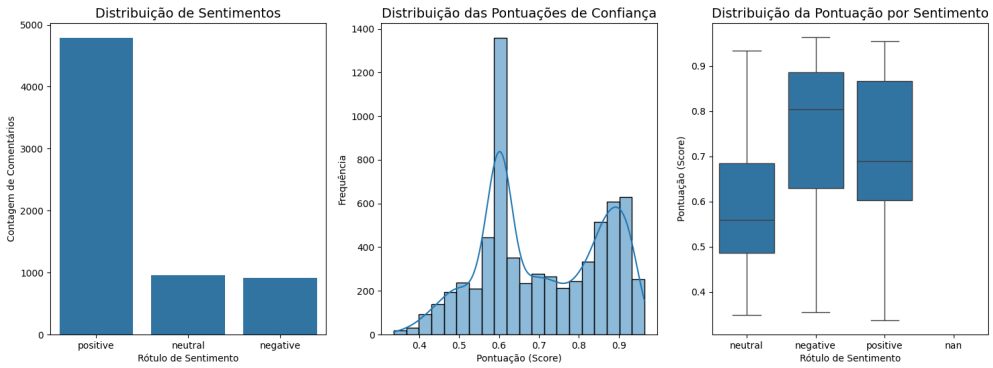

Figura 1: Visualizações da Análise de Sentimento: (a) Distribuição de Rótulos, (b) Distribuição dos Scores de Confiança, (c) Boxplot de Scores por Sentimento.

3.2 Resultados da Modelagem de Tópicos

A modelagem de tópicos identificou 50 temas distintos no corpus. As visualizações interativas fornecem insights profundos sobre a estrutura desses temas.

3.2.1 Mapa de Distância Intertópica

O mapa da Figura 2 (Intertopic Distance Map) posiciona os tópicos em um espaço 2D com base em sua similaridade semântica. Cada círculo representa um tópico, e seu tamanho é proporcional à sua frequência.

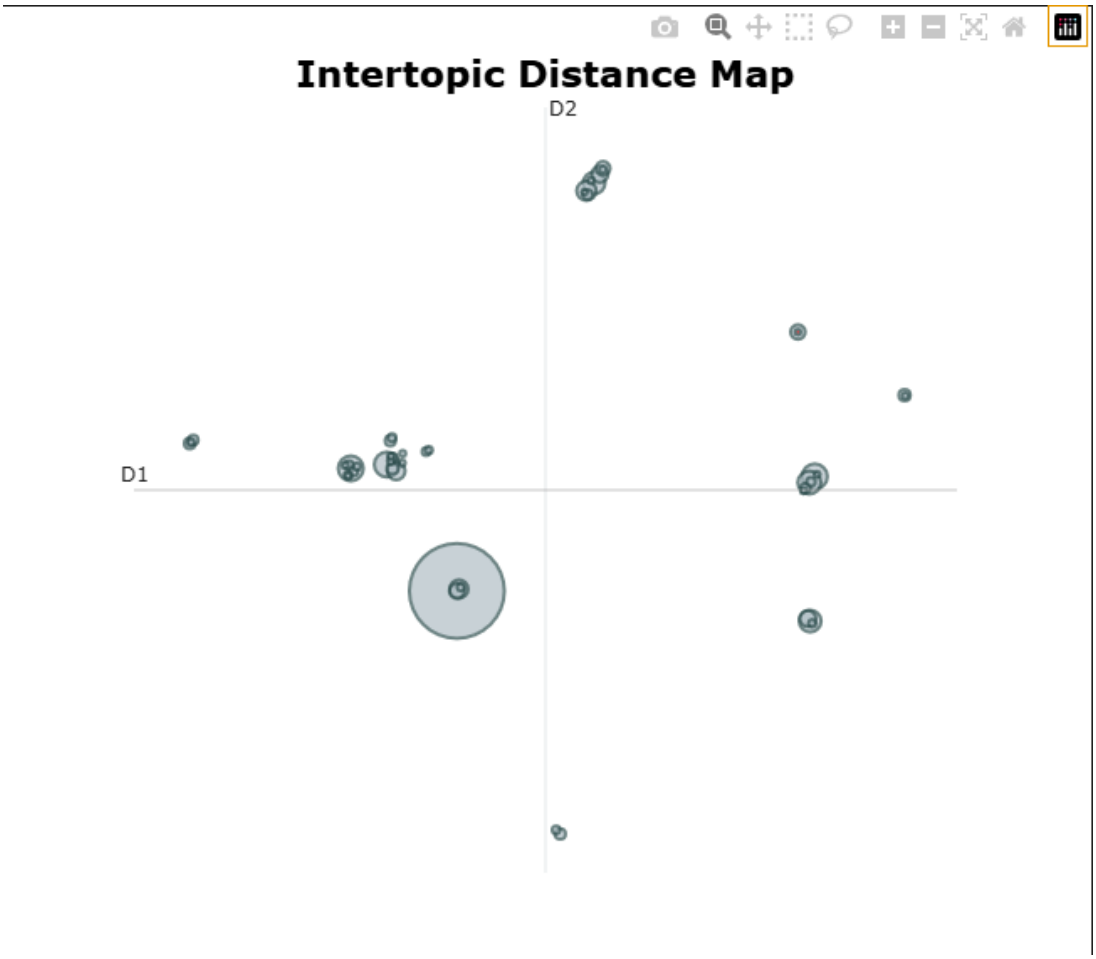

Figura 2: Mapa de Distância Intertópica. Círculos próximos são semanticamente similares.

A análise do mapa revela um **tópico dominante** (círculo grande), que representa a maioria dos comentários e é composto principalmente por expressões de apoio e emojis positivos (ex: aplausos, corações). Os demais tópicos, menores e mais dispersos, formam clusters que abordam assuntos mais específicos, como política, críticas a serviços públicos e pedidos de justiça.

3.2.2 Clusterização Hierárquica e Matriz de Similaridade

O dendrograma de clusterização hierárquica (Figura 3) e o mapa de calor de similaridade (Figura 4) complementam a análise, mostrando como os tópicos se relacionam e podem ser agregados em meta-tópicos. Por exemplo, tópicos relacionados a críticas sobre concursos públicos e serviços de saúde, embora distintos, mostram-se semanticamente próximos e poderiam ser agrupados sob um tema maior de "Reivindicações de Serviços Públicos".

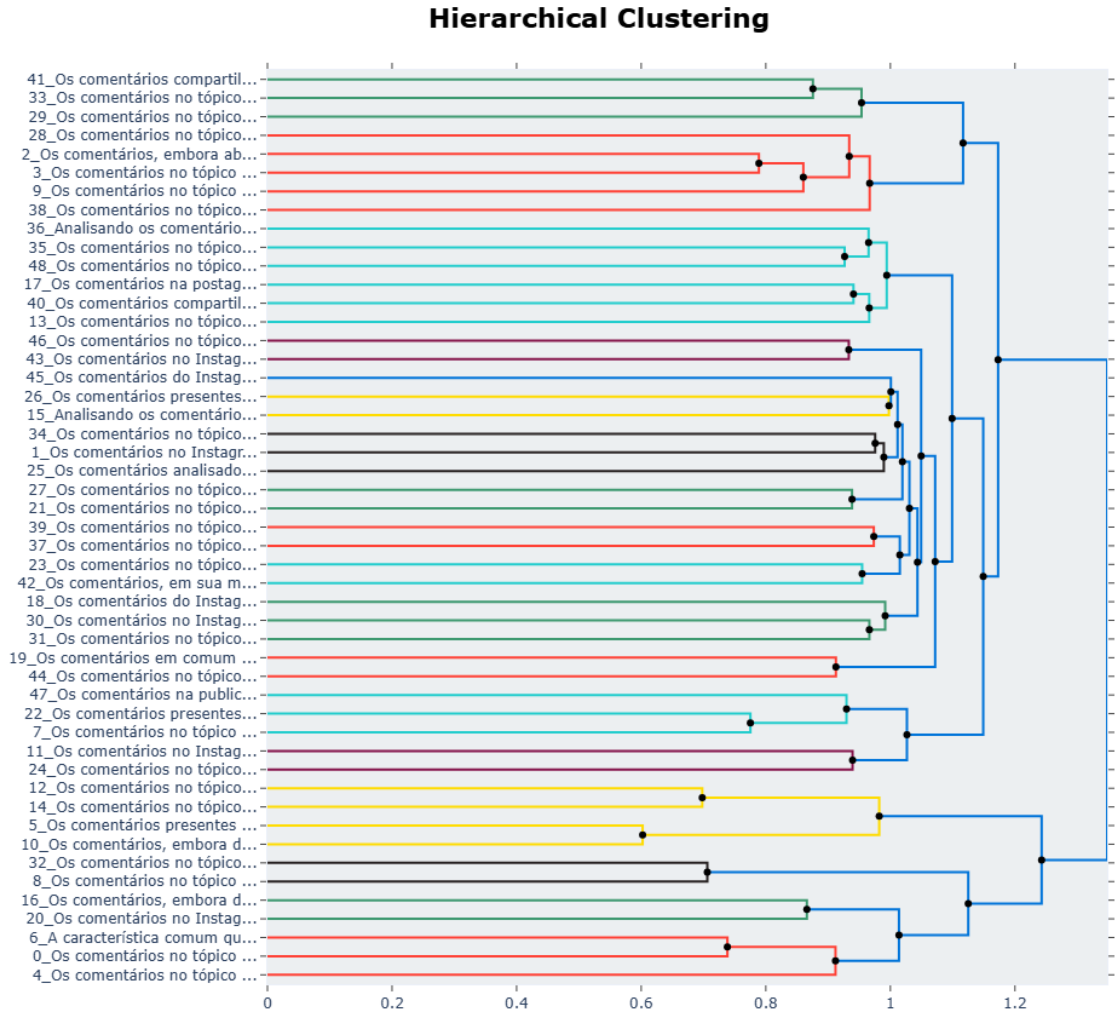

Figura 3: Dendrograma da Clusterização Hierárquica dos Tópicos.

Com base na análise do dendrograma de clusterização hierárquica, a modelagem de tópicos mais coesa e distinta agruparia os comentários em quatro clusters principais. A justificativa para essa escolha reside em observar as distâncias horizontais em que os agrupamentos são formados. Ao traçar uma linha de corte vertical imaginária no eixo de distância em aproximadamente 1.1, identificamos quatro agrupamentos distintos que são fundidos em etapas posteriores com uma distância (dissimilaridade) significativamente maior, indicado pelas longas linhas horizontais que os conectam. O primeiro tópico hierárquico (topo, em tons de verde e vermelho) representa um tópico coeso. O segundo (meio-superior, em ciano, amarelo e preto) é o maior e mais diverso grupo, mas ainda assim claramente distinto dos demais. O terceiro (meio-inferior, em marrom e roxo) e o quarto (base do gráfico, em amarelo, cinza e vermelho) formam os dois últimos tópicos. Esta divisão em quatro grupos otimiza a modelagem, pois maximiza a similaridade dos comentários dentro de cada grupo, ao mesmo tempo que garante que os tópicos sejam significativamente diferentes entre si.

Os itens do primeiro tópico hierárquico (tópicos: 2, 3, 9, 29, 33, 36, 38, 41), têm em comum a descrição de comentários que expressam um forte sentimento de apoio,

concordância e aprovação. A análise foca em reações positivas, como elogios a iniciativas e concordância com o conteúdo postado. Além disso, muitos itens descrevem ações que os usuários desejam realizar ou estão com dificuldade para fazer, como encontrar um link para inscrição ou cadastro, o que os une pelo tema de "ação do usuário". Quanto aos itens do segundo tópico hierárquico (tópicos: 1, 17, 21, 23, 25, 26, 28, 34, 35, 37, 39, 40, 42, 43, 45, 46), a principal característica deste grupo é a contextualização. As análises aqui presentes se concentram em identificar a origem dos comentários (plataforma, tipo de postagem) e frequentemente mencionam figuras de autoridade, como governadores, políticos ou outras personalidades públicas. O tema central é descrever a quem os comentários se dirigem e de onde vêm. Já os itens do terceiro tópico hierárquico (tópicos: 7, 11, 12, 14, 18, 19, 22, 24, 30, 31, 44, 47) agrupa as análises que se aprofundam nas qualidades e sentimentos dos comentários. Os textos descrevem o tom (irônico, crítico, de apoio), o estilo da linguagem, e as emoções predominantes, como admiração, insatisfação ou preocupação. O elo comum é a qualificação do discurso, indo além do tema para capturar a nuance e a subjetividade das mensagens. Por fim, os itens do quarto tópico hierárquico (tópicos: 4, 5, 6, 8, 10, 16, 20, 27, 32) são unidos por um forte viés de crítica, insatisfação e debate. As análises descrevem comentários que abordam problemas sociais, fazem reclamações, apontam divergências e expressam descontentamento com temas como gestão pública, segurança, educação e corrupção. É o cluster que representa a análise do discurso de oposição e conflito.

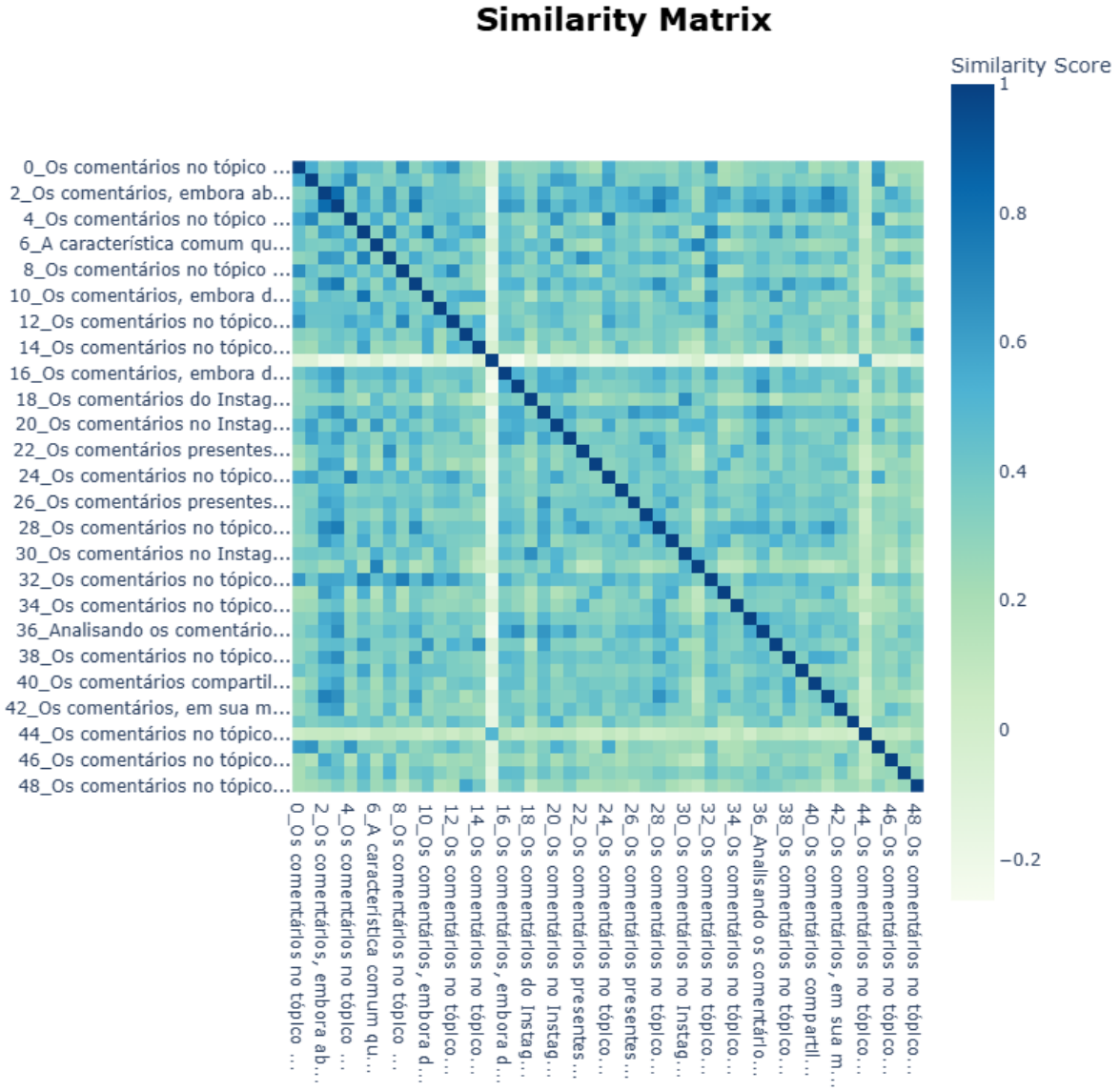

Figura 4: Matriz de Similaridade (Heatmap) entre os Tópicos.

Analisando a matriz de similaridade, a modelagem de tópicos agruparia os comentários em aproximadamente quatro clusters principais, que são visualmente identificáveis como "quadrados"de cores mais escuras (tons de azul) ao longo da diagonal. A lógica para essa divisão é que os itens dentro de cada um desses quadrados demonstram uma alta pontuação de similaridade mútua, indicando que compartilham o mesmo tema central. O primeiro agrupamento pode ser visto no canto superior esquerdo (itens 0 a 8). Um segundo grupo coeso é observável no meio da matriz (aproximadamente itens 20 a 26). Um terceiro cluster é notado logo abaixo (itens 30 a 36). Finalmente, o agrupamento mais forte e coeso parece ser o do canto inferior direito (itens 40 a 48), que exibe os tons de azul mais escuros e consistentes, sugerindo uma altíssima similaridade interna. As áreas de cores mais claras (verdes/amarelas) entre esses blocos azuis representam a baixa similaridade entre os diferentes tópicos, justificando a sua separação em grupos distintos.

4 Conclusão

A aplicação do pipeline híbrido de NLP revelou com sucesso as principais características do corpus de comentários do Instagram. A análise de sentimentos indicou uma recepção majoritariamente positiva ao conteúdo dos Reels, com alta confiança do modelo classificador.

A modelagem de tópicos híbrida, combinando BERTopic e Gemini, provou ser excepcionalmente eficaz. Ela não apenas identificou a estrutura temática dos dados—com um tópico principal de reações positivas e diversos subtópicos específicos—mas também forneceu descrições contextuais e de fácil interpretação para cada tema. Esta abordagem representa um avanço significativo em relação aos métodos tradicionais, permitindo uma análise qualitativa mais profunda e a extração de insights acionáveis com maior facilidade. Os resultados demonstram o potencial da integração de modelos de embedding e modelos generativos para análises de texto complexas.