Trabalho de Mineração e Análise de Redes Sociais - MARS
Rian Wagner Costa - 212050103

1 - Caracterização topológica de redes sociais:

Esta etapa apresenta uma análise de redes baseada nos 5.499 comentários extraídos de 8 vídeos do YouTube, partindo de pesquisas relacionadas com o nome “Elon Musk”. Os vídeos selecionados foram postados por emissoras e agências notórias e com grande público. A análise consiste na construção de um grafo onde os nós representam palavras e as arestas indicam a ocorrência dessas palavras dentro de um mesmo comentário. O objetivo é identificar palavras-chave e compreender sua conectividade dentro da rede de comentários.

Para fins de observação, a Figura 1 apresenta a rede quando montada considerando somente a similaridade dos comentários

Figura 1

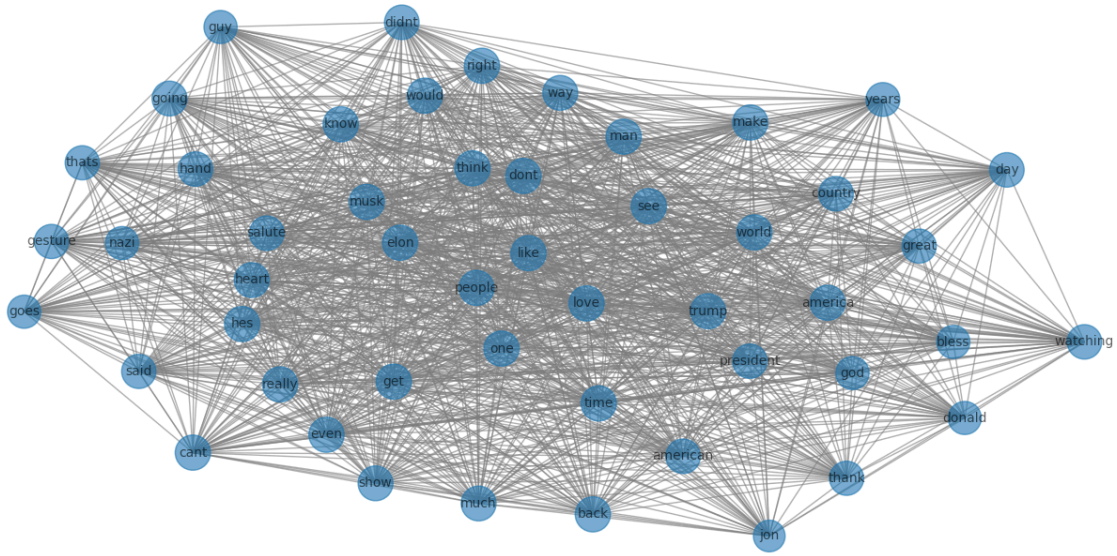
Como pode ser observado, considerando uma similaridade de 0.8, a rede se tornaria visualmente incompreensível. No entanto, essa complexidade evidencia o alto grau de semelhança entre os comentários, que, apesar de transmitirem significados distintos, utilizam um vocabulário comum.

Quando analisada somente a ocorrência de palavras, já é possível observar a rede de uma forma mais limpa, como na Figura 2 onde a rede foi montada considerando as 50 palavras mais repetidas nos comentários.

- **Propriedades da Rede:**
  Número de nós: 50
  Número de arestas: 1142
  Coeficiente de clustering medio: 0.96

Trabalho de Mineração e Análise de Redes Sociais - MARS
Rian Wagner Costa - 212050103

Figura 2



- **Distribuição de graus:**
  Distribuição de graus (top 10): [('elon', 49), ('trump', 49), ('like', 49), ('people', 49), ('america', 49), ('love', 49), ('musk', 49), ('hes', 49), ('man', 49), (‘president’, 49).

- Centralidade:

  Duas medidas de centralidade foram utilizadas para identificar os nós mais importantes:

  Centralidade de Grau**:** Mede a quantidade de conexões diretas que um nó possui.
  Centralidade de grau (top 10): [('elon', 0.9999999999999999), ('trump', 0.9999999999999999), ('like', 0.9999999999999999), ('people', 0.9999999999999999), ('america', 0.9999999999999999), ('love', 0.9999999999999999), ('musk', 0.9999999999999999), ('hes', 0.9999999999999999), ('man', 0.9999999999999999), ('dont', 0.9999999999999999)]

  Centralidade de Eigenvector**:** Mede a influência de um nó considerando a importância de seus vizinhos.
  Centralidade de eigenvector (top 10): [('elon', 0.14599696840301332), ('trump', 0.14599696840301332), ('like', 0.14599696840301332), ('people', 0.14599696840301332), ('america', 0.14599696840301332), ('love', 0.14599696840301332), ('musk', 0.14599696840301332), ('hes', 0.14599696840301332), ('man', 0.14599696840301332), ('dont', 0.14599696840301332)]

Conclusão: A análise permitiu identificar as palavras mais utilizadas e suas relações dentro dos comentários, proporcionando insights sobre os principais temas

discutidos. Redes desse tipo podem ser úteis para análises de sentimento, detecção de spam e tendências em redes sociais.

2 - **Descoberta de Conhecimento em Bases de Dados de Redes Sociais:**

- Coleta de Dados:

Os dados utilizados nesta atividade foram coletados diretamente da plataforma YouTube, via scraping, contendo informações sobre comentários de usuários nos vídeos selecionados anteriormente. Cada entrada de dados inclui os seguintes atributos:

- number: Número identificador do comentário.
- user: Nome de usuário que fez o comentário.
- comment_text: Texto do comentário.
- time_ago: Tempo desde que o comentário foi feito.
- likes: Número de curtidas no comentário.
- number_responses: Número de respostas ao comentário.

Esses dados foram armazenados no formato JSON, o que permitiu uma fácil extração e conversão para um formato tabular adequado para análise.

- Processamento de Dados:

Primeiramente, os dados foram carregados e transformados em um DataFrame do pandas. Para garantir que a análise fosse realizada corretamente, alguns ajustes foram feitos:

- Remoção de campos desnecessários: Como o foco foi analisar os comentários, apenas os campos relevantes foram mantidos.
- Tratamento de dados ausentes: Foram identificados valores ausentes ou inconsistentes e tratados de acordo, com a aplicação de técnicas como substituição por média ou remoção das linhas problemáticas.
- Extração de novas variáveis: A partir do campo “comment_text”, foi gerada uma nova variável “comment_length”, que representa o comprimento do comentário em número de caracteres, como uma métrica adicional para a análise.

As estatísticas básicas para os atributos “likes”, “number_responses” e “comment_length” foram calculadas, incluindo:

- Mínimo: Valor mínimo dos atributos.
- Máximo: Valor máximo dos atributos.
- Média: Valor médio dos atributos.
- Mediana: Mediana dos valores dos atributos.

- Desvio padrão: Medida da dispersão dos valores dos atributos.

Essas estatísticas forneceram uma visão inicial da distribuição dos dados e ajudaram a identificar possíveis outliers ou anomalias.

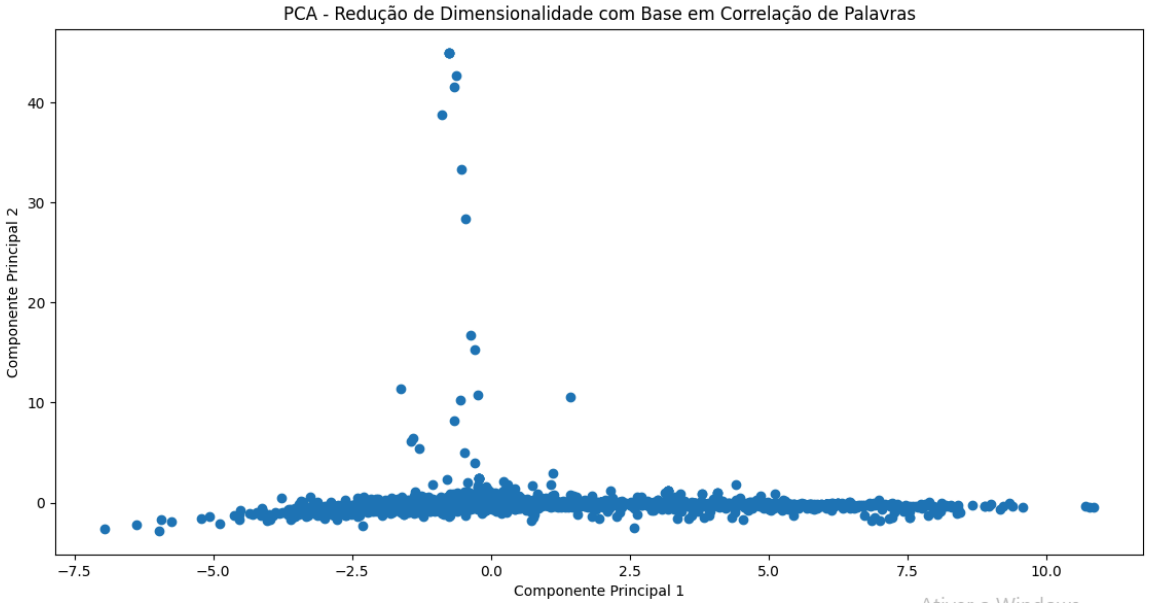
Estatísticas básicas (Atributos numéricos):		
	Curtidas	Número de Respostas
Total de comentários	5499	5499
Média	25.23	1.99
Desvio padrão	102.51	12.76
Valor mínimo	0	0
Valor máximo	997	437

- **Análise de Correlação e Redução de Dimensionalidade:**

A rede de correlação, montada anteriormente, entre as palavras foi construída, levando em consideração a frequência de ocorrências de termos dentro dos comentários. Isso permitiu identificar termos frequentemente utilizados em conjunto, como palavras relacionadas ao tópico dos vídeos.

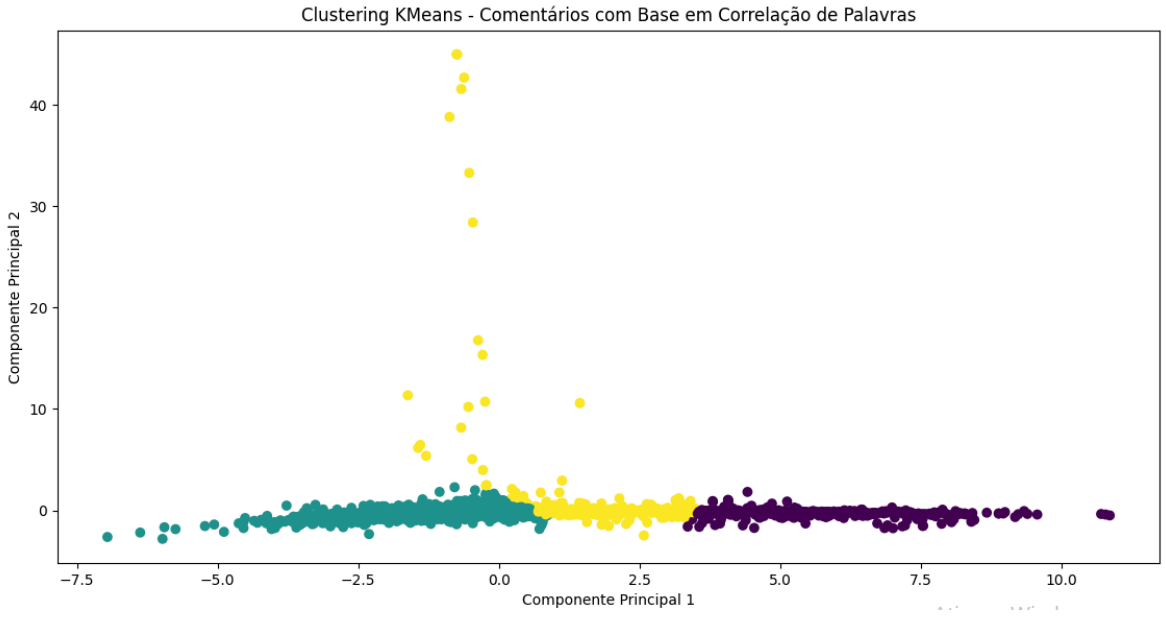
Para reduzir a dimensionalidade dos dados e facilitar a visualização, foi aplicada a técnica de PCA (Análise de Componentes Principais). A redução foi feita para duas componentes principais, mantendo a maior parte da variância dos dados extraídos dos comentários, que foram vetorizados com a técnica de TF-IDF para capturar a correlação das palavras. O gráfico gerado, Figura 3, mostrou como os comentários podem ser agrupados com base nas principais características das palavras presentes nos textos. Essa redução de dimensionalidade permitiu observar os padrões de agrupamento dos comentários, evidenciando os clusters formados com base nas semelhanças de seus conteúdos textuais.

Figura 3



A técnica de K-Means foi utilizada para realizar o clustering dos comentários. O algoritmo foi configurado para dividir os dados em 3 clusters, com base nas duas primeiras componentes principais extraídas do PCA. O gráfico, Figura 4, gerado pela aplicação do KMeans mostrou claramente a separação dos dados em três grupos distintos, com comentários agrupados conforme características semelhantes, como número de curtidas, respostas e comprimento.

Figura 4



Conclusão: Através da aplicação de PCA e KMeans, foi possível reduzir a dimensionalidade dos dados de comentários e agrupar os comentários com base em suas semelhanças textuais. Isso facilita a análise do comportamento dos usuários e permite identificar padrões de discussão dentro dos comentários. A redução de dimensionalidade e o agrupamento dos dados possibilitaram insights valiosos sobre como os comentários podem ser agrupados e categorizados com base nas palavras e temas abordados, ao invés de apenas atributos numéricos como curtidas ou respostas.

3 - Análise de texto em redes sociais:

O objetivo principal desta atividade é aplicar as técnicas de mineração de texto em redes sociais, abordando as etapas desde a coleta dos dados até a análise e visualização dos tópicos. O foco não está na identificação de padrões relevantes, mas sim na execução das técnicas de pré-processamento de texto, transformação em representação vetorial e análise de texto, complementando as atividades anteriores de KDD

- **Pré-processamento dos Textos:**

O pré-processamento dos textos envolveu as seguintes etapas:

- Remoção de stopwords: As palavras comuns que não contribuem significativamente para a análise foram removidas.
- Remoção de pontuação: Todos os sinais de pontuação foram removidos para evitar interferências na análise.
- Transformação para minúsculas: Os textos foram convertidos para letras minúsculas para garantir consistência.
- Lematização: As palavras foram reduzidas à sua forma base (como "running" para "run") para normalizar as variações linguísticas.
- Tokenização: O texto foi dividido em palavras ou tokens individuais.

- **Representação Vetorial:**

Após o pré-processamento, os textos foram transformados em uma representação vetorial. Para isso, utilizamos o TF-IDF (Term Frequency-Inverse Document Frequency) para medir a importância de cada palavra em relação ao corpus de comentários. O TF-IDF é uma técnica de vetorização que ajuda a identificar as palavras mais significativas em um conjunto de textos.

- **Modelagem de Tópicos com LDA (Latent Dirichlet Allocation):**

Trabalho de Mineração e Análise de Redes Sociais - MARS
Rian Wagner Costa - 212050103

O LDA foi utilizado para identificar os tópicos presentes nos comentários. A técnica de modelagem de tópicos foi aplicada para entender os principais assuntos discutidos nos comentários, com base nas palavras mais frequentes, expostas nas Figuras 5, 6 e 7, e no contexto em que aparecem.

Figura 5 - Tópico 0

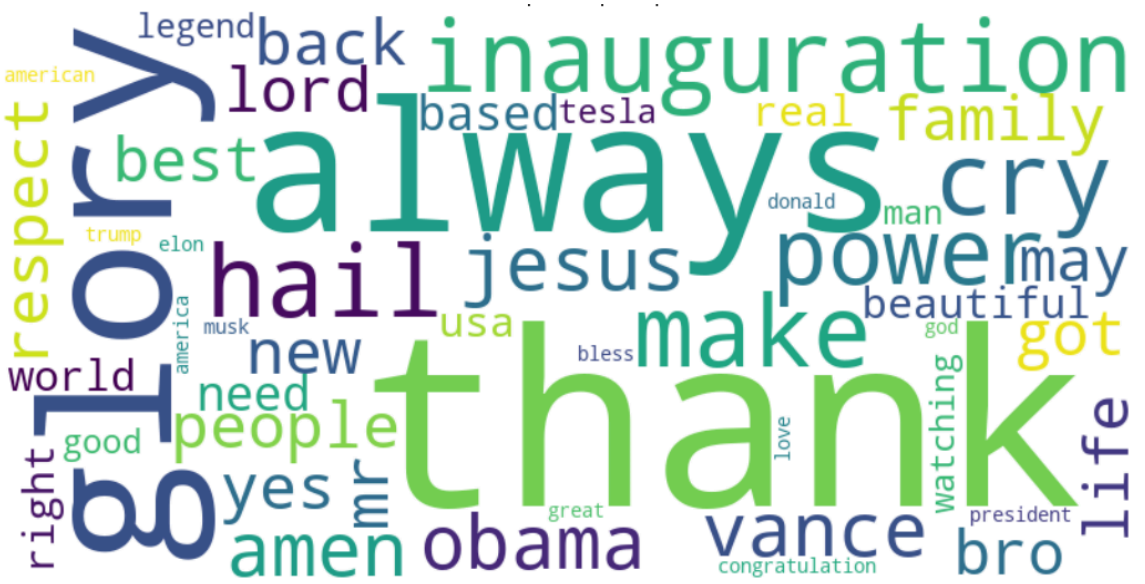


Figura 6 - Tópico 1

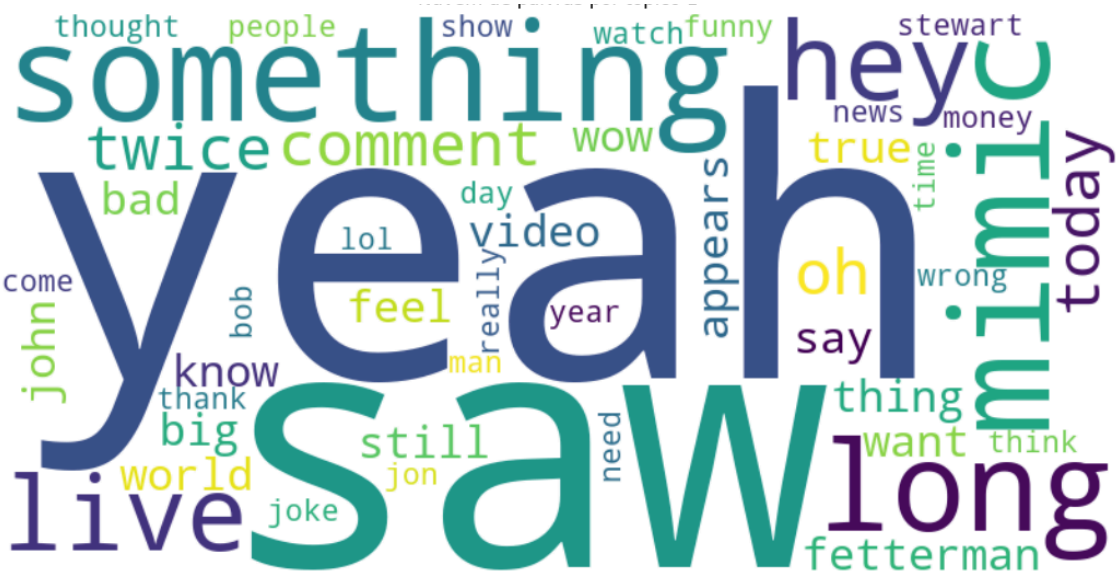
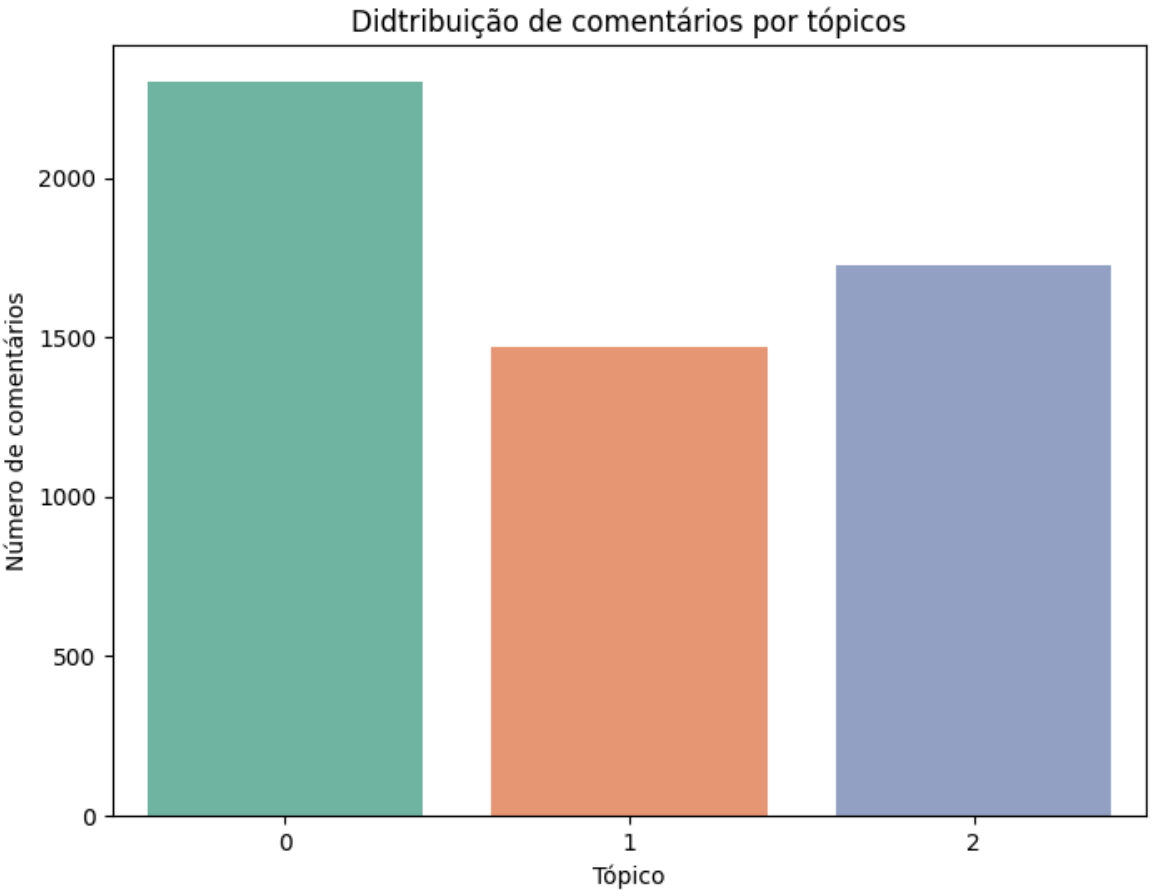


Figura 7 - Tópico 2

Para melhor visualizar a estrutura de tópicos do LDA, a Figura 8 trata da distribuição de comentários em cada tópico.

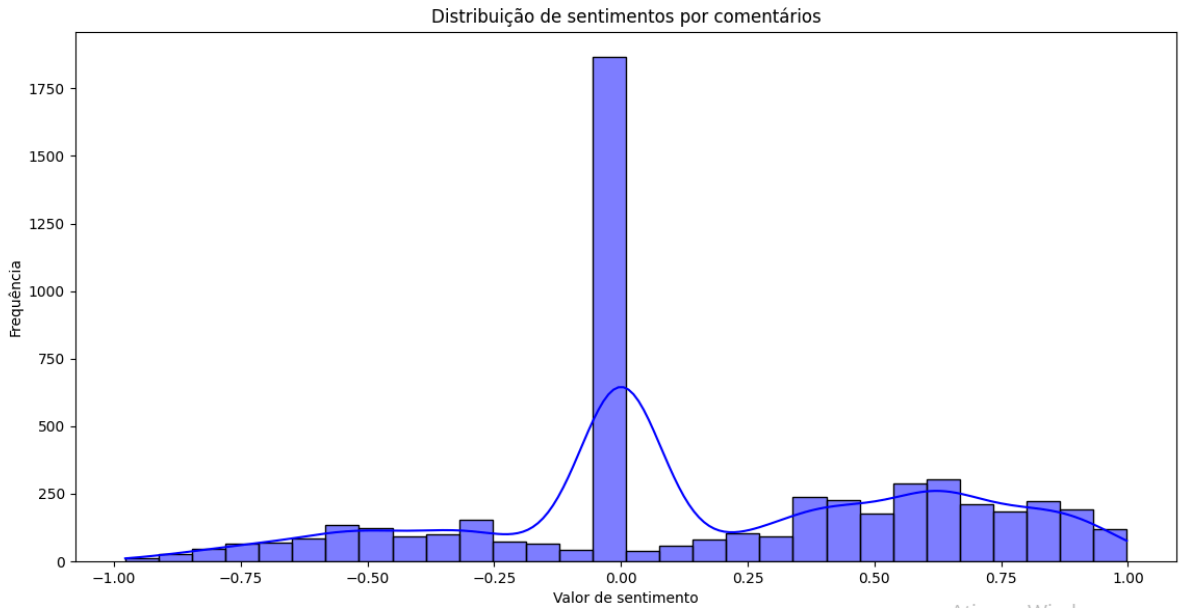
Figura 8



- Análise de Sentimento:

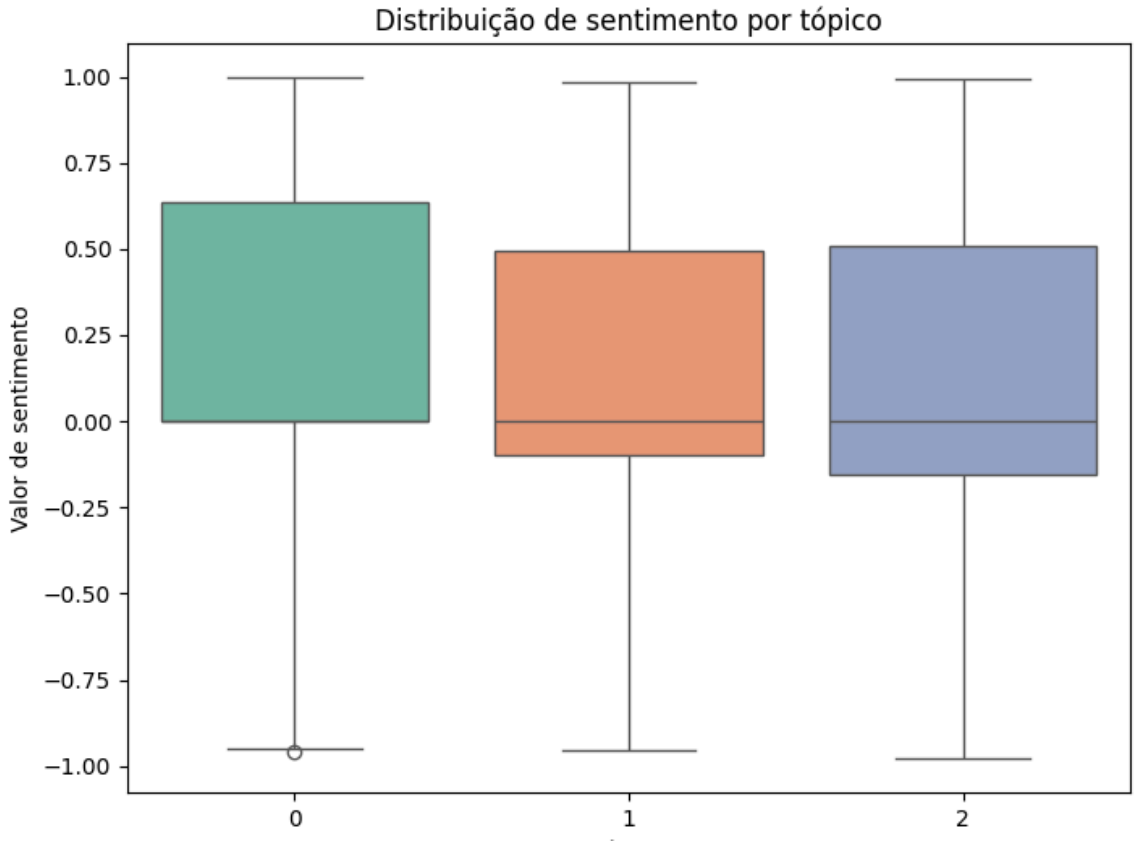
A análise de sentimento foi aplicada para avaliar o tom emocional dos comentários coletados, classificando-os em categorias como positivo, negativo ou neutro, Figura 9. Para isso, utilizamos uma ferramenta de análise de sentimentos, que pode ser baseada em modelos pré-treinados ou em abordagens simples como o VADER (Valence Aware Dictionary and sEntiment Reasoner), que é um léxico de sentimentos amplamente utilizado para textos curtos, como os comentários em redes sociais.

Figura 9



Para melhor visualização a Figura 10 trata da distribuição de sentimentos por tópicos

Figura 10



Conclusão: A aplicação da análise de sentimento foi eficaz para identificar a polaridade emocional dos comentários. A visualização da distribuição dos sentimentos forneceu uma visão clara de como os usuários se sentem em relação ao tema abordado nos vídeos. A integração dessa técnica com a modelagem de tópicos permite uma análise mais robusta dos dados textuais, pois possibilita a compreensão não apenas dos tópicos discutidos, mas também do tom e das emoções associadas a esses tópicos.