

Análise de Sentimento em Comentários de Vídeos em português do TikTok e Correlação com Hashtags: Uma Investigação dos Padrões de Reação e Temas Emergentes

1stRafael Vinícius dos Santos
Departamento de Cincia da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
ORCID: 0000-0002-6534-5785

Abstract—A análise de sentimento em mídias sociais tem se tornado uma ferramenta para entender as opiniões e comportamentos dos usuários a partir de suas interações em plataformas digitais. Este trabalho investiga os padrões de sentimento expressos em comentários de vídeos no TikTok em português, focando na correlação com as hashtags associadas. O TikTok, uma das plataformas de vídeo mais populares globalmente, oferece um vasto conjunto de dados gerados pelos usuários, refletindo suas reações e emoções em relação a diversos temas. Utilizando técnicas de processamento de linguagem natural e análise de dados, este estudo busca identificar e categorizar os sentimentos predominantes (positivos, negativos e neutros) e correlacioná-los com as hashtags utilizadas. Os resultados revelam padrões emergentes nos temas que geram diferentes tipos de reações e fornecem informações sobre as tendências atuais. As conclusões deste estudo têm implicações para a compreensão das dinâmicas de interação nas mídias sociais.

I. INTRODUÇÃO

A ascensão da mídia social tem gerado um imenso volume de dados produzidos pelos usuários, oferecendo uma ampla visão sobre suas opiniões, interesses e comportamentos. Entre essas plataformas, o aplicativo de vídeos curtos, TikTok, destaca-se pelo seu crescimento exponencial e pela sua capacidade de engajamento. Desde a sua fundação, o TikTok tem se consolidado como um dos aplicativos de vídeos curtos mais baixados e utilizados no mundo, com mais de um bilhão de usuários ativos mensais. No Brasil, o aplicativo conta com aproximadamente 80 milhões de usuários adultos, número menor apenas que Estados Unidos e Indonésia, posicionando-se como uma das principais redes sociais do país (Revista Exame, 2023).

A análise de sentimento é uma técnica de processamento de linguagem natural (PLN) que visa identificar o sentimento transmitido por um determinado texto. Este trabalho se limitou a analisar somente se o comentário realizado era positivo ou negativo, não visando extrair opiniões ou sentimentos mais complexos. Essa técnica pode ser utilizada para entender as atitudes dos usuários em relação a produtos, serviços e eventos. No contexto das mídias sociais, a análise de sentimento pode revelar como os usuários se sentem em relação a diversos

temas, pesquisadores a compreender melhor as dinâmicas das interações online.

A motivação deste trabalho foi o fato da rede social em específico estar em grande ascensão e de que a o uso de técnicas de análise de sentimentos em redes sociais é algo novo, uma vez que as redes sociais tomaram grande proporção há menos de 20 anos. Apesar da existência de alguns trabalhos realizando esta análise tomando em base dados do twitter [1] ou Reddit [2], são muito raros trabalhos voltados para o Tiktok ou para a língua portuguesa.

Nesse contexto, este trabalho propõe uma investigação detalhada da análise de sentimento em comentários em português de vídeos do TikTok sobre tópicos emergentes, com o objetivo de compreender os padrões de reação dos usuários e sua correlação com as hashtags associadas aos vídeos. Por meio da coleta e análise sistemática de dados, pretendemos identificar os principais temas e tópicos que geram comentários positivos ou negativos, contribuindo para uma compreensão mais profunda dos assuntos que geram este tipo de reação nos usuários.

Para isso, foi realizada a coleta de aproximadamente 2 mil vídeos e cerca de 100 mil comentários. As hashtags emergentes foram selecionadas com base em análises do Google Trends, TikTok Ads e do site oficial do TikTok e para a tarefa de análise de sentimentos, um modelo BERT (Bidirectional Encoder Representations from Transformers) treinado para o português foi selecionado. O BERT é o estado-da-arte em PLN, sendo capaz de capturar contextos bidirecionais em textos, o que o torna especialmente eficaz para a análise de sentimento.

II. REFERENCIAL TEÓRICO

Estudos relacionados à análise de sentimentos, também conhecida como *Opinion Mining*, têm ocupado uma parcela significativa das pesquisas em processamento de linguagem natural desde o início dos anos 2000. O objetivo central dessa abordagem é desenvolver ferramentas automatizadas capazes de extrair informações subjetivas a fim de criar

conhecimento estruturado [3]. A diversidade de técnicas propostas para realizar tais tarefas inclui diferentes estratégias de pré-processamento de dados e ferramentas baseadas em aprendizado de máquina, com abordagens supervisionadas e não supervisionadas [4]

As técnicas supervisionadas, dependem de sentenças pré rotuladas de acordo com sua polaridade, podendo ser **positiva**, **negativa** ou **neutra**, enquanto abordagens não supervisionadas possuem uma abordagem lexica, ou seja, assumem que palavras individuais possuem uma polaridade prévia, independente do seu contexto.

Em 2017, Vaswani [5] propôs uma nova arquitetura de rede neural chamada *transformer*, que aprende as relações contextuais entre palavras e possibilitou a criação de modelos como *GPT* e *BERT*, os quais revolucionaram as tarefas de processamento de linguagem natural. [6]

Apesar do grande esforço em desenvolvimento de modelos e para análise de sentimentos em redes sociais, como o trabalho de Giachonou [1], a quantidade de modelos e estudos voltados para o português são baixas. Uma revisão das técnicas utilizadas para análise de mídias sociais em português brasileiro foi feita por Cirqueira [7]

III. DESENVOLVIMENTO

O presente capítulo objetiva descrever de forma detalhada os processos e atividades realizadas durante o desenvolvimento deste trabalho.

A. Seleção inicial das hashtags

A seleção das hashtags utilizadas no estudo foi realizada através de uma abordagem multifacetada, incorporando dados de duas fontes principais: Google Trends e TikTok Ads. O período de análise foi o primeiro quadrimestre de 2024, garantindo a relevância e a atualidade das hashtags escolhidas.

1) *Google Trends*: Inicialmente, foi realizado um estudo dos tópicos mais pesquisados no domínio brasileiro do Google (google.br) utilizando o Google Trends. Esta ferramenta permite analisar a popularidade das consultas de pesquisa ao longo do tempo. O Google Trends permite filtrar os resultados por região geográfica, categoria e período de tempo, o que possibilitou um enfoque específico no Brasil durante os primeiros quatro meses de 2024.

2) *TikTok Ads*: Paralelamente, foi utilizada a plataforma TikTok Ads, o portal oficial do TikTok que divulga as hashtags mais relevantes e em alta na plataforma. Esta fonte reflete diretamente as tendências e os interesses dos usuários do TikTok, além de fornecer dados atualizados sobre o engajamento do conteúdo.

Ademais, a ferramenta do tiktok ads permite também realizar um recorte geográfico, temporal e por categorias pré estabelecidas pela própria plataforma, indicando quais as hashtags são tendências dentro de uma categoria específica como saúde, política, entretenimento entre outras.

O processo de seleção de hashtags através do TikTok Ads incluiu:

- Monitoramento das Hashtags em Alta: Foram monitoradas as hashtags mais populares no TikTok Ads durante o primeiro quadrimestre de 2024.
- Avaliação de Relevância: As hashtags foram avaliadas quanto à sua relevância e frequência de uso na plataforma, garantindo que apenas as hashtags com maior impacto e engajamento fossem selecionadas.
- Validação Cruzada: Houve uma validação cruzada com as hashtags obtidas do Google Trends para identificar convergências e divergências, assegurando uma seleção abrangente e representativa das tendências mais significativas.

3) *Integração das Fontes*: Após a coleta e avaliação das hashtags de ambas as fontes, foi realizada uma integração dos dados. As hashtags, criando um conjunto robusto e diversificado de hashtags para análise subsequente. Este processo garantiu que as hashtags utilizadas no estudo fossem representativas das tendências atuais e refletissem um amplo espectro de interesses dos usuários tanto no Google quanto no TikTok.

B. Coleta dos vídeos e enriquecimento do conjunto de hashtags

Para a coleta dos vídeos do TikTok com base nas hashtags inicialmente selecionadas, foi utilizada a plataforma APIfy. Esta plataforma realiza a conexão com diversas APIs, incluindo a do tiktok e possui vários atores dentro do sistema, sendo cada ator responsável por realizar algum tipo de coleta específico. Foi utilizado o ator para coleta de posts criando assim o dataset de vídeos contendo as seguintes informações:

- URL do Vídeo: Servindo como um identificador único para cada vídeo.
- Quantidade de Comentários
- Quantidade de Curtidas
- Autor do vídeo
- Data de Publicação
- Descrição do Vídeo

1) *Enriquecimento do Conjunto de Hashtags*: A descrição dos vídeos desempenha um papel crucial no que tange o enriquecimento do conjunto de hashtags. Durante o processo de coleta, as descrições dos vídeos foram analisadas de forma a identificar todas as hashtags contidas no vídeo, incluindo aquelas que não foram selecionadas a princípio mas apareceram em algum dos vídeos coletados.

A partir desta combinação de dados, foi possível criar a relação das hashtags que apareciam com maior frequência e identificando assim quais as maiores tendências entre os vídeos coletados inicialmente e também possível criar para cada vídeo do dataset, uma lista de todas suas hashtags.

O pré-processamento das hashtags também passou pelo tratamento de texto de forma a retirar caracteres especiais e acentuações permitindo o melhor agrupamento de tags que poderiam sofrer com eventuais escritas diferentes.

C. Coleta dos comentários

A partir da coleta dos vídeos e criação do dataset de vídeos, foi criada uma lista contendo todas URLs de vídeos coletados,

de forma a tornar possível a coleta dos comentários também a partir da plataforma APIfy, utilizando um ator específico para coleta de comentários do tiktok a partir da URL do vídeo.

Este ator permite a realização da coleta de vários comentários de um vídeo incluindo as respostas aos comentários. Foi realizada a coleta de uma parte dos comentários de cada vídeo sendo uma média de 60 comentários coletados por vídeos

O dataset de comentários contém para cada registro:

- ID do comentário
- Texto
- URL do vídeo
- Autor do comentário

D. Análise de Sentimentos

A análise de sentimentos em textos, particularmente em contextos de mídias sociais, requer o uso de modelos avançados de processamento de linguagem natural (PLN). Vários modelos foram estudados para esta tarefa, mas atualmente os modelos baseados em transformers, como o BERT (Bidirectional Encoder Representations from Transformers), são considerados o estado da arte em muitas aplicações de PLN [6].

1) *Modelos Baseados em Transformers*: Os transformers são uma arquitetura de rede neural desenvolvida para lidar com sequências de dados, como textos, e são particularmente eficazes para tarefas de compreensão de linguagem. A principal inovação dos transformers é a atenção bidirecional, que permite que o modelo considere o contexto completo de uma palavra em relação a todas as outras palavras em uma sequência. Isso contrasta com os modelos anteriores, que utilizavam uma atenção unidirecional ou eram baseados em redes neurais recorrentes.

O BERT, é um modelo baseado em transformers que foi pré-treinado em uma grande quantidade de texto e pode ser ajustado (fine-tuned) para tarefas específicas, como análise de sentimentos.

2) *Seleção do Modelo*: Dentro da plataforma Hugging Face, que é um importante repositório para modelos de PLN, diversos modelos baseados em BERT estão disponíveis e prontos para uso. Para a tarefa específica de análise de sentimentos em comentários de redes sociais em português, foram considerados vários modelos ajustados (fine-tuned) para esta língua.

Após uma análise cuidadosa dos modelos disponíveis, o modelo que se destacou foi o pysentimiento. Este modelo é construído sobre o BERT, mais especificamente o modelo roBERTa, que é uma variação do BERT que foi otimizada através de mudanças na arquitetura e treinamento. O pysentimiento foi aprimorado com um fine-tuning adicional utilizando um corpus de posts em português do Twitter, o que o torna particularmente bem adaptado para lidar com o tipo de linguagem informal e frequentemente abreviada encontrada em comentários de redes sociais.

3) *Pysentimiento*: O pysentimiento se distingue por:

- Treinamento Especializado: O modelo foi ajustado com um conjunto de dados específico de posts em português

do Twitter, o que o torna eficaz para entender o contexto e as nuances da linguagem informal e gírias utilizadas nas redes sociais.

- Desempenho em Tarefas de Classificação: O desempenho superior do pysentimiento em tarefas de classificação de sentimentos foi observado, especialmente em comparação com outros modelos de PLN disponíveis para o português.

E. Datasets finais

Após a conclusão da tarefa de análise de sentimentos para todos os comentários coletados, o modelo de análise de sentimentos categorizou cada comentário em uma das três classes: Positivo, Negativo ou Neutro. Com base nesses resultados, foram gerados dois conjuntos de dados para análise:

- **Dataset de Vídeos**

- URL do Vídeo.
- A lista de hashtags associadas a cada vídeo.

- **Dataset de Comentários**

- Texto do Comentário: O conteúdo textual de cada comentário.
- Vídeo Relacionado: A URL do vídeo ao qual o comentário está associado, estabelecendo a relação entre o comentário e o vídeo.
- Polaridade do Sentimento: A classificação do sentimento do comentário, determinada pelo modelo de análise de sentimentos (Positivo, Negativo ou Neutro).

Ademais, também é preciso levar em conta a relação entre cada um dos conjuntos de dados da seguinte maneira:

- Um vídeo pode conter vários comentários.
- Um comentário só pode pertencer a um vídeo.
- Um vídeo pode conter várias hashtags.
- Uma hashtag pode estar contida em vários vídeos.

A partir de toda esta preparação de dados, categorização dos sentimentos de cada comentário e a relação entre os dados coletados, podemos analisar os resultados que serão mais amplamente demonstrados no próximo capítulo

IV. RESULTADOS

Após a coleta e o tratamento dos dados descritos no capítulo anterior, torna-se possível realizar uma análise detalhada dos resultados para identificar as correlações entre os comentários e os assuntos emergentes no TikTok.

A. Hashtags

A priori, vamos discutir acerca dos dados coletados para em seguida transcorrer sobre o resultado da análise de sentimentos e suas correlações.

Foram coletados um total de 1900 vídeos distintos da plataforma TikTok, selecionados com base nas hashtags iniciais identificadas através do Google Trends e TikTok Ads. Após o enriquecimento da base de dados, foram identificadas 4470 hashtags distintas. Considerando as repetições, o total de

TABLE III
HASHTAGS POSITIVAS

Hashtag	Quantidade de Comentários Positivos	Proporção
lula	5396	3,52
bolsonaro	5610	3,62
pablomarcial	3548	3,24
brasil	2187	3,11
maternidade	2736	3,81
roca	2400	3,0
futebol	1958	5,31
fitness	1402	3,04

A análise de postagens em redes sociais mostra-se um mecanismo crucial para compreender o comportamento social, à medida que o número de usuários continua a crescer e reflete padrões cada vez mais representativos da vida real. A capacidade de analisar sentimentos e tendências em plataformas como o TikTok oferece uma visão valiosa sobre as dinâmicas de interação dos usuários e suas reações a diferentes temas.

Entretanto, algumas limitações foram identificadas durante o estudo. Um desafio significativo é a dificuldade de rotular grandes quantidades de dados para aprimorar os modelos de PLN, uma vez que esse processo de rotulação é predominantemente manual e, portanto, trabalhoso. Adicionalmente, a coleta de dados no TikTok enfrenta restrições devido ao acesso limitado proporcionado pela API da plataforma, o que pode restringir a abrangência e a profundidade das análises. Apesar dessas dificuldades, o estudo mostrou que as ferramentas e métodos atuais são adequados para realizar análises valiosas e fornecer informações importantes sobre a interação e o sentimento dos usuários nas redes sociais.

Para futuras pesquisas, recomenda-se o desenvolvimento e o treinamento de modelos mais sofisticados de processamento de linguagem natural (PLN) que possam identificar com maior precisão o sentimento específico transmitido pelos textos. Além da análise geral de sentimentos, seria benéfico focar na detecção de discursos de ódio, que representa um problema significativo nas redes sociais contemporâneas.

O discurso de ódio não apenas afeta a qualidade das interações online, mas também pode ter implicações profundas no bem-estar dos usuários e na dinâmica social geral. Identificar e compreender as bolhas de usuários que propagam esse tipo de discurso é um passo crucial para mitigar o problema e promover um ambiente digital mais saudável.

REFERENCES

- [1] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–41, 2016.
- [2] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence," *Journal of Infection and Public Health*, vol. 14, no. 10, pp. 1505–1512, 2021.
- [3] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, *Sentiment analysis in social networks*. Morgan Kaufmann, 2016.
- [4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, Jul. 2002, pp. 79–86. [Online]. Available: <https://aclanthology.org/W02-1011>
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd0
Paper.pdf
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
- [7] D. Cirqueira, M. F. Pinheiro, A. Jacob, F. Lobato, and A. Santana, "A literature review in preprocessing for sentiment analysis for brazilian portuguese social media," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018, pp. 746–749.