Mineração da Informação na Web e Redes Sociais
Tópicos Especiais em Ciência de Dados
Trabalho 3

Pedro Galhardo Germiniani[1] **e Rafael Vinícius Xavier Pimenta**[1]

[1]Faculdade de computação – Universidade Federal de Uberlândia (UFU)

Abstract. *This study explores web and social media information mining, focusing on sentiment analysis related to the Brazilian Unified Health System (SUS), public health, and vaccination. Using the Twitter API (v2) and a pre-trained model (cardiffnlp/twitter-xlm-roberta-base-sentiment), tweets in Portuguese were collected, processed, and classified into positive, negative, and neutral sentiments. Despite API limitations, such as restricted data collection and the need for a larger training dataset, preliminary results reveal significant trends in public opinion. The work emphasizes that with a more extensive dataset, the analysis could achieve higher accuracy and robustness, serving as a foundation for future research.*

Resumo. *Este trabalho aborda a mineração de informações na web e redes sociais, com foco na análise de sentimentos relacionados ao Sistema Único de Saúde (SUS), saúde pública e vacinação no Brasil. Utilizando a API do Twitter (v2) e um modelo pré-treinado (cardiffnlp/twitter-xlm-roberta-base-sentiment), foram coletados e analisados tweets em português, seguidos de pré-processamento e classificação de sentimentos (positivo, negativo e neutro). Apesar das limitações impostas pela API, como restrições na quantidade de dados coletados e a necessidade de um conjunto maior para treinamento do modelo, os resultados preliminares indicam tendências relevantes na opinião pública. O trabalho destaca que, com uma base de dados mais ampla, a análise poderia alcançar maior precisão e robustez, sendo um ponto de partida para futuras investigações.*

1. Introdução

A análise de sentimentos em redes sociais tem se tornado uma ferramenta valiosa para compreender a opinião pública sobre temas relevantes, como saúde pública e políticas governamentais. Este trabalho visa investigar as percepções dos usuários do Twitter em relação ao SUS, saúde pública e vacinação no Brasil, utilizando técnicas de mineração de dados e processamento de linguagem natural.

Apesar do potencial da abordagem, enfrentamos desafios significativos, como as limitações impostas pela API do Twitter, que restringem o volume de dados coletados. Além disso, a eficácia do modelo de análise de sentimentos depende de um conjunto de dados amplo e diversificado para treinamento, o que não foi totalmente alcançado neste estudo. Essas limitações impactaram o desempenho do modelo, mas não invalidam a metodologia proposta. Com uma base de dados mais extensa, os resultados poderiam ser mais precisos e representativos, destacando a importância de futuros trabalhos nessa direção.

2. Ferramental Técnico

Credenciais e Acesso à API

Para a integração com a plataforma X, foram obtidas as seguintes credenciais de acesso:

- **API Key**
- **API Key Secret**
- **Bearer Token** (necessário para requisições na API v2)
- **Access Token and Secret**

O processo de obtenção consistiu em:

1. Acesso ao portal de desenvolvimento oficial da plataforma X
2. Criação de uma conta de desenvolvedor vinculada à organização
3. Registro de uma aplicação (”X App”) e extração das chaves de autenticação
4. Geração das credenciais de acesso

Ambiente de Desenvolvimento

- **Linguagem**: Python 3.11 (devido à sua ampla adoção em projetos de ciência de dados e suporte a bibliotecas de NLP)
- **Plataforma**: Google Colab (ambiente baseado em nuvem com suporte à GPU)
  - Facilita o uso de modelos de linguagem grandes (LLMs)
  - Oferece recursos computacionais robustos gratuitamente

2.1. Principais Bibliotecas Utilizadas

Acesso e Coleta de Dados

- **Tweepy**: Biblioteca para acesso à API v2 da plataforma X (Twitter)
  - Utilizada a classe `Client` para buscas por tweets com palavras-chave específicas
- **requests**: Para requisições HTTP à API do Twitter
- **json**: Manipulação de dados no formato JSON

Processamento de Dados

- **Pandas**: Manipulação, estruturação e limpeza dos dados coletados
- **re**: Expressões regulares para pré-processamento textual
- **nltk**: Processamento de linguagem natural (incluindo remoção de stopwords)

Análise de Sentimentos

- **Transformers** e **torch** (PyTorch):
  - Carregamento e inferência com o modelo pré-treinado `cardiffnlp/twitter-xlm-roberta-base-sentiment`

Visualização de Dados

- **Matplotlib** e **Seaborn**: Geração de gráficos (distribuição de sentimentos, evolução temporal)
- **WordCloud**: Criação de nuvens de palavras

Utilitários

- **os**: Interação com o sistema operacional
- **time**: Manipulação de tempo e pausas na execução

3. Contextualização e Análise dos Dados

Coleta de Dados

A coleta de dados foi realizada através da API v2 do Twitter, seguindo os seguintes parâmetros e procedimentos:

- **Endpoint utilizado**: `/tweets/search/recent`
- **Termos de busca**: ”SUS”, ”saúde pública” e ”vacinação”
- **Filtros aplicados**:
  - Idioma: português (`lang:pt`)
  - Geolocalização: raio de 2000 km a partir das coordenadas -14.2350 (latitude) e -51.9253 (longitude)
- **Parâmetros de paginação**:
  - Limite de 100 tweets por requisição
  - Máximo de 3 requisições consecutivas
  - Utilização do parâmetro `next_token` para navegação entre páginas
- **Controle de taxa de requisições**:
  - Tratamento do erro 429 (rate limit)
  - Pausas automáticas baseadas no cabeçalho `reset` da resposta
- **Armazenamento**: Dados salvos em arquivo JSON (`tweets_saude_v2.json`) contendo:
  - Texto dos tweets
  - Metadados (data de criação, métricas de engajamento)
  - Informações de localização quando disponíveis

Pré-processamento

Foi realizado uma limpeza textual com as seguintes etapas:

1. Remoção de:
   - URLs
   - Menções (@usuário)
   - Hashtags
   - Emojis
   - Pontuação
2. Normalização:
   - Conversão para minúsculas
   - Remoção de stopwords usando `nltk`

Análise Exploratória

Dados Coletados

Foram analisados **99 registros** de tweets, contendo ao todo **189 frases**, com as seguintes métricas principais:

Table 1. Métricas Básicas dos Tweets

Métrica	Valor
Quantidade de registros	99
Quantidade de frases	189
Tamanho médio das frases	118.79 caracteres
Média de retweets	468.5
Média de replies	0.1

Padrões de Engajamento

Observa-se uma disparidade significativa entre:

- Alto volume de **retweets** (468.5 por tweet)
- Baixíssima quantidade de **replies** (0.1 por tweet)

Isto sugere:

- Conteúdo **polarizante** que gera compartilhamento mas não diálogo
- Possível presença de contas automatizadas (*bots*)
- Baixa qualidade de interação humana genuína

Características do Texto

- Tamanho médio de **118.79 caracteres** indica:
  - Textos relativamente longos para padrões do Twitter
  - Possível inclusão de links ou mídias
- Razão de **1.9 frases por tweet** sugere:
  - Estrutura fragmentada de comunicação
  - Padrão típico de discussões acaloradas

Tarefa de mineração (Classificação de sentimentos):

A análise de sentimentos foi conduzida mediante a utilização de um modelo pré-treinado amplamente reconhecido no âmbito da inteligência artificial: o *cardiffnlp/twitter-xlm-roberta-base-sentiment*, integrante da renomada biblioteca HuggingFace Transformers, referência global no desenvolvimento de arquiteturas de linguagem natural.

Este modelo, treinado de forma supervisionada com base em pequeno corpus de publicações oriundas do Twitter, devido a limitação de requisições possíveis, apresenta elevada acurácia na categorização de enunciados em três classes distintas de polaridade emocional: positiva, negativa e neutra.

Procedeu-se à implementação do modelo por meio de inferência disponibilizado pela biblioteca Transformers, com suporte otimizado para execução em GPU — recurso este que, quando disponível, proporciona significativa melhoria no desempenho computacional.

Ressalte-se que essa abordagem se revela especialmente relevante mapeamento da opinião pública a respeito do cenário atual em que o pais se encontra.

3.1. Visualização dos dados:

Os resultados oriundos da análise de sentimentos foram representados por meio de recursos gráficos consagrados na literatura de análise de dados, permitindo não apenas a visualização intuitiva, mas também a extração de inferências relevantes com base empírica.

Inicialmente, procedeu-se à elaboração de gráficos de barras demonstrando a distribuição proporcional das polaridades emocionais identificadas nos tweets coletados. Em seguida, foi gerada uma nuvem de palavras com base na frequência léxica dos termos mais recorrentes, proporcionando uma visão panorâmica das expressões mais representativas nos discursos analisados.

Essas visualizações revelam tendências significativas no imaginário social em relação a temas de alta relevância para a saúde pública, como o Sistema Único de Saúde (SUS), os programas de vacinação e a temática sensível da saúde mental — informações que podem subsidiar desde políticas públicas até ações judiciais envolvendo o dever estatal de prestação adequada de serviços essenciais.

.

4. Conclusão

O trabalho proporcionou uma experiência valiosa no uso de técnicas de mineração de dados e análise de sentimentos em redes sociais, destacando tanto os pontos positivos quanto os desafios encontrados. A utilização da API do Twitter e do modelo cardiffnlp/twitter-xlm-roberta-base-sentiment permitiu a obtenção de insights preliminares sobre a opinião pública em relação ao SUS e temas correlatos.

No entanto, as limitações da API, como a restrição no número de requisições e a consequente escassez de dados, impactaram o desempenho do modelo. Ficou evidente que um conjunto de dados maior seria essencial para melhorar a precisão e a confiabilidade dos resultados. Apesar disso, a metodologia mostrou-se promissora, e com ajustes futuros, como a coleta de mais dados ou o uso de APIs alternativas, o trabalho poderia alcançar seu potencial pleno. Em suma, este estudo serve como um passo inicial para análises mais robustas e abrangentes no futuro.

References

[1] Pandas Development Team. *Pandas Documentation*. 2025. Disponível em: https://pandas.pydata.org. Acesso em: 27 mar. 2025.

[2] Python Software Foundation. *Python Standard Library - collections*. 2025. Disponível em: https://docs.python.org/3/library/collections.html. Acesso em: 27 mar. 2025.

[3] Matplotlib Development Team. *Matplotlib Documentation - pyplot API*. 2025. Disponível em: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html. Acesso em: 27 mar. 2025.

[4] BARBIERI, F.; ANKE, L. E.; CAMACHO-COLLADOS, J. **XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond**. In: PROCEEDINGS OF THE 13TH LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 2022, Marseille. **Anais...** Marseille: European Language Resources Association, 2022. p. 258-266. Disponível em: https://aclanthology.org/2022.lrec-1.27. Acesso em: 01 mai. 2025.

[5] BARBOSA, L.; ALVES, D.; RIBEIRO, F. **Twitter-XLM-RoBERTa-base for Sentiment Analysis**. Hugging Face, 2022. Disponível em: https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment. Acesso em: 01 mai. 2025.

[6] OpenAI. *ChatGPT*. 2025. Disponível em: https://chatgpt.com. Acesso em: 27 mar. 2025.