Trie of Rules para Visualização de Textos de Redes Sociais

Luiza Bolonha Vieira Onofre, Patrick Marques Ciarelli

[1]Departamento de Engenharia Elétrica – Universidade Federal do Espírito Santo (UFES)
Caixa Postal 29075-910 – Vitória – ES – Brasil

luiza.bolonha@outlook.com, patrick.ciarelli@ufes.br

Abstract. *Data visualization is essential for interpreting large volumes of data, especially texts from social media. This work proposes the use of the Trie of Rules as an efficient and interpretable solution for analyzing and visualizing short texts. The methodology includes text preprocessing and hierarchical organization of relevant terms, allowing for clear identification of patterns and relationships. Functionalities were proposed to improve the quality of data analysis, such as the use of a Large Language Model (LLM) to summarize the content of selected social media posts. The results indicate a clean, flexible, and effective visualization that seeks to overcome the limitations of common approaches.*

Resumo. *A visualização de dados é essencial para interpretar grandes volumes de informação, especialmente textos de redes sociais. Este trabalho propõe o uso do método Trie of Rules como uma solução eficiente e interpretável para análise e visualização de textos curtos. A metodologia inclui pré-processamento textual e organização hierárquica dos termos relevantes, permitindo identificar padrões e relações de forma clara. Também foram propostas funcionalidades para melhorar a qualidade da análise dos dados, como o uso de um Large Language Model (LLM) para resumir o conteúdo de postagens selecionadas de redes sociais. Os resultados esperados indicam uma visualização limpa, flexível e eficaz, que busca superar as limitações das abordagens comuns.*

1. Introdução

A era digital atual é caracterizada pela hiperconectividade e pela produção acelerada de dados. Em 2025, mais de 5,2 bilhões de pessoas estavam ativas nas redes sociais (cerca de 64% da população mundial) [Insights 2025]. No Brasil, os usuários passam, em média, 3 horas e 49 minutos por dia conectados, colocando o país no topo do ranking global de tempo médio diário [Statista 2025]. A cada minuto, em todo o mundo, são compartilhadas cerca de 41,6 milhões de mensagens no WhatsApp, 241 milhões de e-mails e mais de 360 mil postagens no X (antigo Twitter) [Domo 2023], o que evidencia o grande volume e a complexidade dos dados textuais gerados.

Com isso, cresce a necessidade de ferramentas eficazes de visualização de dados, capazes de traduzir rapidamente grandes volumes de dados em conhecimento útil [da Silva 2019]. Contudo, a capacidade analítica não avança no mesmo ritmo da geração de dados [Keim et al. 2008], tornando essenciais abordagens visuais que facilitem a exploração e interpretação de informações complexas.

Neste contexto, este trabalho propõe o uso do método *Trie of Rules* [Kudriavtsev et al. 2024] para analisar textos curtos de forma estruturada e intuitiva. Ori-

ginalmente, este método foi projetado para a visualização de regras de associação, permitindo uma análise rápida e intuitiva das relações entre os elementos, por meio de conexões hierárquicas, além de carregar algumas informações extras. Este método mostrou-se mais efetivo do que outros métodos conhecidos de visualização [Kudriavtsev et al. 2024]. Neste trabalho, são propostos alguns ajustes, como uma etapa de pré-processamento dos textos, para utilizar o *Trie of Rules* na visualização de textos curtos, além de algumas funcionalidades para melhorar a experiência, como o uso do ChatGPT para sumarização do conteúdo. Para avaliar o método proposto, foi utilizado um conjunto de postagens da rede social X e comparado com alguns métodos clássicos de visualização de dados.

2. Fundamentação Teórica

Avaliar a eficácia de métodos de visualização é uma tarefa complexa, pois envolve diferentes tipos de dados, contextos de uso e interações dos usuários. Para guiar o desenvolvimento de interfaces mais eficazes, [Shneiderman et al. 1996] propuseram o mantra da visualização: "visão geral primeiro, depois amplie e filtre, e então detalhes sob demanda". Esse mantra orienta a criação de ferramentas que permitam uma exploração progressiva e intuitiva dos dados, facilitando a compreensão de grandes volumes de informação sem perder o contexto. Complementando essa abordagem, os autores também desenvolveram o modelo TTT, que relaciona sete tipos de dados (unidimensionais, bidimensionais, tridimensionais, temporais, multidimensionais, árvores e redes) às tarefas de interação mais adequadas para sua exploração.

As sete tarefas consideradas desejáveis no modelo TTT são: (1) **visão geral**, que oferece uma panorâmica do conjunto, permitindo compreender o contexto antes de explorar detalhes; (2) **zoom**, para ampliar áreas de interesse e examinar informações com mais precisão; (3) **filtragem**, que remove dados irrelevantes e reduz a complexidade visual; (4) **detalhes sob demanda**, exibindo informações adicionais apenas quando solicitado; (5) **visualização de relações**, destacando conexões e padrões entre elementos; (6) **histórico**, que permite desfazer ou refinar ações anteriores; e (7) **extração**, possibilitando salvar subconjuntos de dados para análises futuras. Essas tarefas fornecem um conjunto de critérios úteis para avaliar e projetar métodos de visualização eficazes, ajustados à natureza dos dados e às necessidades dos usuários.

2.1. Comparação entre Métodos Clássicos de Visualização

A **Nuvem de Palavras** é uma técnica visual simples e atrativa que destaca as palavras mais frequentes de um texto por meio do tamanho da fonte, oferecendo uma visão inicial do conteúdo [Heimerl et al. 2014]. Seu apelo estético e a facilidade de uso a tornam um método popular para apresentações e análises exploratórias rápidas. No entanto, essa abordagem ignora o contexto de uso das palavras, as relações semânticas e a sequência textual, dificultando análises mais profundas.

A **Árvore de Palavras** proporciona uma visualização mais contextualizada ao organizar as ocorrências de uma palavra-chave como a raiz de uma árvore, cujos ramos mostram palavras subsequentes, permitindo explorar padrões e significados com base no contexto [Wattenberg and Viégas 2008]. A interatividade e a estrutura hierárquica facilitam a análise semântica, no entanto, a técnica depende da escolha de uma palavra central, o que pode limitar a exploração do conteúdo e introduzir viés. Em textos com muita variação contextual, a árvore tende a se expandir excessivamente e perder clareza visual.

Tabela 1. Comparação de métodos com base no modelo TTT. Adaptado de [Gan et al. 2014].

Características	Nuvem de Palavras	Grafo de Coocorrência	Árvore de Palavras
Visão geral	x	x	x
Zoom		x	x
Filtragem	x	x	x
Detalhes			x
Relações		x	x
Histórico	x	x	x
Extração		x	x

Por último, o **Grafo de Coocorrência**, de modo geral, representa relações de proximidade entre palavras em um corpus. Segundo [Chen et al. 2016], essa abordagem baseia-se no princípio de que "quanto maior a frequência de coocorrência entre dois termos, mais próxima é a relação temática entre eles", sendo útil para mapear tendências e estruturas de conhecimento em dados textuais. Embora ofereça *insights* valiosos sobre as estruturas discursivas e relações semânticas, o grafo pode se tornar complexo em corpora extensos, exigindo técnicas adicionais de filtragem e agrupamento para manter a legibilidade e a eficácia da visualização.

A análise dos métodos populares de visualização de textos evidencia que cada um adota abordagens e objetivos distintos. Para avaliar sua eficácia de forma sistemática, a Tabela 1 apresenta uma comparação baseada nas sete características do modelo TTT de [Shneiderman et al. 1996]. Embora todos os métodos ofereçam funcionalidades relevantes, apenas a Árvore de Palavras atende a todos os critérios analisados. Entretanto, sua visualização extensa pode comprometer a interpretação rápida das informações.

Essa análise evidencia a ausência, entre os métodos clássicos, de uma solução que combine as funcionalidades do modelo TTT com uma apresentação visual que facilite a compreensão e a extração de informações de forma rápida e intuitiva. Esse cenário aponta para a necessidade do desenvolvimento de novas abordagens.

2.2. Regras de Associação

As regras de associação descrevem relações entre itens com base em sua coocorrência em transações [Brusso 2000]. Elas são expressas na forma $X \Rightarrow Y$, que indica que a ocorrência de X implica em uma provável ocorrência de Y [Agrawal et al. 1993]. Suas aplicações incluem análise de dados, classificação, *marketing* cruzado, agrupamento, entre outros [Kumbhare and Chobe 2014]. Seus principais indicadores são as métricas de **suporte** e **confiança**, embora outras métricas, como ***lift***, possam ser utilizadas.

O **suporte** representa a frequência com que os itens da regra aparecem juntos no conjunto de dados [Mining 2006], Equação 1, e a **confiança** expressa a probabilidade condicional de encontrar Y dado que X já ocorreu [Mining 2006], Equação 2.

$$\text{Suporte}(X \Rightarrow Y) = \frac{\text{Número de transações com ambos } X \text{ e } Y}{\text{Número total de transações}}. \quad (1)$$

$$\text{Confiança}(X \Rightarrow Y) = \frac{\text{Número de transações com ambos } X \text{ e } Y}{\text{Número de transações com } X}. \quad (2)$$

Esses indicadores ajudam a avaliar a relevância e a força das associações entre os itens. No entanto, o processo de identificar essas associações em grandes volumes de dados exige uma abordagem sistemática e eficiente. É nesse contexto que surge a Mineração de Regras de Associação (*Association Rule Mining* - ARM), que visa descobrir automaticamente tais relações significativas em conjuntos de dados extensos. ARM é uma técnica fundamental no campo da Mineração de Dados, na qual algoritmos específicos, como o Apriori e o *FP-Growth* (*Frequent Pattern Growth*), são utilizados para gerar e avaliar regras de associação de forma eficiente [Karthikeyan and Ravikumar 2014].

O Apriori [Agrawal et al. 1993] é um dos algoritmos mais conhecidos para ARM. Ele gera conjuntos candidatos iterativamente e os filtra pelo suporte mínimo. Apesar de simples, é custoso para grandes volumes de dados. O *FP-Growth* [Han et al. 2000] supera essa limitação com a *FP-Tree*, uma estrutura compacta que reduz redundâncias ao condensar transações semelhantes. Com ela, padrões são extraídos por meio de mineração condicional, evitando múltiplas varreduras e a geração explícita de candidatos. O *FP-Max* [Grahne and Zhu 2003], por sua vez, foca apenas nos itemsets máximos frequentes, eliminando subconjuntos redundantes. Ele aplica estratégias de poda para evitar a expansão de ramos irrelevantes, oferecendo eficiência superior em bases densas.

2.3. Método *Trie of Rules*

O método *Trie of Rules*, proposto por [Kudriavtsev et al. 2024], é voltado para a visualização eficiente do resultado de ARM na forma de grafos. Ele estende a estrutura da *FP-tree* ao organizar as regras como caminhos em um grafo hierárquico, onde os nós representam itens e as arestas indicam suas relações. Cada trajeto da raiz até um nó terminal corresponde a uma regra, com métricas, como suporte e confiança, incorporadas aos atributos visuais dos nós, o que facilita a identificação de padrões e a análise das relações entre os itens. A visualização pode ser personalizada, por exemplo, com o tamanho dos nós refletindo a confiança e a tonalidade da cor indicando o *lift*. A Figura 1a) apresenta uma aplicação do método nos dados "*Online Retail Logs*" [Chen 2015].

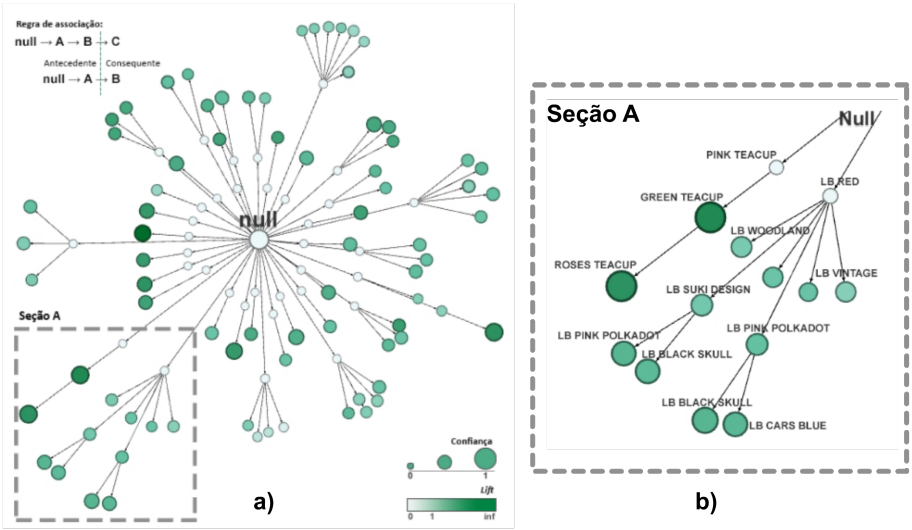

Figura 1. Visualização do *Trie of Rules*. Adaptado de [Kudriavtsev et al. 2024].

A Figura 1b) apresenta uma ampliação da Seção A da Figura 1a), cujos padrões de visualização revelam agrupamentos e correlações significativas, como a associação

entre modelos de lancheiras (do inglês, *lunch bag*) ou a combinação preferida de xícaras (do inglês, *teacup*). Isso evidencia o potencial do método tanto para análise de regras de associação quanto para aplicações como análise de textos, na qual palavras podem assumir o papel de itens e suas relações semânticas serem visualizadas como regras.

3. Metodologia

A proposta deste trabalho é aplicar técnicas de pré-processamento nos textos para extrair as palavras mais relevantes, empregar técnicas de ARM e, em seguida, utilizar o método do *Trie of Rules* para estruturar visualmente as relações entre esses termos.

A Figura 2 apresenta um fluxograma com as principais etapas do processo. O fluxo tem início no pré-processamento dos dados, seguido pela aplicação do método *Trie of Rules*, que gera um arquivo como resultado. Esse arquivo pode ser visualizado em softwares como o Gephi e/ou analisado pela funcionalidade de sumarização de agrupamentos de postagens, proposta neste trabalho.

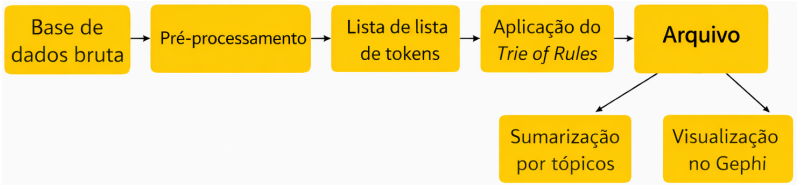

Figura 2. Fluxograma geral do método proposto.

3.1. Conjunto de Dados

Para realizar a avaliação do método proposto, foi utilizado um conjunto de postagens da rede social X (antigo Twitter) coletadas sobre o tema vacinação no período de 01/05/2025 a 07/06/2025, com 2701 e 769 postagens para os meses de maio e junho, respectivamente. As coletas foram feitas com a ferramenta web *Export Comments* [Export Comments 2026]. A coleta foi realizada utilizando vários termos relacionados ao tema vacinação, como vacina e vacinação, e diferentes nomes de doenças, como sarampo, catapora, meningite, coronavírus, entre outros, além do uso de operações lógicas entre os termos, como AND e OR. Compõem esse arquivo informações como o texto da postagem, métricas de engajamento, dados relacionados ao autor e a data de criação.

3.2. Pré-processamento dos Dados

O pré-processamento dos dados é a etapa inicial do método proposto e se faz fundamental para a extração de conhecimento, seleção, organização e preparação dos dados para as fases posteriores [Goldschmidt and Passos 2005]. Esta etapa garante que os textos estejam devidamente preparados para a aplicação do *Trie of Rules* a textos curtos.

Os textos usados neste trabalho foram de postagens na rede social X e, portanto, foi necessário um pré-processamento rigoroso, dada a informalidade, erros ortográficos e gramaticais, o uso de gírias e a alta variabilidade linguística desse tipo de conteúdo.

Inicialmente, foi feita a remoção de postagens duplicadas com base em similaridade semântica e um limiar de similaridade cosseno de 0,92, utilizando a biblioteca spaCy

[Honnibal and Montani 2026]. Esta etapa foi relevante para evitar que postagens de pessoas e *bots*, que inundam as redes sociais, polarizassem as análises. Em seguida, aplicou-se uma etapa de limpeza textual que envolveu a remoção de URLs, menções, *hashtags*, *emojis* e a substituição de abreviações comuns de palavras por suas formas completas.

A seguir, foi utilizada a seguinte sequência de técnicas em cada postagem: tokenização [Schütze et al. 2008], que segmenta o texto em unidades analisáveis, implementada com o uso da biblioteca spaCy; remoção de *stopwords*, palavras com pouca carga semântica, cuja exclusão contribui para a redução de ruído e melhora a precisão dos algoritmos, feita a partir da lista de *stopwords* em português da biblioteca NLTK [Bird et al. 2009]; reconhecimento de termos compostos e de entidades nomeadas (*Named Entity Recognition* - NER), também com o uso da ferramenta spaCy, cuja meta é detectar e classificar automaticamente nomes próprios (como pessoas, organizações ou locais), enriquecendo a representação textual com a extração de informações estruturadas [Nadeau and Sekine 2007]; análise morfossintática utilizando as *POS tags* do spaCy, que identifica a classe gramatical de cada termo, permitindo filtrar palavras menos informativas [Jurafsky and Martin 2008], mantendo apenas substantivos, nomes próprios e adjetivos. Após estes passos, os termos (*tokens*) duplicados que sobraram em cada postagem foram eliminados, a fim de evitar distorções na geração das regras de associação.

As etapas anteriores garantem que apenas os termos mais informativos fossem considerados, contribuindo para uma visualização mais concisa, interpretável e eficaz, especialmente na análise de textos curtos como postagens em redes sociais.

Após o pré-processamento, cada postagem foi representada como uma lista de *tokens*, resultando em um conjunto de dados estruturado como uma lista de listas. Essa estrutura serviu como entrada para o método *Trie of Rules*, na qual cada lista de *tokens* foi considerada como uma transação.

3.3. Aplicação do *Trie of Rules*

O código-fonte do *Trie of Rules* foi obtido do repositório oficial no GitHub dos autores do método [ARM-interpretation 2021][1] . Para a execução do método, foi necessário definir um algoritmo para ARM e um valor mínimo de suporte. Nos experimentos, foi observado que o algoritmo Apriori se mostrou o mais adequado ao contexto de postagens em redes sociais, por permitir uma análise mais descritiva, mesmo que mantenha certas duplicidades. O resultado da execução do código do *Trie of Rules* foi um objeto da classe `Trieofrules`, que pode ser exportado em formato `.graphml`, contendo a estrutura do grafo e métricas de regras de associação, como suporte e confiança.

3.4. Visualização dos Grafos

Para a visualização dos grafos, utilizou-se a ferramenta Gephi [Bastian, M. et al. 2009][2]. Algumas configurações foram necessárias para garantir a visualização adequada. Na aba "Visão Geral" do software, recomenda-se aplicar, em sequência, os algoritmos de distribuição `ForceAtlas2` e `Yifan Hu`. Na seção "Aparência", o tamanho dos nós pode ser configurado por *ranking*, com base no valor de suporte, enquanto a cor dos nós pode ser

[1]Disponível em: `https://github.com/ARM-interpretation/Trie-of-rules`. Acesso em: 16 maio 2026.

[2]Disponível em: `https://gephi.org`. Acesso em: 16 maio 2026.

atribuída conforme os valores de confiança, facilitando a identificação visual das regras mais relevantes. Em "Rótulos", é necessário verificar se a coluna `value` está selecionada como rótulo principal dos nós, garantindo a correta exibição dos termos. Já na aba "Visualização", recomenda-se desativar as arestas curvas e ativar os rótulos dos nós, ajustando o tamanho da fonte conforme necessário. Após essas configurações, a visualização pode ser atualizada e exportada em formatos como `.pdf` ou `.png`.

3.5. Outras Funcionalidades para Visualização e Análise dos Dados

Além da proposta descrita para visualização de dados textuais utilizando o *Trie of Rules*, neste trabalho também foram desenvolvidas algumas adaptações e procedimentos para melhorar a visualização e a análise dos dados.

Para facilitar a visualização dos resultados gerados pelo método *Trie of Rules*, foi desenvolvido um *script* que permite identificar e separar *clusters* na visualização. Neste trabalho, um *cluster* é definido como sendo um ou mais ramos que partem de um mesmo nó e que, por sua vez, está ligado ao nó raiz. A partir do arquivo `.graphml` original, são gerados dois novos arquivos: um contendo apenas os ramos formados por um único nó ligado à raiz (definido como *cluster* central, representando termos que ocorrem isoladamente) e outro contendo os ramos compostos por múltiplos nós (*clusters* periféricos). Este desacoplamento dos nós do grafo permite uma visualização mais limpa.

Outra funcionalidade proposta foi a de sumarização de postagens por tópicos. Com base nos termos presentes em um *cluster*, são localizadas as postagens que contêm todos os *tokens* de ao menos uma das ramificações que compõem o *cluster* estudado. Essas postagens são armazenadas em arquivos `.csv` e, a partir dos termos presentes na estrutura, é gerado um *prompt* de comando que é utilizado em uma ferramenta de *Large Language Model* (como o ChatGPT-4[3]) para realizar a sumarização do conteúdo. Esse processo permite a segmentação temática das postagens e a geração de resumos claros e objetivos sobre os assuntos discutidos. A seguir, tem-se um exemplo de *prompt* adotado, onde a palavra "dengue" é o nó antecedente das palavras"ano", "Butantan" e "Anvisa":

> Considere os tweets a seguir, que foram identificados como relacionados ao assunto **dengue**. Eles mencionam com frequência palavras como: **ano, Butantan, Anvisa**. Com base nesse conteúdo, gere um pequeno texto de 1 a 2 parágrafos, com uma síntese clara e objetiva sobre o que está sendo discutido nos tweets. O texto deve ajudar qualquer pessoa a entender rapidamente o teor das conversas, mesmo que ela não tenha lido os tweets originais. Retorne somente o resumo.

Dessa forma, o processo desenvolvido possibilita a segmentação das publicações em grupos temáticos coesos e a geração de resumos objetivos para cada um deles.

3.6. Gerando Visualizações com Outras Abordagens

Para comparar o método proposto com visualizações já consolidadas, foi necessário preparar os conjuntos de dados de forma apropriada, de forma a possibilitar a geração dos resultados com as visualizações descritas na Seção 2.1.

Para a Nuvem de Palavras e Grafo de Coocorrência, foi preciso um pré-processamento simplificado ao conjunto de dados em relação ao utilizado no método proposto. Nele, foi removida a identificação de entidades e termos compostos, mas manteve-se os demais passos. A Nuvem de Palavras foi construída a partir das 1000 palavras mais

[3]Disponível em: `https://chatgpt.com`. Acesso em: 16 maio 2026.

frequentes (dentre as pré-processadas), usando o site `wordcloud.online`. A Árvore de Palavras utiliza o conteúdo textual completo da base no site `jasondavies.com/wordtree`, permitindo explorar os contextos de uso dos termos.

Para o Grafo de Coocorrência, foram selecionadas as 100 palavras pré-processadas mais recorrentes, considerando como arestas suas coocorrências com todas as palavras dentro das transações. Os dados do grafo foram importados no Gephi. Na aba de Visão Geral, foram aplicados os cálculos de Grau Ponderado Médio e Modularidade; os nós foram coloridos por classe de modularidade e redimensionados de acordo com o grau de entrada ponderado. A distribuição adotada foi a *Circle Pack Layout*, seguida de ajustes de expansão e exibição dos rótulos. Assim, os nós do grafo são visualizados na forma de comunidades (*clusters*) de palavras.

4. Resultados

Com o objetivo de avaliar a eficácia do método proposto, uma série de visualizações foi realizada. Todo o código deste estudo foi desenvolvido em Python[4].

4.1. Comparação entre os Métodos

A Figura 3 apresenta o *Trie of Rules* gerado a partir das postagens realizadas em maio de 2025. Ela foi obtida com as regras de associação extraídas via algoritmo Apriori, utilizando suporte mínimo de 0,009. Nesta visualização, o tamanho do nó representa o suporte, isto é, a frequência relativa do termo no *corpus*, e a cor do nó codifica a confiança, métrica que expressa a probabilidade de um termo ocorrer dado que outro já apareceu. Assim, um nó grande indica um termo encontrado várias vezes no texto e uma coloração verde mais intensa sinaliza associações mais confiáveis entre os termos conectados. O nó raiz (nó "0") serve de referência para o maior valor de suporte e de confiança possível.

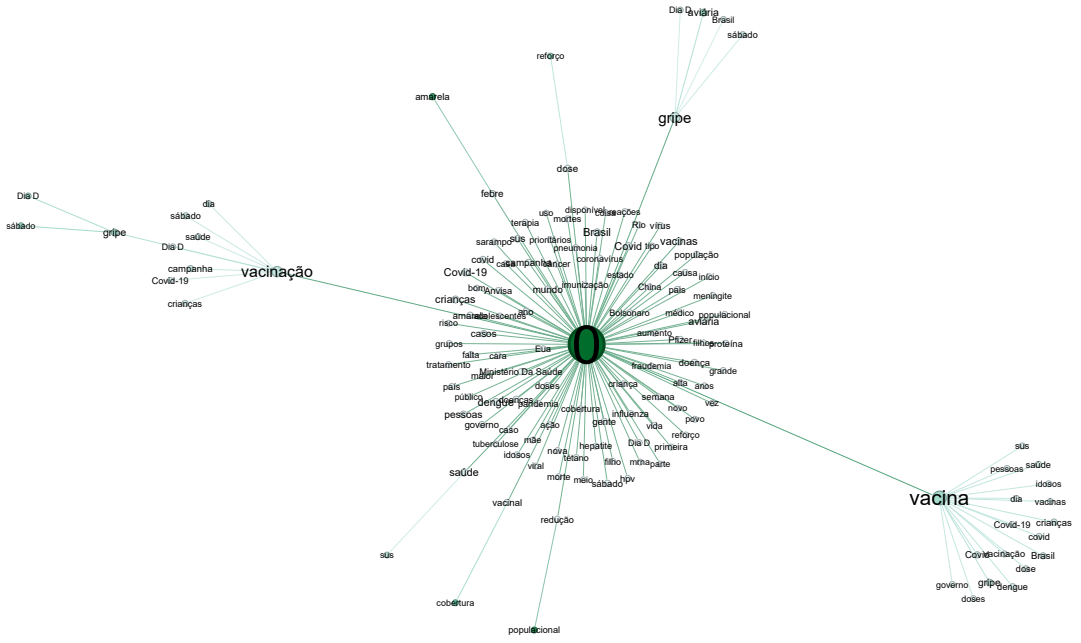

Figura 3. *Trie of Rules* das postagens do mês de maio de 2025 sobre vacina.

[4]Disponível em: https://github.com/luiza-bolonha/Pre-Processing-ToR. Acesso em: 18 maio 2026.

O valor mínimo de suporte adotado foi definido por meio de testes com diferentes limiares. Observou-se que valores significativamente mais altos resultavam em poucas regras de associação, devido à diversidade temática presente nos dados utilizados.Por outro lado, valores menores produziam visualizações excessivamente densas, com grande quantidade de termos, dificultando a interpretação e compreensão dos resultados.

Observa-se um conjunto de termos ligados diretamente ao nó raiz, mas que não estão conectados a outros termos. Isto é um indicativo de que, embora estes termos sejam frequentes, em relação ao valor de suporte mínimo, não existe uma confiança forte o suficiente para identificar a ocorrência de outras palavras uma vez que estes termos tenham ocorrido. Por outro lado, existem ramificações mais profundas que conectam sequências de termos, revelando tópicos específicos presentes nas postagens.

Das ramificações presentes na Figura 3, é possível extrair uma visão panorâmica, como em "vacina" e "vacinação", evidenciando a centralidade desses assuntos na cobertura analisada e possibilitando extrair algumas relações interessantes. Na análise da palavra "vacina", por exemplo, observam-se várias conexões derivadas dela, como "gripe", "dengue" e "Covid-19", indicando um aumento da mobilização pública em torno de ações específicas para imunização de diversas doenças.

A visualização também revela conexões menores, porém significativas, como "febre amarela", "meningite", "hepatite" e "HPV" indicando atenção a doenças específicas que, embora não dominem o debate, permanecem presentes na cobertura. Palavras como "falta", "disponível" e "cobertura" sugerem discussões sobre logística e acesso, enquanto "Dia D" e "campanha" apontam para esforços de mobilização. Esses ramos ajudam a compor um quadro mais amplo das preocupações e ações relacionadas à imunização.

Para efeito comparativo, com os mesmos dados de maio de 2025, foram geradas as visualizações de Nuvem de Palavras, Árvore de Palavras e Grafo de Coocorrência, Figuras 4, 5 e 6, respectivamente. A Nuvem de Palavras permite resgatar rapidamente as principais temáticas, com destaque visual para termos como "Vacina", "Vacinação" e "Gripe", embora não revele como esses termos se associam entre si ou com os demais. Já a Árvore de Palavras fornece o contexto das discussões, permitindo identificar temas recorrentes, como vacinação contra gripe e covid-19, mas exige a definição de um termo central (no caso, "Vacinação") além de ser mais complexa para obter uma visão panorâmica dos assuntos. O Grafo de Coocorrência oferece uma visão relacional dos termos, evidenciando agrupamentos semânticos e conexões relevantes entre tópicos. Comunidades como "Vacina" e "Doses", ou ainda "Gripe" e "Covid", tornam-se visualmente perceptíveis. Apesar da riqueza interpretativa, a visualização pode se tornar densa e exigir ferramentas interativas para facilitar a leitura, especialmente em contextos com alto volume de dados.

Diante das diferentes abordagens exploradas, observa-se que cada método de visualização possui vantagens e limitações específicas. A estrutura *Trie of Rules* destaca-se por evidenciar hierarquias e associações frequentes entre termos, permitindo uma leitura panorâmica das principais construções textuais. No entanto, pode-se tornar uma visualização densa e de difícil leitura caso o suporte mínimo esteja subdimensionado. A Nuvem de Palavras oferece uma visão quantitativa e imediata dos termos mais recorrentes, embora careça de profundidade semântica. A Árvore de Palavras proporciona acesso direto

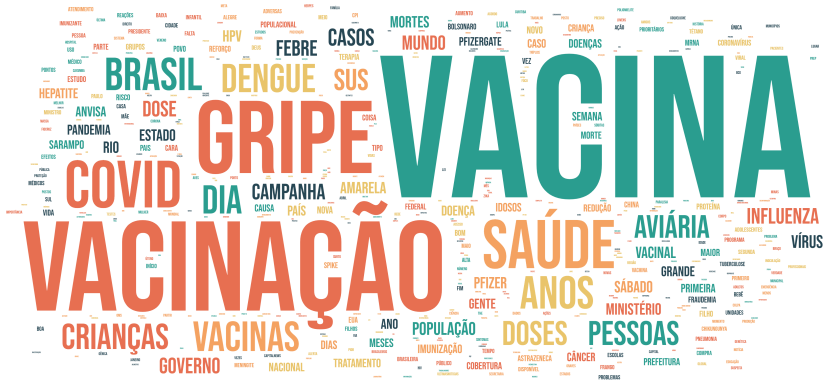

Figura 4. Nuvem de Palavras de postagens de maio de 2025 sobre vacina.

Figura 5. Árvore de Palavras de postagens de maio de 2025 sobre vacina.

aos contextos em que os termos são empregados, mas sua interface extensa e textual pode dificultar a identificação rápida de padrões. Já o Grafo de Coocorrência combina frequência e contextos semânticos, oferecendo uma visão relacional mais rica entre os tópicos, embora a densidade de informações possa comprometer a clareza visual.

4.2. Contribuições Adicionais para Método

A sumarização de publicações foi uma funcionalidade desenvolvida para enriquecer a análise do método proposto, de forma a contextualizar e aprofundar as informações dos temas presentes no *Trie of Rules* sem a necessidade da leitura de várias postagens. Como descrito na Seção 3.5, a partir da seleção de um *cluster*, são identificadas as postagens que contêm termos de ao menos uma de suas ramificações, as quais são então enviadas ao ChatGPT com um *prompt* estruturado para gerar um resumo do tema.

Outra funcionalidade proposta foi a separação do *Trie of Rules* em um *cluster* central, composto por nós isolados conectados apenas à raiz, e *clusters* periféricos, que abrangem as demais ramificações da visualização. Essa divisão busca reduzir a carga de informação exibida simultaneamente, facilitando a leitura da estrutura geral e permitindo uma análise mais direcionada dos temas emergentes em diferentes regiões do grafo. Enquanto os nós isolados geralmente representam tópicos amplos ou recorrentes, as ramificações indicam discussões mais estruturadas, com maior densidade semântica e potencial analítico para compreensão de contextos, argumentos e narrativas específicas.

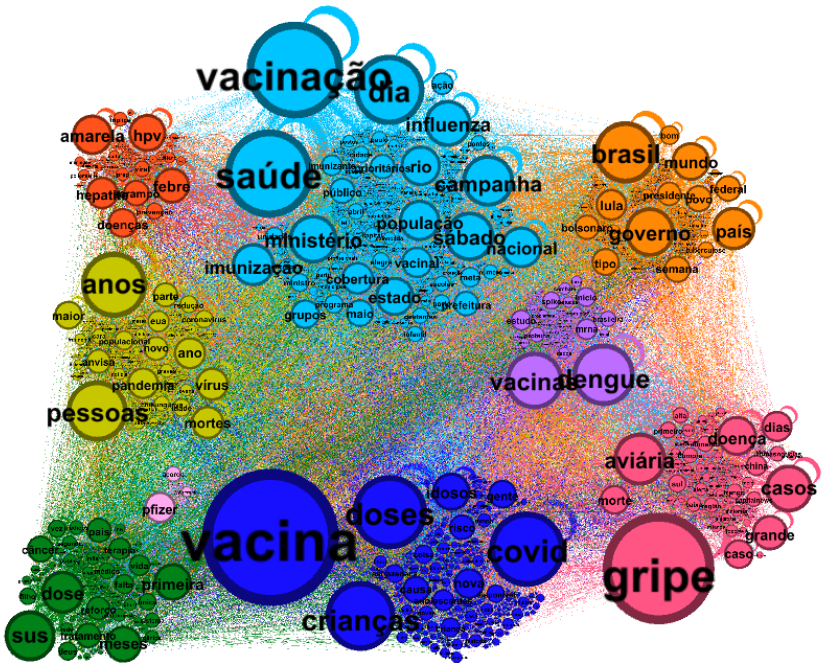

Figura 6. Grafo de Coocorrência de postagens de maio de 2025 sobre vacina.

Para avaliar essas funcionalidades, a Figura 7 apresenta os *clusters* periféricos extraídos de postagens da primeira semana de junho de 2025, utilizando o algoritmo Apriori e suporte mínimo de 0,008, determinado também empiricamente. Enquanto a Figura 8 exibe o *cluster* central, contendo os termos isolados. A escolha de um intervalo de tempo curto de dados favorece a identificação de temas pontuais, muitas vezes negligenciados em análises com um tempo mais longo. Na Figura 7, observam-se ramificações distantes da temática predominante de saúde, como os grupos de termos “fake” e “café”, ou “granja” e “comercial”.

Por se tratar de um tema que foge do padrão, é interessante descobrir o contexto que esses termos foram citados. Dessa forma, o *cluster* com o nó central “fake” foi sumarizado, com o seguinte resultado:

> Os tweets discutem a ação da Anvisa que determinou o recolhimento de três marcas de “café fake”, ou seja, cafés considerados impróprios para o consumo humano. A medida foi tomada após a identificação de impurezas e substâncias não permitidas nesses produtos, levantando preocupações sobre a qualidade e segurança alimentar. O termo “fake” tem sido usado para destacar a gravidade da adulteração desses alimentos.
> As conversas também refletem indignação e desconfiança dos consumidores em relação à fiscalização de produtos alimentícios no país. A repercussão nas redes sociais evidencia uma demanda por mais transparência e controle na cadeia produtiva de alimentos, com o “café fake” tornando-se símbolo dessa preocupação com fraudes no consumo cotidiano.

A estratégia de sumarização dos *clusters* demonstra ser relevante para a compreensão contextual de temas emergentes nas postagens, e o uso dos termos dos *clusters* ajuda a filtrar as postagens relacionadas ao tema. Como exemplificado com o caso “café fake”,

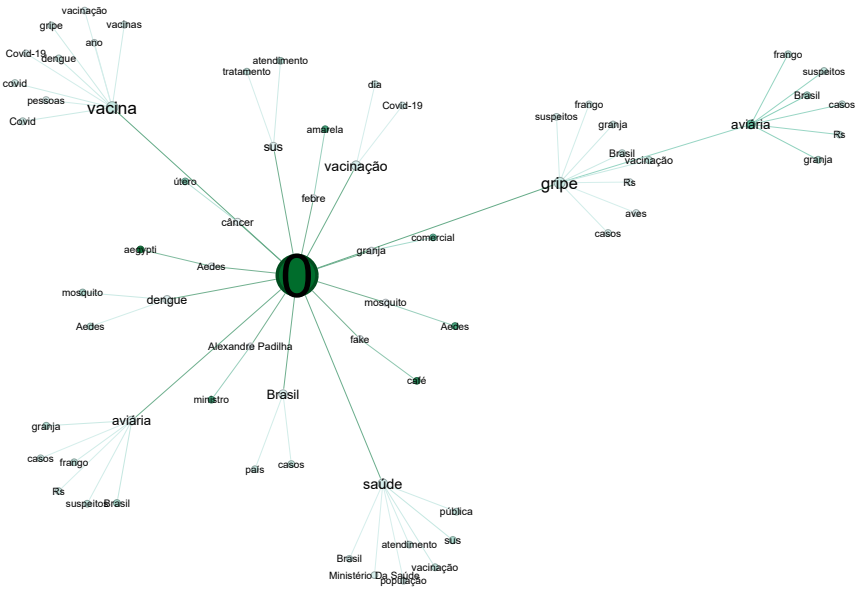

Figura 7. *Clusters* periféricos das postagens da primeira semana de junho.

essa abordagem permite não apenas sintetizar grandes volumes de dados, mas também capturar nuances e preocupações específicas manifestadas pelos usuários.

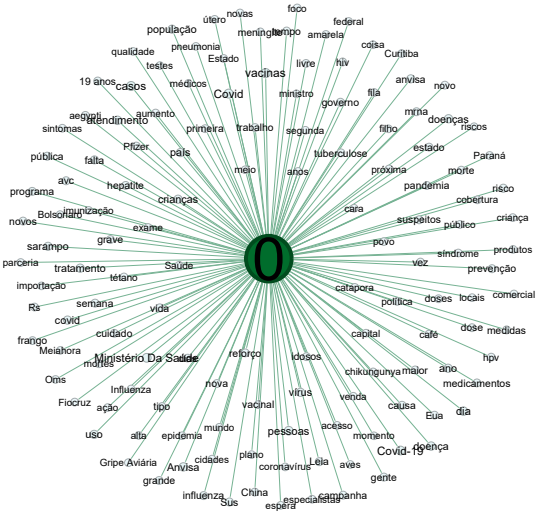

Figura 8. *Cluster* central das postagens da primeira semana de junho.

5. Conclusão

Este trabalho propôs um método de visualização para textos curtos baseado no *Trie of Rules*. Foram feitas adaptações específicas para o pré-processamento textual, além de funcionalidade de sumarização de tópicos e reconhecimento e separação dos *clusters*.

O método proposto demonstrou bom desempenho, pois ele permite ter uma visão geral dos dados, visualizar as relações hierárquicas entre os termos, explorar os dados com maior ou menor abrangência (em função do valor do suporte), extrair os *clusters* relevantes e obter, a partir deles, uma seleção de textos dos quais consegue obter maiores

detalhes por meio da sumarização. Além disso, uma mesma estrutura de grafo pode ser visualizada de diferentes formas, por meio da mudança do método de distribuição dos nós e/ou o uso de diferentes métricas para representar a cor e o tamanho dos nós.

Para trabalhos futuros, serão investigadas estratégias de automatização do suporte mínimo, mecanismos para detecção de variações temporais nos textos, a investigação do uso de radicais das palavras para agrupar termos com semânticas semelhantes, comparar o método proposto com outras abordagens de mineração de regras e o uso de outras métricas, como o *lift*, para enriquecer a representação visual.

Referências

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216.

ARM-interpretation (2021). Trie-of-rules: Visualization method for association rule mining (arm). Repositório no GitHub. Acessado em: 16 mar. 2026.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. `https://gephi.org`. Acessado em: 16 mar. 2026.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc. Acessado em: 16 mar. 2026.

Brusso, M. J. (2000). Access miner: uma proposta para a extração de regras de associação aplicada à mineração do uso da web. Master's thesis, Universidade Federal do Rio Grande do Sul.

Chen, D. (2015). Online retail. `https://doi.org/10.24432/C5BW33`. Acessado em: 16 mar. 2026.

Chen, X., Chen, J., Wu, D., Xie, Y., and Li, J. (2016). Mapping the research trends by co-word analysis based on keywords from funded project. *Procedia computer science*, 91:547–555.

da Silva, F. C. C. (2019). Visualização de dados: passado, presente e futuro. *Liinc em Revista*, 15(2):205–223.

Domo (2023). Data never sleeps 11.0. `https://www.domo.com/learn/infographic/data-never-sleeps-11`. Acessado em: 16 de mar. 2026.

Export Comments (2026). Export facebook, instagram, twitter, youtube, tiktok, vimeo comments to csv / excel. Acessado em: 16 mar. 2026.

Gan, Q., Zhu, M., Li, M., Liang, T., Cao, Y., and Zhou, B. (2014). Document visualization: an overview of current research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(1):19–36.

Goldschmidt, R. and Passos, E. (2005). *Data mining: um guia prático*. Gulf Professional Publishing.

Grahne, G. and Zhu, J. (2003). High performance mining of maximal frequent itemsets. *6th International workshop on high performance data mining*, 16:34.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12.

Heimerl, F., Lohmann, S., Lange, S., and Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *47th Hawaii International Conference on System Sciences*, pages 1833–1842.

Honnibal, M. and Montani, I. (2026). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Acessado em: 16 mar. 2026.

Insights, S. (2025). New global social media research summary 2025. `https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/`. Acessado em: 16 mar. 2026.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, 2 edition.

Karthikeyan, T. and Ravikumar, N. (2014). A survey on association rule mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(1):2278–1021.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). *Visual Analytics: Scope and Challenges*, pages 76–90. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kudriavtsev, M., McCarren, A., Lee, H., and Bezbradica, M. (2024). Efficient visualization of association rule mining using the trie of rules.

Kumbhare, T. A. and Chobe, S. V. (2014). An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies*, 5(1):927–930.

Mining, W. I. D. (2006). Data mining: Concepts and techniques. *Morgan Kaufinann*, 10(559-569):4.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Shneiderman, B., Plaisant, C., Hesse, B., Heine, L., Rose, A., and Harkey, D. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343.

Statista (2025). Social networks - statistics & facts. `https://www.statista.com/topics/1164/social-networks`. Acessado em: 16 mar. 2026.

Wattenberg, M. and Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics*, 14(6):1221–1228.