PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Informática

Daniel Machado Osório Pereira

Análise da Evolução de Tópicos em Redes Sociais Através da Análise Formal de Conceitos

Belo Horizonte

2023

Daniel Machado Osório Pereira

Análise da Evolução de Tópicos em Redes Sociais Através da Análise Formal de Conceitos

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientador: Prof. Dr. Mark Alan Junho Song

Belo Horizonte

2023

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Pereira, Daniel Machado Osório
P436a Análise da evolução de tópicos em redes sociais através da Análise Formal de Conceitos / Daniel Machado Osório Pereira. Belo Horizonte, 2023.
65 f. : il.

Orientador: Mark Alan Junho Song
Dissertação (Mestrado) - Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Rede social na Internet. 2. Conceitos - Análise. 3. Modelos matemáticos - Programação. 4. Algorítmos computacionais. 5. Twitter (Rede social on-line). 6. Conteúdo gerado pelo usuário. 7. Banco de dados. 8. Processamento de linguagem natural (Computação). I. Song, Mark Alan Junho. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.011
Ficha catalográfica elaborada por Daniela Luzia da Silva Gomes - CRB 6/2505

Daniel Machado Osório Pereira

Análise da Evolução de Tópicos em Redes Sociais Através da Análise Formal de Conceitos

Dissertação apresentada ao Programa de Pós-Graduação em Informática como requisito parcial para qualificação ao Grau de Mestre em Informática pela Pontifícia Universidade Católica de Minas Gerais.

Prof. Dr. Mark Alan Junho Song – PUC Minas (Orientador)

Prof. Dr. Wladmir Cardoso Brandão – PUC Minas

Prof. Dr. Sérgio Vale Aguiar Campos – DCC-UFMG

Belo Horizonte, 11 de Setembro de 2023.

Aos meus pais pelo amor e apoio incondicional,
à minha irmã pelo companheirismo,
e à Embraer por sempre me incentivar.

AGRADECIMENTOS

Muitos ajudaram a concluir esse trabalho. Em especial, o Professor Mark, pelo incentivo, apoio e por tão bem se adaptar ao distanciamento causado pela pandemia.

Contribuições importantes também foram feitas pelo Professor Wladmir, sempre transmitindo confiança e perseverança.

O companheirismo e senso de humor do senhor Pedro Pongelupe Lopes foi essencial para atravessar este desafio. Agradeço pelo apoio e incentivo para iniciar esse trabalho.

"It gets easier. Every day it gets a little easier, but you gotta do it everyday, that's the hard part."

Monkey from BoJack Horseman

RESUMO

As redes sociais atualmente geram trocas de informações *online* entre seus usuários que as alimentam com conteúdos não filtrados e que abordam diversos assuntos e acontecimentos recentes. Identificar entre esses conteúdos tópicos que estão sendo abordados por atores da rede social e entender a evolução desses tópicos durante um período de tempo é o objetivo deste trabalho. Essa dissertação propõe uma abordagem que, utilizando Análise Formal de Conceitos (AFC), consiga extrair conhecimento da evolução de um tópico dentro da rede social Twitter, identificando a hierarquia existente entre os tópicos e analisando-a.

Palavras-chave: Análise formal de conceitos, Análise de redes sociais, Twitter

ABSTRACT

Social networks currently generate exchanges of information online between their users, who feed them with unfiltered content and that address various topics and recent events. Identifying among these specific contents that are being considered by social network actors and understanding the related evolution over a period of time is the objective of this work. This project proposes an approach that, using Formal Concept Analysis (FCA), can extract knowledge of the evolution of a topic within the Twitter social network, identifying an existing hierarchy between topics and analyzing it.

Keywords: Formal concept analysis, Social network analysis, Twitter.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AFC – Análise Formal de Conceitos

PLN – Processamento de Linguagem Natural

API – *Application Programming Interface*

SUMÁRIO

1 INTRODUÇÃO

A *Internet* evoluiu significativamente ao longo dos anos, provendo um conjunto diversificado de serviços e sistemas. Atualmente, é um espaço híbrido que incorpora diferentes meios de comunicação e aplicações capazes de atingir um grande público em todo o mundo. Entre as aplicações mais populares estão as redes sociais (ZHANG et al., 2012). Isso porque elas permitem aos usuários gerar uma quantidade significativa de conteúdo que exemplifica suas impressões e vivências. No entanto, existe uma rede social em específico que se destaca por forçar seus usuários a se expressarem de forma concisa, a rede social Twitter.

Nessa rede os usuários se expressam através de *tweets* - um corpo textual com tamanho máximo de 280 caracteres na sua versão gratuita. Essa limitação de caracteres pode ser um desafio, mas também pode levar a uma expressão mais clara e objetiva de ideias. Apesar das limitações, o Twitter tem se mostrado uma plataforma importante para a disseminação de ideias e opiniões, com uma grande variedade de tópicos e discussões ocorrendo diariamente. É uma rede social que, embora imponha uma limitação de caracteres, ainda assim oferece uma plataforma de grande alcance para a expressão de opiniões e ideias.

Através do uso de *hashtags* e outras ferramentas, é possível alcançar uma audiência ainda maior, aumentando o alcance e a influência de um *tweet*. *Tweet* é um modelo textual curto e eficiente que permite aos usuários relatar rapidamente suas vivências e compartilhar informações em tempo real, tornando o Twitter um dos meios de divulgação de informação mais rápido e eficiente (CATALDI; CARO; SCHIFANELLA, 2010).

O *tweet*, devido à sua natureza textual curta, permite aos usuários reportar de forma imediata o que estão vivenciando no momento em que a postagem é realizada. Essa característica é uma das principais razões pelas quais o Twitter se tornou tão popular entre os usuários.

Diferentemente de um jornalista, por exemplo, que deve seguir protocolos rigorosos para garantir a qualidade de sua matéria, o usuário do Twitter pode relatar suas vivências sem se preocupar tanto com a qualidade gramatical ou ortográfica de sua escrita. Isso faz com que o Twitter se torne provavelmente o meio mais rápido e eficiente de divulgação de informações em todo o mundo. Como resultado, notícias importantes podem ser transmitidas em tempo real, proporcionando uma atualização imediata do que está acontecendo ao redor do mundo.

Por causa dessa capacidade única de fornecer informações em tempo real, o Twitter se tornou uma plataforma de destaque para empresas, organizações e indivíduos que desejam estar atualizados sobre os principais eventos. Além disso, muitos jornalistas passaram a utilizar o Twitter como uma ferramenta para acompanhar notícias e reportá-las em tempo real, fornecendo atualizações rápidas e precisas para seus seguidores.

Entretanto, a crescente popularidade do Twitter torna mais difícil extrair informações relevantes dado o grande número de *tweets*. No entanto, essa tarefa é essencial para muitas empresas que desejam entender o que seus clientes estão dizendo sobre seus produtos ou serviços, a fim de melhorar suas estratégias de *marketing* e atendimento ao cliente.

Por este motivo, criou-se a base de dados RepLab2013, com o apoio da Conference and Labs of the Evaluation Forum (CLEF). Essa base de dados é utilizada para testar sistemas que avaliam a reputação online de empresas e organizações. Para isso, foi gerado uma base de *tweets* que falam sobre entidades (empresas, organizações, celebridades). Logo, a base é utilizada com o objetivo de agrupar *tweets* de acordo com o assunto abordado e identificar *tweets* que são positivos ou negativos para as entidades. No entanto, esse agrupamento foi realizado de forma manual, um processo trabalhoso e propenso a erros, especialmente considerando o grande número de *tweets* gerados diariamente (AMIGO et al., 2013).

Uma alternativa para solucionar esse desafio é utilizar técnicas computacionais, como a Análise Formal de Conceitos (AFC) e o Processamento de Linguagem Natural (PLN). A AFC é uma técnica matemática que permite a organização de informações em conceitos hierárquicos, enquanto o PLN é um ramo da inteligência artificial que se concentra no desenvolvimento de sistemas capazes de compreender e processar a linguagem humana. A combinação dessas técnicas pode permitir a análise automática de grandes conjuntos de *tweets*, ajudando as empresas a entenderem melhor a opinião de seus clientes e adaptar suas estratégias.

A AFC, introduzida por Rudolf Wille em 1981, fornece uma base matemática para que sejam desenvolvidos métodos de análise de dados para obtenção de informações (WILLE, 1982). Logo, como as redes sociais possuem uma grande quantidade de dados que possuem informações relevantes, a AFC é aplicada na área Análise de Redes Sociais. Entretanto, os autores observam que poucos trabalhos utilizam AFC em contextos de alta-dimensionalidade, como é o caso do contexto das redes sociais, pelo fato dos algoritmos não se comportarem bem quando o volume de dados é alto (MISSAOUI; KUZNETSOV; OBIEDKOV, 2017).

Desta forma, nesta dissertação, o objetivo é apresentar uma abordagem que utiliza PLN para encontrar grupos de palavras recorrentes em *tweets* e depois analisar como esses

grupos de palavras se relacionam. A relação entre esses termos é medida utilizando AFC, através das métricas de suporte e confiança. Além disso, também é analisado como esses termos variam ao longo do tempo.

1.1 Justificativa

À medida que as redes sociais se tornaram uma das fontes mais rápidas de compartilhamento de informações, as publicações dos usuários das redes sociais acabaram se tornando uma forma rápida de obter informações sobre notícias, eventos e acontecimentos recentes. A explosão das mídias sociais, a característica não estruturada e dinâmica dos dados trocados e a grande quantidade de informações destacam a necessidade de criar serviços automáticos que avaliem e extraiam informações relevantes desses dados.

A AFC define uma relação entre os conceitos formais que possibilita a elaboração de um reticulado conceitual. O reticulado conceitual é uma representação gráfica que estrutura os dados de forma hierárquica e que explora correlações, semelhanças, anomalias ou mesmo inconsistências nas estruturas de dados. Representações hierárquicas são alternativas interessantes para abordar a tarefa de Detecção de Tópicos, pois os tópicos são inerentemente hierárquicos e a descoberta de tais estruturas hierárquicas garante um grande avanço para a detecção de tópicos (ZENG et al., 2011). Porém, as técnicas normalmente utilizadas para a detecção de tópicos utilizam representações de tópicos simples, sem indicar a hierarquia existente entre os tópicos.

A AFC permite o relacionamento do *tweet* com os tópicos detectados e com o instante de tempo em que o *tweet* foi publicado. Dessa forma, AFC nos permite derivar uma estrutura de conhecimento hierárquica, que leva em consideração instantes de tempos distintos, e com isso determinar quais tópicos são mais relevantes em cada instante.

Nesta dissertação é proposta a utilização da AFC e PLN para a tarefa de Detecção e Evolução de Tópicos na rede social Twitter.

1.2 Objetivos

Nesta seção são apresentados os objetivos (gerais e específicos) elaborados para esta dissertação.

1.2.1 Objetivo geral

Para solucionar o problema apresentado, propõe-se a metodologia de Detecção e Evolução de Tópicos aplicada à teoria AFC, identificando os tópicos através da PLN e

identificando a relação entre eles usando AFC, sendo que a fonte de informação são *posts* da rede social Twitter.

O principal objetivo da dissertação é detectar tópicos discutidos na rede social Twitter, para responder às seguintes questões: I) é possível identificar como os tópicos variam ao longo do tempo? II) os tópicos identificados refletem com os acontecimentos e notícias que ocorreram no mesmo instante de tempo?

Para alcançar o objetivo, a proposta para esta pesquisa consistiu em empregar a AFC e PLN. É importante ressaltar que a aplicação da AFC permite a avaliação das métricas relativas ao suporte e confiança, que indicam como os tópicos variam em diferentes instantes de tempo de acordo com a variação das métricas.

Três estudos de caso foram desenvolvidos para que o objetivo fosse alcançado. O primeiro utilizou a RepLab 2013 para analisar a entidade BMW. Além disso, também foi criada uma base de dados coletando *tweets* através da *Application Programming Interface* (API) do Twitter. O segundo estudo de caso consistiu na análise de *tweets* que discutem sobre a vacina no Brasil, durante o mês de janeiro de 2021, verificando quais termos se relacionam com vacinas e como eles evoluem ao longo do tempo. O terceiro analisou *tweets* sobre a eleição presidencial no Brasil em 2022, com foco nos dois principais candidatos, Bolsonaro e Lula.

1.2.2 Objetivos específicos

Para atingir o objetivo proposto foram desenvolvidos os seguintes objetivos específicos:

1. Identificar o atual panorama de pesquisa para realização de detecção de tópicos através de uma revisão sistemática da literatura;
2. Complementar a base de dados RepLab 2013, buscando o corpo textual dos *tweets* e a data em que eles foram publicados;
3. Criar uma base de dados com *tweets* sobre um tema específico e durante um período de tempo contínuo;
4. Avaliar a variação dos tópicos ao longo do tempo e verificar se a variação reflete com os acontecimentos que ocorreram durante o mesmo período de tempo.

1.2.3 Organização da dissertação

Esta dissertação está assim organizada: No Capítulo 2 é apresentada a fundamentação teórica sobre o tema. O Capítulo 3 contém os trabalhos relacionados a essa

dissertação. No Capítulo 4 é descrita a metodologia proposta nesta dissertação. O Capítulo 5 apresenta os resultados obtidos e as discussões derivadas desses resultados. O capítulo 6 encerra a dissertação com as conclusões e propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção é apresentado os conceitos necessários para compreender AFC e as técnicas utilizadas para obter e pré-processar os *tweets*.

2.1 Análise Formal de Conceitos

A AFC é uma teoria matemática para a formação de conceitos oriundos de um conjunto de dados, resultando em um modelo teórico que organiza as informações e expõe relações entre os dados.

2.1.1 Contexto Formal

Um contexto formal é composto por três grupos: objetos, atributos e as relações entre eles, tal como apresentado na Tabela 1. Nela, os *tweets* correspondem aos objetos, os tópicos são atributos e a relação entre ambos é definida pela incidência X que mostra as características que um certo *tweet* possui (tópico). Nesse caso, o tópico é um conjunto de palavras presente no corpo textual do *tweet*.

Um contexto formal é matematicamente definido como uma tripla (O, A, I), em que O é o conjunto de objetos, A é o conjunto de atributos e I é a relação de incidência binária entre esses objetos e atributos, de forma que $I \subseteq O \times A$. Um contexto formal é comumente representado em forma de tabela, cujas linhas correspondem aos objetos, as colunas correspondem aos atributos e as interseções destes representam as incidências.

Considerando um contexto formal $K = (O, A, I)$, quando um objeto a pertencente a O e um atributo a pertencente a A possuem a relação de incidência i pertencente a I, significa que o objeto o tem o atributo a. Isso pode ser dito pela notação $(o, a) \in I$ ou mesmo oIa, em que $o \in O$ e $a \in A$.

2.1.2 Conceito Formal

Um conceito formal é um par (A, B), sendo A um grupo de objetos e B um grupo de atributos. A relação existente entre eles implica no fato de A ser objetos relacionados

	Used BMW	Pay Online	BMW X5	BMW M3
Tweet 1	**X**			
Tweet 2		**X**		**X**
Tweet 3	**X**	**X**		**X**
Tweet 4			**X**	

Tabela 1 – Exemplo de um contexto formal

Objetos	Atributos
{Tweet 1, Tweet 2, Tweet 3, Tweet 4}	{}
{Tweet 4}	{BMW X5}
{Tweet 1, Tweet 3}	{Used BMW}
{Tweet 2, Tweet 3}	{Pay Online, BMW M3}
{}	{Used BMW, Pay Online, BMW X5, BMW M3}

Tabela 2 – Conceitos existentes no contexto formal da Tabela 1

com B e B ser atributos que se relacionam com os objetos de A. Logo podemos afirmar que B implica em A e que A implica em B. O par (A, B) só pode ser considerado um conceito formal se $A = B'$ e $B = A'$. Essa relação é explicitada de pelo operador de derivação (') definido a seguir:

$$\begin{aligned} A' &= \{a \in A | \forall o \in A, (o, a) \in I\} \\ B' &= \{o \in O | \forall a \in B, (o, a) \in I\} \end{aligned} \tag{2.1}$$

Para exemplificar, usando a Tabela 1, os objetos A = {Tweet 2, Tweet 3} e atributos B = {Pay Online, BMW M3} ao serem submetidos à operação descrita acima, terão o resultado $A' = B$ = {Pay Online, BMW M3}. Repare que B' = A = {Tweet 2, Tweet 3}. Portanto, {{Tweet 2, Tweet 3}, {Pay Online, BMW M3}} é um conceito. Todos os conceitos encontrados na Tabela 1 estão exibidos na Tabela 2. É possível reparar a existência de um conceito com o conjunto de atributos vazio e um conceito com o conjunto de objetos vazio. Eles são chamados de *infimum* e *supremum*, respectivamente.

2.1.3 Regras de implicação

As implicações são dependências entre elementos de um conjunto obtidos de um contexto formal. Dado o contexto $K = (O, A, I)$ as regras de implicação assumem a forma $B \rightarrow C$ se e somente se $B, C \subset A$ e $B' \subset C'$. Uma regra de implicação $B \rightarrow C$ é considerada válida se e somente se, todo objeto que possuir os atributos de B também possuir os atributos de C.

Regra	Suporte	Confiança
Pay Online → BMW M3	50%	100%
Used BMW → Pay Online, BMW M3	25%	50%

Tabela 3 – Exemplo de regras com suporte e confiança

Considere que $K = (O, A, I)$ seja um contexto formal. A partir deste contexto pode-se definir regras, da seguinte forma: $r : A \rightarrow B(s, c)$, sendo que $A, B \subseteq A$ e $A \cap B = \varnothing$.

Também pode ser definido o suporte às regras, que é dado por (AGRAWAL; SRIKANT, 1994):

$$s = supp(r) = \frac{|A' \cap B'|}{|O|}$$

e a confiança que é dada por:

$$c = conf(r) = \frac{|A' \cap B'|}{|A'|}$$

A Tabela 3 mostra 2 regras existentes no contexto da Tabela 1. A regra Pay Online → BMW M3 possui suporte de 50% pelo fato de em 2 *tweets* essa regra acontecer, dentro de um total de 4 *tweets*. Já a confiança é de 100%, visto que sempre que um *tweet* possui Pay Online ele também possui BMW M3. Define-se implicações como regras que apresentam 100% de confiança, ou seja, a regra Pay Online → BMW M3 é uma implicação.

2.2 Análise Formal de Conceitos Triádicos

2.2.1 Contexto Triádico

Um contexto triádico β é uma quádrupla (X, Y, Z, I), onde X, Y e Z são conjuntos não vazios e I uma relação ternária entre X, Y e Z. O conjunto X representa os objetos, o conjunto Y os atributos e o conjunto Z representa a condição para que ocorra a relação entre objeto e atributo. Nesse contexto I é interpretado como a incidência entre os outros conjuntos, mostrando que um objeto x (x ∈ X) possui certo atributo y (y ∈ Y) dependendo de certa condição z (z ∈ Z) (BELOHLAVEK; OSICKA, 2010).

As Tabelas 4 e 5 mostram um contexto triádico também sobre termos presentes em *tweets*. A diferença em relação ao contexto diádico apresentado anteriormente é o acréscimo da condição mês, que fornece uma dimensão temporal ao conceito. Logo, o contexto triádico possui 3 dimensões, que no exemplo das Tabelas 4 e 5 são os *tweets*, os termos e a data de publicação dos *tweets*.

	Mês 1				Mês 2			
	UsedBmw	PayOnline	BmwX5	BmwM3	UsedBmw	PayOnline	BmwX5	BmwM3
Tweet 1	**X**					**X**	**X**	
Tweet 2		**X**		**X**	**X**			
Tweet 3	**X**	**X**		**X**			**X**	
Tweet 4			**X**		**X**			**X**

Tabela 4 – Exemplo de um contexto triádico

	Mês 1	**Mês 2**
Tweet 1	Used BMW	Pay Online, BMW X5
Tweet 2	Pay Online, BMW M3	Used BMW
Tweet 3	Used BMW, Pay Online, BMW M3	BMW X5
Tweet 4	BMW X5	Used BMW, BMW M3

Tabela 5 – Conceitos existentes no contexto triádico da Tabela 4

2.2.2 Conceito Triádico

Um conceito triádico, obtido a partir de um contexto triádico β (X_1, X_2, X_3, I), é uma tripla representada por (C_1, C_2, C_3), sendo que $C_1 \subseteq X_1$, e o mesmo vale para $C_2 \subseteq X_2$ e $C_3 \subseteq X_3$. Logo, o conceito triádico é composto por C_1, também denominado *extent*, C_2 *intent* e C_3 *modus*.

Para encontrar um conceito triádico é preciso identificar um grupo de atributos que possua um grupo de objetos e que se relacionam de acordo com um grupo de condições. A Tabela 6 mostra conceitos triádicos identificados no contexto triádico da Tabela 4.

Todo contexto triádico possui no mínimo 3 conceitos (triviais). Esses conceitos aparecem nas 3 primeiras linhas da Tabela 6. A similaridade desses conceitos é que todos possuem uma dimensão vazia, seja a *extent*, *intent* ou *modus*. Considerando o contexto triádico $\mathbb{K} := (K_1, K_2, K_3, Y)$, podemos encontrar os 3 conceitos triviais através das seguintes equações (LEHMANN; WILLE, 1995):

$$\begin{aligned} o_1 &:= ((K_2 \times K_3)^{(1)}, K_2, K_3) \\ o_2 &:= (K_1, (K_1 \times K_3)^{(2)}, K_3) \\ o_3 &:= (K_1, K_2, (K_1 \times K_2)^{(3)}) \end{aligned} \tag{2.2}$$

Objetos	Atributos	Condições
{}	{Mês 1, Mês 2}	{Used BMW, Pay Online, BMW X5, BMW M3}
{Tweet1, Tweet2, Tweet3, Tweet4}	{}	{Used BMW, Pay Online, BMW X5, BMW M3}
{Tweet1, Tweet2, Tweet3, Tweet4}	{Mês 1, Mês 2}	{}
{Tweet1, Tweet3}	{Mês 1}	{Used BMW}
{Tweet2, Tweet4}	{Mês 2}	{Used BMW}
{Tweet1, Tweet3}	{Mês 2}	{BMW X5}
{Tweet3}	{Mês 1}	{Used BMW, Pay Online, BMW M3}

Tabela 6 – Conceitos Triádicos

Regra de Associação	Suporte	Confiança
(Pay Online → BMW M3) Mês 1	50%	100%
(Pay Online → BMW X5) Mês 2	25%	100%
(BMW X5 → Pay Online) Mês 2	25%	50%

Tabela 7 – Regras de Implicação da Tabela 4

2.2.3 Regras de Implicação Triádicas

A maioria dos estudos da AFC é focada na representação de regras em contextos diádicos. O precursor no estudo de extrair regras de implicação em um contexto triádico foi Klaus Biedermann (BIEDERMANN, 1997). Em seguida, outro trabalho (GANTER; OBIEDKOV, 2004) apresentou outra abordagem para extração de regras chamada Regra de Associação de Condição de Atribuição.

A Regra de Associação de Condição de Atribuição descrita por Biedermann pode ser descrita por $(C_1 \rightarrow C_2)c(sup, conf), C_1, C_2 \subseteq K_3$ and $A \subseteq K_2$. Ou seja, se a condição de C_1 ocorrer nos atributos de A então a condição de C_2 também irá ocorrer para os atributos de A, com certo suporte e confiança. A Tabela 7 apresenta regras de associação de condição de atribuição encontradas no contexto triádico descrito na Tabela 4.

Observando a primeira regra de associação (Termo 2 → Termo 4) Mês 1 que possui 100% de suporte e 50% de confiança, é possível afirmar que sempre que um *tweet* possuir o termo 2 ele também possuirá o Termo 4, e isso ocorre em 50% da base de dados analisada.

2.3 Coleta de *tweets*

Nesta dissertação foram criadas bases de dados com *tweets* extraídos através da API do Twitter. Com isso, foi possível avaliar temas específicos como a eleição presidencial no Brasil, extraindo *tweets* durante um período de tempo.

Esta tarefa foi realizada com a criação de robôs que diariamente buscavam *tweets* na API do Twitter. A API não fornece *tweets* anteriores a um período de 7 dias, por isso

foi necessário a criação dos robôs que buscavam *tweets* diariamente por um determinado período de tempo.

A API também limita a quantidade de *tweets* que podem ser extraídos diariamente. Em consequência, um número restrito de *tweets* pôde ser coletado. Logo, essas bases de *tweets* que construímos não possuem uma ordem de grandeza de milhões de *tweets*, e sim uma ordem de milhares de *tweets*.

2.3.1 Pré-processamento

Bases de dados textuais, como a RepLab 2013, precisam ser pré-processadas antes de serem analizadas. As etapas realizadas neste trabalho são as seguintes: *tokenization*, *stop word removal* e *stemming*.

- Tokenization: consiste em dividir uma sentença em palavras, removendo os sinais de pontuação, que não são significativos;
- Stop word removal: consiste em remover palavras como artigos e preposições, pelo fato dessas palavras não serem significativas para análises textuais (YOGISH; MANJUNATH; HEGADI, 2019);
- Stemming: consiste em transformar a palavra em seu radical. Exemplo: a palavra correr é transformada para corr;
- N-Gram: é uma sequência de n palavras em sequência extraídas de uma base textual, nessa dissertação foram extráidos bigrams e trigrams, ou seja, duas ou três palavras em sequência (BANERJEE; PEDERSEN, 2003).

As etapas descritas acima foram aplicadas através do pacote NLTK da linguagem Python. O processo de Tokenization consiste em transformar sentenças, que são um conjunto de palavras, em uma lista de palavras, que se agrupadas novamente voltam a formar uma sentença. O pacote NLTK possui uma lista pronta de *stop words*, como *"the", "a", "an", "in"*, logo essas palavras da lista são removidas da base de dados que está sendo pré-processada, já que essas palavras não são significativas para a análise. Isso possibilita que a base de dados após o pré-processamento tenha um tamanho reduzido e ainda reduz o tempo de análise (CONTRERAS; HILLES; ABUBAKAR, 2018).

Para realizar o *stemming* foi empregado o algoritmo de Porter por apresentar a menor taxa de erro. Mesmo sendo desenvolvido em 1980 o algoritmo de Porter ainda é muito utilizado, visto que possui mais de 12 mil citações, sendo que mais de 1 mil das citações foram feitas no século XXI. O fato do *stemmer* Porter poder ser aplicado a

um variado número de idiomas justifica o grande interesse da comunidade científica pelo algoritmo (WILLETT, 2006).

O algoritmo de Porter basicamente define regras para identificar plurais e tempos verbais nos sufixos das palavras. O algoritmo passa por várias iterações até que chegue no radical final da palavra, por exemplo, a palavra *Generalizations* é reduzida para *Generalization* na primeira iteração, na segunda iteração para *Generalize*, na terceira iteração para *General* e na quarta e última iteração o algoritmo define o radical *Gener* (PORTER, 1980).

2.3.2 Redução de Contexto Formal

A AFC é uma teoria matemática que trata bem desafios de extrair conhecimento de bases de dados. Porém, quando essa base possui muitos atributos, gera um contexto formal muito extenso. Para evitar essa situação se fazem necessárias técnicas para reduzir o contexto formal.

Para diminuir o tamanho do contexto vindo de uma base de dados extraída do Twitter podem ser utilizadas técnicas de seleção e balanceamento de atributos de um contexto formal (RECUERO, 2008). Com isso é possível selecionar apenas os atributos relevantes e reduzir o tamanho do contexto formal. A técnica frequência é capaz de reduzir o tamanho de um contexto.

Essa técnica pode ser dividida em duas etapas. Primeiro serão descartados os atributos que menos se relacionam com os objetos e depois serão selecionados os atributos que mais se relacionam com objetos. É necessário selecionar os atributos que mais se relacionam com os objetos pois a próxima etapa da técnica não garante que esses atributos serão escolhidos (RECUERO, 2008).

A segunda etapa da técnica consiste em selecionar o atributo que possui o maior número de relações e depois remover esse atributo e todos os objetos que se relacionam com ele do contexto formal. Essa etapa se repete até todos os objetos terem sido removidos ou quando o número de atributos desejados ter sido alcançado.

As Tabelas 8 e 9 exemplificam como funcionariam as iterações da segunda etapa. A primeira iteração escolhe o Termo 2, por possuir o maior número de relações com tweets, e remove os tweets que possuíam esse termo e a segunda iteração escolhe o Termo 1, por possuir o maior número de relações com tweets entre os termos restantes. A Tabela 10 mostra o resultado final da técnica, que é um contexto reduzido.

	Termo 1	Termo 2	Termo 3	Termo 4	Termo 5	Termo 6
Tweet 1	**X**	**X**			**X**	**X**
Tweet 2	**X**		**X**			**X**
Tweet 3		**X**		**X**	**X**	
Tweet 4		**X**	**X**			
Tweet 5	**X**					
Tweet 6		**X**	**X**	**X**	**X**	

Tabela 8 – Primeira iteração

	Termo 1	Termo 3	Termo 4	Termo 5	Termo 6
Tweet 2	**X**	**X**			**X**
Tweet 5	**X**				

Tabela 9 – Segunda iteração

	Termo 1	Termo 2
Tweet 1	**X**	**X**
Tweet 2	**X**	
Tweet 3		**X**
Tweet 4		**X**
Tweet 5	**X**	
Tweet 6		**X**

Tabela 10 – Resultado final

3 TRABALHOS RELACIONADOS

Existem diversos trabalhos que são relevantes ao contexto deste estudo. Tratam-se de trabalhos no contexto de detecção de tópicos em redes sociais, evolução de tópicos e classificação de corpos textuais. Alguns desses trabalhos são descritos nessa seção.

Zhang et al. (ZHANG et al., 2012) detalha como a detecção de tópicos na Internet é um desafio pelo fato das informações produzidas serem sucintas e não descreverem de forma adequada o real contexto que está sendo abordado. Para solucionar essa característica das informações produzidas na Internet os autores utilizaram a técnica *pseudo-relevance feedback*, que consiste em adicionar informação ao dado que está sendo analisado.

Com essa estratégia os autores conseguiram aprimorar as informações produzidas na Internet, aprimorando o contexto que essas informações estão tratando, e com isso conseguir identificar dentro dessas informações quais se tornarão no futuro mais presentes dentro da Internet. Essa dissertação também busca detectar tópicos de conteúdos produzidos na Internet, porém não utilizamos a técnica *pseudo-relevance feedback*, visto que a base de dados RepLab 2013 já nos fornece o contexto em que os conteúdos analisados estão inseridos.

Cataldi et al. (CATALDI; CARO; SCHIFANELLA, 2010) utilizou da técnica de detecção de tópicos para identificar tópicos emergentes na comunidade do Twitter. Os autores conseguiram realizar a identificação considerando que se o tópico ocorre frequentemente no presente e era raro no passado, e assim caracterizavam-os como emergentes. Para incrementar a estratégia abordada foi feita uma análise dos autores desses tópicos emergentes através do algoritmo *Page Rank*, para garantir que o tópico emergente não está presente apenas em alguma bolha da comunidade do Twitter. Por último, foi criado um grafo que conecta o tópico emergente com outros tópicos que se relacionam com ele, e que por isso possuem uma chance maior de também se tornarem tópicos emergentes. Diferentemente do trabalho descrito acima, essa dissertação tem como objetivo utilizar a detecção de tópicos para agrupar tweets em entidades, garantindo que os tweets agrupados abordem o mesmo assunto.

Também levando em conta a questão de avaliar os autores das redes sociais, o tra-

balho de Miao et al. (MIAO et al., 2016) desenvolveu um *framework* que encontra usuários relevantes da rede social para que seja feita a análise somente do conteúdo publicado por esses autores. Mesmo com essa análise restrita, o *framework* desenvolvido consegue identificar 92% dos *trending topics* existentes no Twitter, mostrando que o *framework* se destaca em relação aos outros já propostos.

Dragoş et al. (SM.; C.; DF., 2017) apresentam uma abordagem que investiga o comportamento de usuários de uma plataforma de aprendizagem utilizando AFC. Para identificar o perfil desses estudantes é analisado o *log* gerado pela plataforma, que contém informações sobre as ações que cada estudante está executando na plataforma.

O uso de AFC por Dragoş et al. ocorre para considerar o instante de tempo que as ações são executadas pelos estudantes. É relevante para traçar o perfil dos estudantes entender se ele está executando as ações de forma atrasada, adiantada ou no tempo certo. Portanto, AFC pode ser considerada como alternativa para estudar eventos temporais.

Cigarrán et al. (CIGARRAN; CASTELLANOS; GARCíA-SERRANO, 2016) utilizou de AFC para agrupar tweets de acordo com os tópicos encontrados. Por isso a escolha da base de dados RepLab 2013, que já faz esse agrupamento dos tweets em entidades, baseado no conteúdo textual do tweet. Por usar AFC o trabalho ainda consegue obter um reticulado conceitual dos tópicos encontrados, obtendo uma visão hierárquica dos tópicos, sendo isso um diferencial em relação a outras técnicas. A proposta esteve entre os melhores resultados do fórum RepLab 2013, provando a eficácia da AFC para o desafio de detecção de tópicos.

Recuero et al. (RECUERO, 2008) apresentou uma técnica para balanceamento de contextos formais com foco na recuperação de informação sobre dados de motores de busca na Internet. A técnica consiste em balancear esses dados, selecionando apenas as informações que possibilitem a criação de *clusters* dentro desses dados. Nesta dissertação foi utilizada a técnica para balancear um contexto formal gerado a partir de dados da rede social Twitter, possibilitando a redução desse contexto formal.

Amigó et al. (AMIGO et al., 2013) descreve a organização e os resultados da RepLab 2013, que tem como foco monitorar a reputação de empresas e indivíduos através da opinião de usuários do Twitter. Isso é feito através da divisão dos tweets em entidades, sendo que cada entidade engloba uma empresa ou um indivíduo. Dentro das entidades é avaliado se o tweet apresenta aspectos positivos ou negativos a entidade. Nesta dissertação não será observado se os tweets possuem aspecto positivo ou negativo a entidade, o foco será na detecção de tópicos presentes nos tweets e como eles variam ao longo do tempo.

Castellanos et al. (CASTELLANOS; CIGARRAN; GARCíA-SERRANO, 2017) utiliza de AFC para detectar tópicos na Rep Lab 2013. Para detectar quais conceitos encontrados refletiam em tópicos, eles utilizaram a técnica de calcular a estabilidade do

conceito, que determina se um conceito continuará existindo mesmo que uma quantidade aleatória de tweets seja removido. Ou seja, garantir que o conceito não existe apenas por um grupo pequeno de tweets (objetos).

Willett et al. (WILLETT, 2006) descreve como o algoritmo de Porter para stemming evoluiu ao longo do tempo, reduzindo a taxa de erro ao buscar o radical das palavras e aumentando o número de idiomas que o algoritmo consegue abranger. Com isso, mesmo com o fato de o algoritmo ter sido desenvolvido no século passado ele ainda é amplamente utilizado atualmente, tendo um grande número de citações em trabalhos recentes. Este trabalho foi desenvolvido à luz da análise do algoritmo Porter, principalmente pelo fato da RepLab 2013 possuir tweets publicados nos idiomas Inglês e Espanhol.

Ali et al. (ALI; LI; PEDRYCZ, 2023) utiliza AFC para entender os níveis de granularidade da base temporal sobre eventos capturados de câmeras de vigilância. Para relacionar os eventos foram feitas implicações Fuzzy intuicionistas, que consideraram tanto o momento quanto o lugar que os eventos ocorreram. O objetivo é que essas implicações forneçam a periodicidade que eventos similares ocorrem, possibilitando assim a previsão de eventos futuros.

Nesta dissertação também tratamos uma base temporal, que consiste de tweets que foram publicados em certo instante de tempo. Porém, utilizamos a própria AFC para relacionar esses tweets ao longo do tempo, ao contrário do trabalho citado acima que utilizou de implicações Fuzzy.

Ren et al. (REN; LI; ZHAI, 2023) propõe o uso de atributos negados ao construir um contexto formal. Com isso é possível gerar regras de implicação que relacionam os atributos originais com suas negações, garantindo uma melhor tomada de decisão ao analisar um certo conjunto de dados. Nesta dissertação também utilizamos regras de implicação para extrair conhecimento de uma base de dados, porém não consideramos a negação de atributos ao construir um contexto formal.

Murshed et al. (MURSHED et al., 2022) criou uma base de dados de tweets com possibilidade de serem tweets de usuários praticando cyberbullying. Foi utilizada a API do Twitter para extrair esses tweets da rede social e depois foi feita uma classificação manual desses tweets, entre tweets que continham ou não cyberbullying.

Nesta dissertação também utilizamos a API do Twitter para criar uma base de dados de *tweets* para ser analisada, porém não foi necessária a classificação manual desses *tweets*.

Portanto, essa dissertação contribui ao inserir uma nova abordagem sobre a detecção de tópicos em redes sociais, que é o uso da AFC para detectar tópicos e ainda avaliar a evolução temporal dos mesmos. Assim como nos trabalhos relacionados, emprega-se a

abordagem de analisar publicações feitas em redes sociais e conseguir extrair informações relevantes desse conteúdo.

4 METODOLOGIA

Neste capítulo é apresentada a metodologia de pesquisa proposta para esta dissertação. Para isso, foram realizadas as etapas descritas na Figura 4.

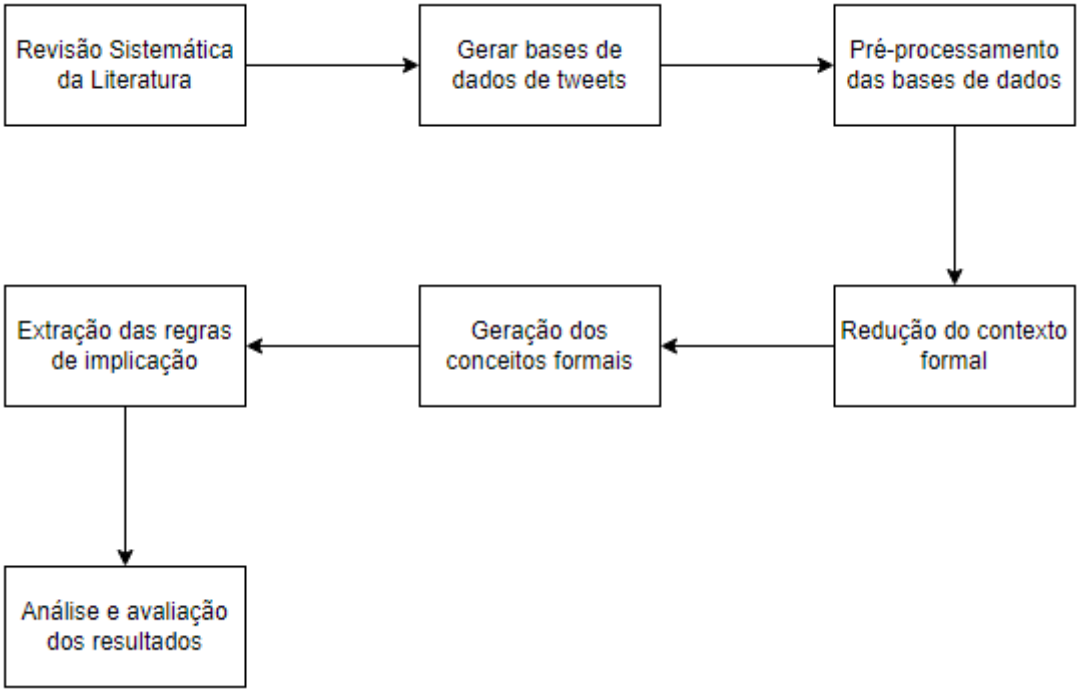

Figura 1 – Metodologia adotada

4.1 Revisão Sistemática da Literatura

Em primeiro lugar, foi feito um levantamento bibliográfico através de uma Revisão Sistemática da Literatura (RSL). O objetivo dessa RSL foi verificar quais técnicas são utilizadas para realizar detecção de tópicos em redes sociais, como o Twitter, e como a AFC e a PLN vem sendo empregadas no contexto de redes sociais. Também foi feito um estudo para a construção do referencial teórico dessa dissertação, através da leitura sobre AFC e PLN.

tweet_id	tweet_text	tweet_timestamp	entity_id
205888692580126000	#radensaleh is not a myth. Lea	2012-05-25 05:11:04+00:00	RL2013D01E001
207430942028079000	The new BMW 3 Series is awar	2012-05-29 11:19:25+00:00	RL2013D01E001
208204757779759000	@GEAGarratt BMW hand over	2012-05-31 14:34:17+00:00	RL2013D01E001
208283774251831000	I asked Sauber about more inf	2012-05-31 19:48:16+00:00	RL2013D01E001
208347487822557000	In another life, I found motorc	2012-06-01 00:01:26+00:00	RL2013D01E001

Figura 2 – Exemplo de tweets com a data de publicação e sua entidade

4.2 Base de dados RepLab2013

A Campanha de Avaliação Replab 2013 é um fórum internacional de experimentos e avaliação no campo de *Online Reputation Management*. Um dos desafios abordados no fórum é a classificação de *tweets* dentro de entidades, que identificam os temas que o *tweet* aborda.

A base de dados Replab 2013 consiste em um grupo de *tweets* relacionados com 61 entidades que foram extraídos entre 1º de junho de 2012 e 31 de dezembro de 2012. Essas entidades estão divididas em quatro domínios, sendo eles: automóveis, entidades financeiras, universidades e música/artistas.

Essa base de dados foi escolhida pelo fato de que os trabalhos (AMIGO et al., 2013) (CASTELLANOS; CIGARRAN; GARCíA-SERRANO, 2017) (CIGARRAN; CASTELLANOS; GARCíA-SERRANO, 2016) que abordam detecção de tópicos utilizam a RepLab 2013 para validar os modelos propostos e também por a base ter rotulado os *tweets* a entidade que ele pertence. Esse processo de atribuir rótulos aos *tweets* foi feito de forma manual, sendo por isso uma ótima forma para validação de novas metodologias propostas.

Nesta dissertação, a metodologia proposta é utilizar a AFC e PLN para identificar tópicos em *tweets* e avaliar como esses tópicos variam ao longo do tempo. Esse resultado é comparado com os acontecimentos e notícias referentes ao período de tempo analisado. Observa-se quais desses acontecimentos foram mais relevantes comprovando ou não a eficácia da metodologia.

4.3 Criação das bases de tweets sobre vacina e eleição presidencial

Para criar ambas as bases foi utilizado um script Python que executava diariamente e coletava os *tweets* mais relevantes do dia, utilizando o filtro da própria API do Twitter. Esse filtro faz com que apenas *tweets* que tenham tido amplo alcance dentro da rede social sejam retornados.

Com esse *script* foram obtidos 105 *tweets* sobre a vacinação que ocorreu no Brasil no período de 5 de janeiro de 2021 e dia 17 de janeiro de 2021. Esse número menor de *tweets* se justifica ao filtro utilizado na API do Twitter que retornava apenas os *tweets* que tiveram um maior alcance.

Já o *script* que coletou *tweets* sobre a eleição presidencial no Brasil coletou 3634 *tweets* no período de 23 de julho de 2022 e 8 de setembro de 2022. O grande período de coleta foi um diferencial para obter melhores resultados.

Os desafios encontrados durante esse processo são as limitações que a API do Twitter possui, limitando o número de requisições que podem ser feitas por cada usuário. Atualmente as limitações estão ainda maiores, visto que não é possível utilizar a API do Twitter de forma gratuita, o que atrapalha a reprodutibilidade desse trabalho.

4.4 Integração e Pré-processamento da base de dados

Para obter todas as informações necessárias para realizar o trabalho foi feita uma integração com a API do Twitter para recuperar o corpo textual e as datas de publicação dos tweets presentes na Rep Lab 2013. A Rep Lab 2013 não possui essas informações para respeitar a privacidade dos autores dos tweets, que caso excluam os tweets irão impedir que sua postagem continue sendo utilizada por trabalhos que utilizam a Rep Lab 2013.

Com a integração realizada foi possível recuperar 32402 tweets e a data de postagem. Com todas as informações necessárias obtidas, a próxima etapa é realizar o pré-processamento da base de dados, de forma com que o corpo textual dos tweets seja transformado em uma lista de palavras que serão analisadas.

Para realizar a tarefa foram utilizadas as técnicas descritas na Seção 2.3.1, que foram aplicadas na base de dados integrada da Rep Lab 2013.

Com isso, foi possível dividir o corpo textual do tweet, que possui como forma original o seguinte texto: “#radensaleh is not a myth. Learn about his life. Bring your kids to Galeri Nasional BMW_Indonesia jer_in”, em vários tópicos, como mostra a Tabela 11. Cada tópico está vinculado a um tweet ID, que serão os objetos do contexto formal. Já os tópicos serão os atributos do contexto formal.

Logo, o resultado final dessa etapa é a criação de uma base de dados que possui o tweet Id, os tópicos extraídos desse tweet, a data em que o tweet foi postado e a entidade a que esse tweet pertence. A Figura 3 mostra um exemplo dessas informações.

Tweet_id	Termo
205888692580126000	radensaleh
205888692580126000	myth
205888692580126000	learn
205888692580126000	lif
205888692580126000	bring
205888692580126000	kid
205888692580126000	galer
205888692580126000	bmw_indonesia
205888692580126000	jer_in

Tabela 11 – Tweet e seus termos

tweet_id	topic	tweet_timestamp	entity
207430942028079000	the	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	new	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	bmw	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	3	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	sery	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	award	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	5	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	star	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	euro	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	ncap	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	crash	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	test	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	read	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	new	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	pag	2012-05-29 11:19:25+00:00	RL2013D01E001
207430942028079000	//t.co/4r8mblje	2012-05-29 11:19:25+00:00	RL2013D01E001

Figura 3 – Informações obtidas no final do processo

4.5 Redução do contexto formal

Após a etapa de pré-processamento será necessário reduzir o contexto formal a fim de promover tópicos que não são significativos por estarem associados a poucos *tweets*. Logo, além de remover atributos do contexto formal essa etapa também ajudará a diminuir o número de conceitos encontrados, facilitando a análise final do trabalho, que é a validação dos conceitos encontrados.

Para realizar essa redução serão utilizadas as técnicas descrita na Seção 2.3, que consistem em retirar os tópicos que possuem relação com poucos tweets e manter os tópicos que se relacionam com a maioria dos tweets e remover conceitos que dependem de um número pequeno de tweets, conceitos que não são coesos.

A escolha das técnicas descritas na Seção 2.3 se deve pelo fato de outros trabalhos (CASTELLANOS; CIGARRAN; GARCíA-SERRANO, 2017) (CIGARRAN; CASTELLANOS; GARCíA-SERRANO, 2016) que aplicaram AFC na Rep Lab 2013 utilizaram essas técnicas e conseguiram bons resultados, estando entre os melhores trabalhos da conferência Rep Lab.

Para realizar a seleção dos tópicos que estão presentes no maior número possível de *tweets* foi utilizado um *script* Python que contava em quantos *tweets* o termo estava presente e fazia a ordenação desses tópicos, listando primeiro aqueles com maior número de aparições. Esses tópicos com maior número de presença em *tweets* foram os selecionados para compor o contexto formal.

Mesmo realizando essa seleção o suporte encontrado nas regras de implicação foi baixo, demonstrando que a base de dados é esparsa. Logo, é necessário um maior foco no processo de agrupar tópicos semelhantes, para que esses grupos possuam uma presença no maior número possível de tweets. Portanto, seria interessante utilizar técnicas de pré-processamento que realizam a tarefa de agrupar palavras semelhantes, sendo que esse grupo formado poderia ser o tópico utilizado dentro do contexto formal, garantindo regras com um suporte maior.

4.6 Lattice Miner

Com os tópicos já definidos é necessário preparar os arquivos que servirão como entrada de dados para a ferramenta Lattice Miner. Essa ferramenta é um *software* para construção, visualização e manipulação de contextos. Foi desenvolvido na Universidade de Québec sob a supervisão da professora Rokia Missaoui. A ferramenta lê arquivos no formato JSON para que consiga gerar o contexto formal. Para isso foi feito um script em Python que gere esses arquivos no formato esperado pela ferramenta.

Esse arquivo JSON precisa informar os objetos do contexto formal, os atributos, a condição e qual a relação entre eles. Os objetos são os IDs dos tweets, os atributos os tópicos encontrados e a condição é o dia em que o tweet foi publicado. A Figura 4 mostra o arquivo gerado para analisar tweets obtidos durante o período eleitoral brasileiro.

Com os arquivos JSON prontos é possível carregá-los dentro da ferramenta Lattice Miner e assim gerar o contexto formal. Com o contexto formal construído é possível extrair as regras de implicações para serem analisadas, gerando assim o resultado esperado. As regras de implicação são fornecidas através de arquivos XML e foram analisadas manualmente, visto que poucas regras foram geradas devido ao suporte baixo que eles apresentaram.

De forma manual as regras foram analisadas com foco nas regras que possuíam algum significado relevante e que seria possível discutir na seção de Resultados desta dissertação. De forma prática foi feito uma busca por regras que exemplificam acontecimentos ou sentimentos dos usuários por alguma empresa, pessoa pública ou evento.

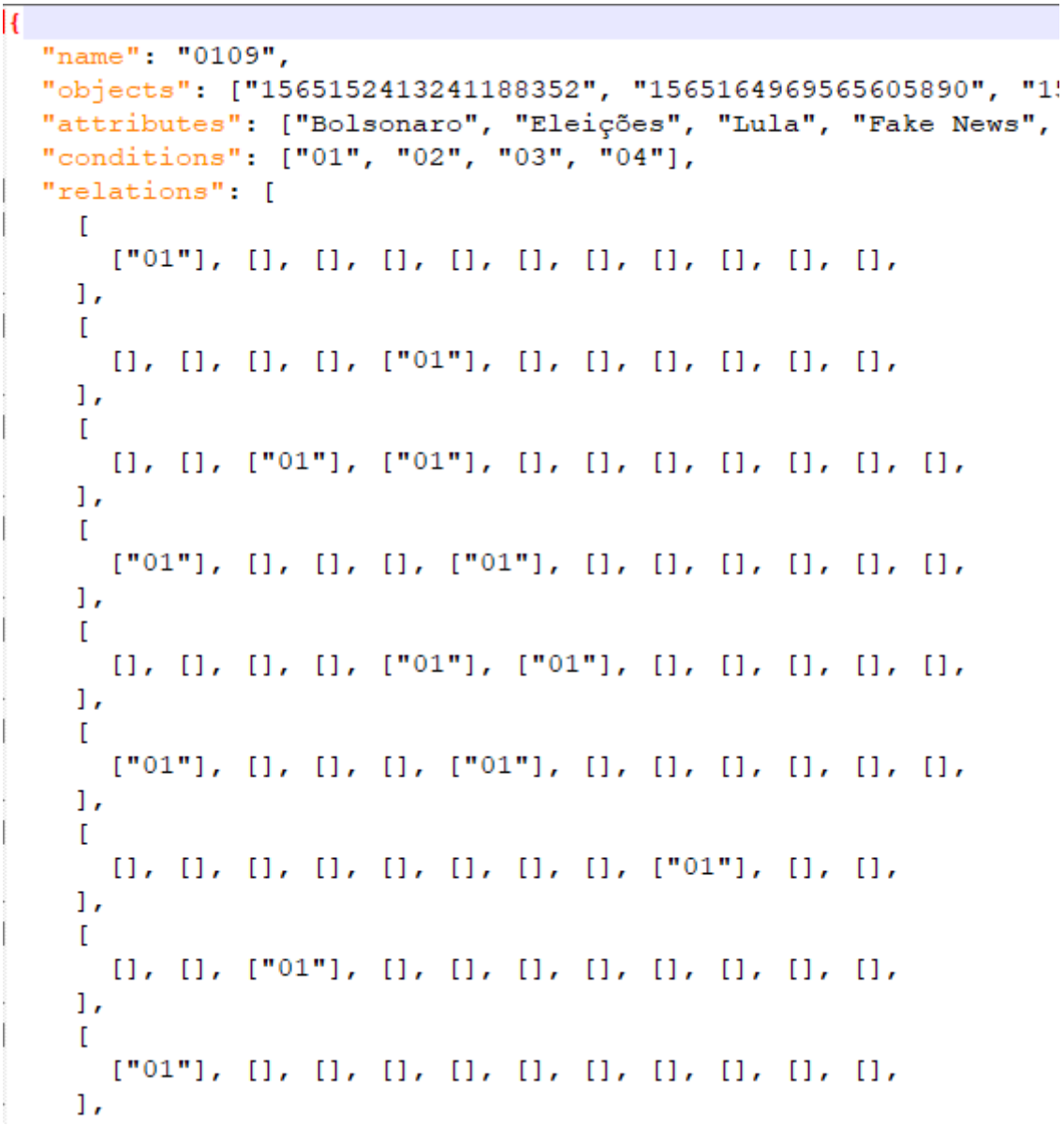

Figura 4 – Exemplo do arquivo JSON gerado

5 RESULTADOS E DISCUSSÕES

5.1 RepLab 2013 - Entidade BMW

O primeiro resultado obtido foi a análise do tópico "BMW vehicles for sale" extraído da entidade BMW. Usando tweets da RepLab 2013 foi analisado quais modelos de carro da BMW estavam sendo anunciados para venda no Twitter. O resultado está na Tabela 12.

Tabela 12 – Tópico For Sale variando ao longo do tempo.

Day	Antecedent	Consequence	Support	Confidence
1	BMW Series	For Sale	0.65%	50%
1	BMW M3	For Sale	1.3%	50%
2	BMW Series	For Sale	0.65%	20%
2	BMW X5	For Sale	0.65%	50%
2	BMW Z4	For Sale	0.65%	100%
3	BMW M3	For Sale	1.3%	33%
3	BMW Z4	For Sale	0.65%	50%
3	BMW X6	For Sale	0.65%	100%
4	BMW Series	For Sale	1.95%	50%

Ao analisar as regras da Tabela 12, fica evidente que a métrica de suporte é baixa para todas as regras identificadas. Isso sugere que os termos presentes nos tweets são dispersos e variados, havendo uma ampla gama de termos mencionados. Essa dispersão de termos está diretamente relacionada à query de busca utilizada na análise. Quando uma query genérica é empregada, é comum que a base de resultados seja esparsa, com diversos termos presentes e poucas ocorrências em comum.

Outro fator que contribui para a base esparsa é o pré-processamento realizado nos dados. Dependendo do método de pré-processamento utilizado, é possível que uma quantidade significativa de termos seja removida da base, resultando em um conjunto de regras com suporte mais alto.

Durante o trabalho foi considerado o uso de técnicas mais avançadas para o pré-processamento das bases, como o uso da WordNet ou Word embedding. WordNet é uma base de palavras que possui as relações semânticas entre as palavras, como sinônimos.

Ela é usada para remover stop words e também para converter palavras ao seu radical. Word embedding é a representação de palavras de forma vetorial, possibilitando que palavras semelhantes sejam identificadas, visto que a representação vetorial das mesmas seria parecida.

Porém, devido ao tempo disponível para realizar o trabalho foi decidido que as técnicas descritas na Seção 2.3 seriam utilizadas. Logo, seria interessante reproduzir esse experimento outra vez no futuro, utilizando técnicas de pré-processamento mais poderosas e que possibilitem resultados finais com maior grau de precisão.

Estas questões específicas foram abordadas de maneira mais aprofundada nos outros dois estudos de caso discutidos nas próximas seções, nos quais foram adotadas abordagens diferentes para a *query* de busca. Estas abordagens podem ter resultado em regras com suporte mais elevado e uma melhor identificação de padrões e relações entre os termos.

Este primeiro resultado mostrou que os modelos de carro anunciados para venda mudam todo dia. Logo a empresa BMW pode analisar essa informação para entender quais modelos são mais vendidos no mercado de carros usados. Esse resultado abrange um período de 4 dias, analisando um período maior é possível que informações mais relevantes sejam extraídas.

Em um segundo momento avaliamos 942 tweets da entidade BMW, que pertence a RepLab 2013, sem especificar nenhum tópico. As regras de implicação encontradas estão dentro de um intervalo de 5 dias, entre 01/06/2012 e 05/06/2012, e são mostradas na Tabela 13.

Tabela 13 – Regras de implicação da entidade BMW.

Day	Antecedent	Consequence	Support	Confidence
1	BMW	Audi	0.46%	1.92%
1	BMW	Buy	1.39%	5.76%
1	BMW	For Sale	0.46%	1.92%
2	BMW	Audi	0.93%	3.27%
2	BMW	For Sale	0.46%	1.63%
2	BMW	Buy	0.93%	3.27%
3	BMW	Audi	0.46%	3.99%
4	BMW	Audi, Mercedes	0.46%	4.34%
4	BMW	Want	0.93%	8.69%
5	BMW	Audi	3.72%	22.22%
5	BMW	Mercedes	2.79%	16.66%

As regras de implicação confirmam que o tópico “BMW vehicles for sale” é relevante até mesmo quando se avalia a entidade como um todo. Outro padrão observado é que

durante os 5 dias os usuários do Twitter incluíram as palavras BMW e Audi em seus *tweets*, porém, no dia 4 e 5 a palavra Mercedes também foi incluída nesses *tweets*. Com essa informação a empresa BMW poderia investigar para entender porquê os usuários do Twitter relacionaram essas marcas de carro em seus *tweets*.

Na última etapa 3897 *tweets* foram analisados, sendo que esses *tweets* foram obtidos através da API do Twitter. A motivação para essa coleta é analisar *tweets* recentes, visto que os *tweets* da RepLab 2013 são do ano de 2012. Outro fator que motivou a coleta é obter uma quantidade maior de *tweets* do que a disponibilizada pela RepLab 2013. As regras de implicação encontradas são de *tweets* coletados durante 5 dias e o resultado está na Tabela 14.

Esses *tweets* foram obtidos sem utilizar o filtro da API do Twitter que retorna apenas *tweets* que tiveram amplo alcance dentro da rede social. Por esse motivo, muitos *tweets* gerados por *bots* foram obtidos, ou seja, não é um *tweet* que expressa o sentimento ou opinião de uma pessoa real que está utilizando a rede social. Com isso, foi possível perceber a importância de utilizar o filtro da API do Twitter para gerar bases de *tweets*.

Tabela 14 – Regras de implicação da entidade BMW de tweets extraídos pela API do Twitter.

Day	Antecedent	Consequence	Support	Confidence
1	Used BMW	Pay Online	0.36%	4.99%
2	BMW M4 CSL	Passion and Confidence	1.09%	8.33%
4	Used BMW	Pay Online	0.36%	16.66%
5	Used BMW	Pay Online	0.36%	50%

Os resultados mostram que o tópico “Used BMW” se relaciona com o tópico “Pay Online”, o que leva a conclusão que pagamentos online estão sendo utilizados no mercado de carros semi-novos. Outra relação é que o modelo BMW M4 CSL se relaciona com “Passion and Confidence”, o slogan de um evento de eSports que a BMW patrocina. Essa informação é relevante para a empresa BMW por mostrar qual modelo de carro está sendo observado pelos usuários que acompanham o evento de eSports, revelando se o objetivo do *marketing* da empresa foi alcançado.

5.2 Vacina COVID

A Tabela 15 mostra os resultados obtidos analisando os tweets que contém o tópico “vacinas”. O dia 1 representa o dia 5 de janeiro de 2021.

Durante a análise dos tweets relacionados às vacinas no Brasil, constatou-se que o

Tabela 15 – Regras de implicação sobre vacinas.

Day	Antecedent	Consequence	Support	Confidence
1	Vacina	Bolsonaro	1%	50%
2	Vacina	Bolsonaro	4%	41%
2	Seringas	Bolsonaro	2%	75%
3	Vacina	Eficácia	4%	41%
7	Vacina	China	6%	40%
8	Vacina	Eficácia	5%	50%
8	Vacina	Bolsonaro	1%	16%
9	Vacina	Eficácia	3%	57%
9	Vacina	Bolsonaro	1%	28%
10	Vacina	Manaus	3%	40%
10	Vacina	Eficácia	1%	20%
11	Vacina	Oxigênio	6%	38%
11	Vacina	Bolsonaro	6%	38%
13	Vacina	Bolsonaro	1%	14%

presidente Jair Bolsonaro foi mencionado quase que diariamente. Essa presença frequente é resultado da intensa campanha liderada pelo presidente para desacreditar a eficácia das vacinas contra a COVID-19 e desencorajar os brasileiros a se vacinarem.

Ao longo da pandemia, Bolsonaro tinha feito declarações controversas, sugerindo que as vacinas podem causar efeitos colaterais graves e afirmando que os brasileiros que já foram infectados com o vírus são naturalmente imunes e, portanto, não precisariam ser vacinados. Além disso, o presidente tinha sido criticado por sua postura em relação à compra de vacinas, que muitas vezes parece hesitante ou incoerente, gerando incertezas na população e prejudicando a luta contra a pandemia.

No segundo dia de análise dos tweets relacionados à vacinação no Brasil, “Bolsonaro” e “seringas” foram um dos tópicos discutidos na plataforma. O assunto em questão era a escassez de seringas disponíveis para iniciar o processo de vacinação contra a COVID-19 no país. Usuários do Twitter mencionaram a falta de ação do governo brasileiro para garantir um número suficiente de seringas, e muitos apontaram a responsabilidade direta de Bolsonaro por essa situação. A implicação do presidente no problema das seringas indica uma falha na gestão da pandemia pelo governo e revela como os assuntos discutidos no Twitter são voláteis, visto que esse assunto só é discutido uma vez, com novos temas surgindo a cada dia em resposta às últimas notícias e desenvolvimentos relacionados à vacinação. Logo, o que ocorreu foi a queda do suporte dessa regra, abrindo espaço para que o suporte de outras regras subissem nos dias seguintes.

A métrica de confiança, que alcançou o valor mais alto de 75%, confirma a forte relação entre a escassez de seringas no Brasil e o presidente Jair Bolsonaro. Essa confiança

significativa indica que três em cada quatro tweets que discutiam o tema das seringas, o nome do presidente também era mencionado. Isso evidencia a percepção dos usuários da rede social de que o presidente Bolsonaro desempenhou um papel importante na questão da falta de seringas.

A eficácia das vacinas foi discutida ao longo dos dias analisados, sendo que essa regra está presente em 4 dias não contínuos. Uma explicação para a discussão sobre a eficácia das vacinas por um período maior de tempo é o fato da eficácia da CoronaVac, primeira vacina a ser aplicada no Brasil, ser baixa, o que gerou dúvida por grande parte da população. Foi necessário um trabalho de conscientização para explicar que o mais importante era a redução da necessidade de internação para pacientes já imunizados.

Nos dias 10 e 11 houve relação das vacinas com os tópicos Manaus e Oxigênio. Nesses dias ocorreu uma falta de oxigênio nos hospitais da cidade de Manaus, e como o oxigênio é um item essencial no tratamento de pacientes com COVID-19 justifica a relação do tópico com as vacinas, visto que as vacinas evitam que pacientes tenham que ser internados em hospitais. No dia 11 o tópico Bolsonaro também está em uma regra de implicação, pelo fato do governo federal ter sido acusado de não auxiliar com o envio de mais oxigênio para a cidade de Manaus. Interessante observar como um acontecimento regional, envolvendo apenas uma cidade brasileira, foi amplamente discutido a nível nacional na rede social Twitter, mostrando como ocorre uma integração entre usuários de várias localidades na rede social.

No dia 11, é relevante destacar que houve uma quantidade significativa de tweets que mencionavam os termos "vacina", "oxigênio"e "Bolsonaro". O suporte das duas regras combinadas atingiu 12%, o que representa o maior suporte em um único dia entre todas as regras avaliadas. Esse número demonstra claramente como os usuários da rede social estavam intensamente envolvidos nas discussões relacionadas à crise em Manaus e à possível influência do presidente Bolsonaro na falta de vacinas no Brasil.

Esses resultados mostram que nossa análise reflete com os acontecimentos e notícias sobre vacinas no Brasil em janeiro de 2021 e trazem reflexões sobre esses acontecimentos, como o exemplo da discussão sobre a eficácia da vacina ter se alongado durante vários dias na rede social Twitter. Logo, o uso dessa técnica analisando outros temas e por longos períodos de tempos também pode trazer bons resultados aos interessados.

5.3 Eleição Presidencial no Brasil

Para analisar a eleição presidencial foram coletados 3634 *tweets* entre o período de 23/07/2022 e 08/09/2022. Os *tweets* foram obtidos através da API do Twitter usando as palavras Lula, Bolsonaro e Eleições. Foi utilizado o filtro da API do Twitter que retorna

apenas *tweets* populares, e por esse motivo o número de *tweets* foi reduzido, porém são *tweets* que geraram grande engajamento na rede social.

Após aplicar as técnicas de PLN e selecionar os N-Gram que possuem significado, foram obtidos os seguintes atributos para o conceito formal: Alexandre de Moraes, bem contra mal, eleições, Bolsonaro, Lula, democracia, pesquisa eleitoral, senador Rogério Carvalho, orçamento secreto, Bolsonaro no flow, forças armadas, *fake news*, primeiro turno, Guilherme de Pádua, genocida, ex-presidiário, Jornal Nacional, corrupção, voto, presidente, entrevista, dinheiro vivo, imóveis comprados, piso salarial enfermagem, Fachin, suspende decreto armas.

A condição do contexto formal criado é o período de tempo em que o *tweet* foi publicado. O período de tempo utilizado foram 4 dias, gerando ao final 12 condições. Como exemplo, a Tabela 16 mostra uma amostra do contexto formal gerado.

Tabela 16 – Exemplo de parte do contexto gerado

	23/07 - 26/07				**27/07 - 30/07**			
ID	Lula	Bolsonaro	Genocida	Ex-presidiário	Lula	Bolsonaro	Genocida	Ex-presidiário
1	**X**			**X**				
2		**X**	**X**					
3					**X**			**X**
4						**X**	**X**	

Com o contexto formal obtido foi possível gerar as regras de implicação e analisar como as regras geradas refletem os acontecimentos da eleição presidencial no Brasil. A Tabela 17 detalha as regras geradas.

A primeira regra de implicação a ser discutida neste estudo é composta pelos termos “Eleições” e “Fake news”. Essa regra foi observada durante um período crucial que coincidiu com a divulgação de possíveis punições aos candidatos que espalhassem notícias falsas durante o período eleitoral, entre os dias 08/08 e 15/08. Durante esse período, também houve um aumento significativo na averiguação da eficácia das redes sociais em detectar e remover notícias falsas, visando evitar que os usuários fossem desinformados. Acredita-se que esses acontecimentos foram refletidos nas discussões do Twitter, o que justifica sua identificação em nosso estudo.

Além disso, é importante destacar que a disseminação de notícias falsas pode ter graves consequências para a democracia, como a interferência no processo eleitoral e a manipulação da opinião pública. Por essa razão, a luta contra as *fake news* se tornou uma preocupação global e a análise desses dados pode contribuir para a compreensão e o combate a esse fenômeno crescente.

Ao analisar a métrica de confiança entre as regras que resultam em fake news, com os antecedentes "Eleições"e "Bolsonaro", é possível observar uma tendência interessante.

Tabela 17 – Regras de implicação que variaram ao longo do tempo

Período de tempo	Antecedente	Consequência	Suporte	Confiança
08/08 até 11/08	Eleições	Fake news	4%	13%
12/08 até 15/08	Eleições	Fake news	2%	16%
08/08 até 11/08	Lula	Primeiro turno	2%	16%
16/08 até 19/08	Lula	Primeiro turno	8%	26%
16/08 até 19/08	Bolsonaro	Fake news	4%	33%
01/09 até 04/09	Bolsonaro	Fake news	3%	22%
20/08 até 23/08	Lula	Ex-presidiário	4%	25%
28/08 até 31/08	Lula	Ex-presidiário	2%	14%
24/08 até 27/08	Lula	Entrevista	2%	25%
28/08 até 31/08	Lula	Entrevista	2%	7%
28/08 até 31/08	Bolsonaro	Dinheiro vivo, Imóveis comprados	2%	10%
01/09 até 04/09	Bolsonaro	Dinheiro vivo, Imóveis comprados	3%	22%

Nota-se que a regra com o antecedente "Bolsonaro"possui uma confiança maior em comparação à regra com o antecedente "Eleições". Essa diferença de confiança sugere que o nome do candidato à presidência, Jair Bolsonaro, estava mais fortemente associado a notícias falsas do que o próprio contexto eleitoral em si. Isso evidencia que os usuários da rede social tinham uma percepção marcante de que o candidato Bolsonaro estava envolvido em disseminação de informações enganosas. Essa associação entre o nome de um candidato político e a propagação de fake news pode ter influenciado a percepção pública e gerado debates e discussões acaloradas nas mídias sociais durante o período das eleições.

Na Tabela 17, a segunda regra de implicação identificada é composta pelos termos “Lula” e “Primeiro turno”. A análise temporal revela que essa regra ocorreu em duas fases distintas, a primeira entre 08/08 e 11/08 e a segunda entre 16/08 e 19/08. Na época, a possibilidade de Lula vencer a eleição no primeiro turno era uma pauta recorrente nos meios de comunicação, uma vez que as pesquisas eleitorais apontavam que ele estava próximo de alcançar 50% dos votos válidos.

É interessante notar como o suporte dessa regra aumentou significativamente, passando de 2% para 8%, refletindo o crescente apoio à candidatura de Lula e o interesse do público nesse cenário eleitoral. Vale destacar que o aumento do suporte está diretamente relacionado à evolução das pesquisas eleitorais, que mostravam a crescente preferência dos eleitores pelo ex-presidente. Em 18 de agosto, uma pesquisa apontou que Lula atingiu 51% dos votos válidos, o que impulsionou ainda mais a discussão sobre a possibilidade de sua vitória no primeiro turno.

É possível observar, portanto, como a métrica de suporte pode ser útil para avaliar

a evolução das discussões na rede social, acompanhando a ascensão de temas que se tornam cada vez mais relevantes para os usuários. Essa análise é fundamental para compreender o impacto da opinião pública nas eleições e na construção da imagem dos candidatos.

A terceira regra de implicação identificada na Tabela 17 é composta pelos termos "Bolsonaro" e "Fake news". Interessante notar que essa regra apareceu em períodos de tempo dispersos, primeiro entre 16/08 e 19/08 e depois entre 01/09 e 04/09. Essa oscilação na frequência de ocorrência dessa regra pode ser explicada por dois eventos relevantes relacionados a Bolsonaro e às notícias falsas.

O primeiro evento ocorreu em meados de agosto, quando veio a público a notícia de que um grupo de empresários estaria orquestrando um golpe de estado em favor do presidente Bolsonaro. Esse fato gerou grande repercussão na mídia e nas redes sociais, com Bolsonaro tratando a notícia como uma *fake news*, o que pode ter influenciado a ocorrência dessa regra na primeira fase analisada.

O segundo evento que pode ter impulsionado a ocorrência dessa regra ocorreu no início de setembro, quando o Tribunal Superior Eleitoral (TSE) multou o presidente Bolsonaro por ter divulgado uma notícia falsa que vinculava Lula ao Primeiro Comando da Capital (PCC). Essa notícia falsa circulou amplamente nas redes sociais, gerando debates e discussões sobre a disseminação de informações inverídicas e seu impacto nas eleições. Esse fato pode ter contribuído para a ocorrência da regra na segunda fase analisada.

Essa regra de implicação mostra como um determinado tema pode surgir na rede social, ser discutido por um período e depois desaparecer temporariamente, para ressurgir novamente em outro momento. Isso evidencia a volatilidade das redes sociais e a dinamicidade das discussões que nelas ocorrem. A compreensão dessas dinâmicas é fundamental para a análise da opinião pública e das estratégias de campanha eleitoral.

A quarta regra de implicação identificada na Tabela 17 é particularmente interessante, já que revela a forte polarização política do Brasil. Essa regra é composta pelos termos "Lula" e "Ex-presidiário", e foi vista logo após um debate televisionado entre os presidenciáveis em que Jair Bolsonaro chamou o candidato Lula de "ex-presidiário".

Esse ato provocou uma discussão acalorada nas redes sociais sobre se Lula poderia ser chamado de inocente após os fatos que desmoralizaram a Operação Lava Jato, a maior investigação sobre corrupção na história do país. Enquanto os apoiadores de Bolsonaro aplaudiram a provocação, os partidários de Lula reagiram com indignação e acusaram o atual presidente de tentar denegrir a imagem do ex-presidente e líder do Partido dos Trabalhadores.

A regra de implicação "Lula" e "Ex-presidiário" ilustra como as redes sociais podem ser usadas como ferramentas para a propagação de discursos políticos e como as

divergências políticas podem levar a discussões polarizadas e acaloradas na internet.

É interessante notar que a quarta regra de implicação identificada na Tabela 17 apresentou uma confiança de 25% entre 20 e 23 de agosto, o que a torna a segunda maior confiança encontrada na análise. Isso destaca o forte vínculo entre o nome do candidato Lula e o termo “ex-presidiário”, sugerindo que a estratégia de Jair Bolsonaro foi efetiva em vincular Lula ao seu passado como condenado pela justiça.

Esse resultado indica como a retórica política pode influenciar as discussões nas redes sociais e como a polarização pode levar a uma divulgação ampla de informações tendenciosas. Além disso, essa regra de implicação destaca a importância da análise de sentimentos e opiniões nas mídias sociais para entender como a retórica política pode influenciar a opinião pública e moldar o debate político em torno de determinados temas e figuras públicas.

A quinta regra de implicação identificada na Tabela 17 é composta pelos termos “Lula” e “Entrevista”. Esse último termo refere-se à entrevista que o Jornal Nacional da TV Globo realizou com todos os presidenciáveis. É interessante observar que essa entrevista foi realizada no dia 25 de agosto e, ainda assim, o assunto continuou sendo discutido no Twitter até o dia 31 de agosto. Esse fato demonstra a relevância da entrevista para os usuários da plataforma, que se engajaram em discussões sobre o desempenho dos candidatos naquele evento.

No caso específico do termo "Lula", é possível inferir que a participação do candidato na entrevista gerou um interesse ainda maior por parte dos usuários, que discutiram seu desempenho e as ideias apresentadas. Isso destaca a importância da mídia tradicional na formação de opinião pública e como as plataformas de mídia social podem amplificar e prolongar o impacto desses eventos na sociedade. Além disso, a persistência da discussão sobre a entrevista no Twitter também mostra como as redes sociais podem ser usadas para monitorar e analisar o engajamento do público em relação a eventos políticos importantes.

É notável como a métrica de confiança dessa regra específica diminuiu significativamente no segundo período de tempo analisado. Essa queda na confiança indica que, após alguns dias da entrevista mencionada, novos termos e assuntos relacionados ao candidato Lula passaram a ganhar relevância na discussão. Isso destaca a volatilidade da rede social, onde os temas e tópicos discutidos podem mudar rapidamente em um curto período de tempo.

Essa dinâmica ágil da rede social reflete a natureza efêmera das conversas e a rápida disseminação de informações. Em poucos dias, outros eventos, declarações ou acontecimentos mais recentes podem capturar a atenção dos usuários e deslocar o foco para assuntos mais recentes. Essa observação reforça a importância de monitorar continuamente as discussões nas mídias sociais para obter uma compreensão atualizada do

panorama de opiniões e tópicos de interesse.

A sexta regra de implicação identificada na Tabela 17 é muito relevante para a compreensão do contexto político brasileiro. Ela é composta pelos termos “Bolsonaro”, “Dinheiro vivo” e “Imóveis comprados” e está relacionada com as matérias publicadas pela mídia brasileira que evidenciaram que muitos dos imóveis comprados pela família Bolsonaro foram adquiridos em dinheiro vivo.

Esse fato gerou questionamentos por parte dos adversários de Jair Bolsonaro, que buscaram entender o motivo da compra utilizando dinheiro em espécie e a origem desse dinheiro. A implicação desses termos mostra como as questões relacionadas à ética na política são relevantes para os usuários do Twitter, que buscam debater e entender as implicações dessas ações por parte dos políticos.

É possível notar como a análise dos tópicos discutidos no Twitter pode fornecer informações valiosas para as campanhas dos presidenciáveis. Ao compreender quais tópicos estão gerando mais interesse e discussão entre os usuários da rede social, as campanhas podem ajustar suas estratégias de comunicação e atuar de forma mais efetiva para conquistar o eleitorado. Além disso, a análise também pode ajudar as campanhas a identificar pontos fracos em suas estratégias e aprimorá-las.

Em resumo, a análise de tópicos no Twitter pode ser uma ferramenta importante para as campanhas eleitorais se conectarem com o eleitorado e compreenderem melhor as suas preocupações e interesses.

6 CONCLUSÕES E TRABALHOS FUTUROS

Nesta dissertação apresenta-se uma abordagem para analisar tópicos discutidos na rede social Twitter. Para identificar esses tópicos utiliza-se a API do Twitter para extrair os tweets e PLN para encontrar os tópicos dentro dos tweets. Finalmente, a AFC foi utilizada para gerar regras de implicação entre esses tópicos, fornecendo as métricas de suporte e confiança.

Com as regras de implicação e métricas geradas foi possível comparar como os usuários do Twitter eram afetados por acontecimentos que estavam acontecendo no momento, como a vacinação contra a COVID-19 e a eleição presidencial brasileira.

Além de avaliar essa abordagem com tweets extraídos através da API também foi usada a base RepLab2013, que consiste em tweets que abordam sobre automóveis, entidades financeiras, universidades e música/artistas. Esses tweets são avaliados por especialistas e classificados manualmente entre essas categorias. Neste estudo avalia-se os tweets que discutiam sobre a marca de automóveis BMW, e com isso foi possível observar os modelos que estavam sendo anunciados à venda no Twitter e opiniões gerais do público sobre a marca.

Já com os tweets extraídos através da API do Twitter foi avaliado como os usuários da rede social estavam discutindo sobre a vacinação contra a COVID-19. Observa-se como o presidente do Brasil na época se relacionava fortemente com as discussões sobre as vacinas, e também como a crise de falta de oxigênio em Manaus se relacionou com o termo vacina, visto que as vacinas poderiam ter evitado tal acontecimento. Portanto, foi possível averiguar que as notícias que envolviam as vacinas eram discutidas na rede social Twitter de forma volátil, com os tópicos variando dia-a-dia.

Finalmente foi avaliada a eleição presidencial brasileira, com foco nos dois candidatos mais relevantes, Lula e Bolsonaro. A análise foi feita durante 1 mês e meio, possibilitando que tópicos que estavam sendo discutidos fossem esquecidos e depois relembrados pelos usuários da rede social, como foi o caso dos tópicos “Bolsonaro” e “Fake news”.

Os resultados mostram que é possível avaliar como os usuários estão reagindo aos acontecimentos relacionados a eleição, informação que é relevante para as campanhas dos

presidenciáveis, que precisam de métricas para avaliar o impacto das suas ações.

Como um trabalho futuro planeja-se a automatização de todo o processo, para que os dados sejam extraídos da rede social e analisados de forma automatizada, gerando as métricas sobre os tópicos discutidos no mesmo momento. Isso é relevante para que essas informações sejam analisadas no mesmo instante de tempo que elas estão sendo discutidas na rede social.

Outro trabalho futuro seria criar uma interface gráfica para a ferramenta, para que essa análise possa ser iniciada por um usuário comum, definindo os parâmetros dos dados que devem ser buscados na rede social e avaliados pela ferramenta em seguida. Com isso, seria viável que esse método pudesse ser usado de forma comercial, por empresas e órgãos que queiram avaliar um tópico específico na rede social Twitter.

Como outro trabalho futuro nós planejamos tornar genérica a escolha de qual rede social o usuário buscaria os dados a serem analisados. Para isso, seria necessário construir funções que comuniquem com as APIs das diversas redes sociais que existem atualmente. Essa abordagem seria interessante visto que cada rede social possui suas peculiaridades, possibilitando que o resultado final da análise seja diferente para cada rede social.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 487–499. ISBN 1-55860-153-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645920.672836>.

ALI, I.; LI, Y.; PEDRYCZ, W. Granular computing approach to evaluate spatio-temporal events in intuitionistic fuzzy sets data through formal concept analysis. Axioms, v. 12, n. 5, 2023. ISSN 2075-1680. Disponível em: <https://www.mdpi.com/2075-1680/12/5/407>.

AMIGO, E. et al. Overview of replab 2013: Evaluating online reputation monitoring systems. In: CEUR Workshop Proceedings. [S.l.: s.n.], 2013. v. 1179. ISBN 978-3-642-40801-4.

BANERJEE, S.; PEDERSEN, T. The design, implementation, and use of the ngram statistics package. In: GELBUKH, A. (Ed.). Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 370–381. ISBN 978-3-540-36456-6.

BELOHLAVEK, R.; OSICKA, P. Triadic concept analysis of data with fuzzy attributes. In: 2010 IEEE International Conference on Granular Computing. [S.l.: s.n.], 2010. p. 661–665.

BIEDERMANN, K. How triadic diagrams represent conceptual structures. In: LUKOSE, D. et al. (Ed.). Conceptual Structures: Fulfilling Peirce's Dream. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p. 304–317. ISBN 978-3-540-69424-3.

CASTELLANOS, A.; CIGARRAN, J.; GARCíA-SERRANO, A. Formal concept analysis for topic detection: A clustering quality experimental analysis. Information Systems, v. 66, p. 24–42, 2017. ISSN 0306-4379. Disponível em: <https://www.sciencedirect.com/science/article/pii/S030643791730087X>.

CATALDI, M.; CARO, L. D.; SCHIFANELLA, C. Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining. New York, NY, USA: Association for Computing Machinery, 2010. (MDMKDD '10). ISBN 9781450302203. Disponível em: <https://doi.org/10.1145/1814245.1814249>.

CIGARRAN, J.; CASTELLANOS Ángel; GARCíA-SERRANO, A. A step forward for topic detection in twitter: An fca-based approach. Expert Systems with Applications, v. 57, p. 21–36, 2016. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417416301038>.

CONTRERAS, J. O.; HILLES, S.; ABUBAKAR, Z. B. Automated essay scoring with ontology based on text mining and nltk tools. In: 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). [S.l.: s.n.], 2018. p. 1–6.

GANTER, B.; OBIEDKOV, S. Implications in triadic formal contexts. In: WOLFF, K. E.; PFEIFFER, H. D.; DELUGACH, H. S. (Ed.). Conceptual Structures at Work. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 186–195. ISBN 978-3-540-27769-9.

LEHMANN, F.; WILLE, R. A triadic approach to formal concept analysis. In: ELLIS, G. et al. (Ed.). Conceptual Structures: Applications, Implementation and Theory. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. p. 32–43. ISBN 978-3-540-49539-0.

MIAO, Z. et al. Cost-effective online trending topic detection and popularity prediction in microblogging. ACM Trans. Inf. Syst., Association for Computing Machinery, New York, NY, USA, v. 35, n. 3, dez. 2016. ISSN 1046-8188. Disponível em: <https://doi.org/10.1145/3001833>.

MISSAOUI, R.; KUZNETSOV, S. O.; OBIEDKOV, S. Formal Concept Analysis of Social Networks. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2017. ISBN 3319641662.

MURSHED, B. A. H. et al. Dea-rnn: A hybrid deep learning approach for cyberbullying detection in twitter social media platform. IEEE Access, v. 10, p. 25857–25871, 2022.

PORTER, M. An algorithm for suffix stripping. In: . [s.n.], 1980. v. 14, n. 3. Disponível em: <https://doi.org/10.1108/eb046814>.

RECUERO, J. Organización de resultados de búsqueda mediante análisis formal de conceptos. In: . [s.n.], 2008. Disponível em: <https://dialnet.unirioja.es/servlet/tesis?codigo=41174>.

REN, X.; LI, D.; ZHAI, Y. Research on mixed decision implications based on formal concept analysis. International Journal of Cognitive Computing in Engineering, v. 4, p. 71–77, 2023. ISSN 2666-3074. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666307423000104>.

SM., D.; C., S.; DF. Şotropa. An investigation of user behavior in educational platforms using temporal concept analysis. In: International Conference on Formal Concept Analysis 2017. [S.l.: s.n.], 2017.

WILLE, R. Restructuring lattice theory: An approach based on hierarchies of concepts. In: RIVAL, I. (Ed.). Ordered Sets. Dordrecht: Springer Netherlands, 1982. p. 445–470. ISBN 978-94-009-7798-3.

WILLETT, P. The porter stemming algorithm: then and now. In: Program electronic library and information systems. [S.l.: s.n.], 2006. v. 40.

YOGISH, D.; MANJUNATH, T. N.; HEGADI, R. S. Review on natural language processing trends and techniques using nltk. In: SANTOSH, K. C.; HEGADI, R. S.

(Ed.). Recent Trends in Image Processing and Pattern Recognition. Singapore: Springer Singapore, 2019. p. 589–606. ISBN 978-981-13-9187-3.

ZENG, J. et al. Semantic multi-grain mixture topic model for text analysis. Expert Syst. Appl., Pergamon Press, Inc., USA, v. 38, n. 4, p. 3574–3579, abr. 2011. ISSN 0957-4174. Disponível em: <https://doi.org/10.1016/j.eswa.2010.08.146>.

ZHANG, J. et al. Detecting topic labels for tweets by matching features from pseudo-relevance feedback. In: Proceedings of the Tenth Australasian Data Mining Conference - Volume 134. AUS: Australian Computer Society, Inc., 2012. (AusDM '12), p. 9–19. ISBN 9781921770142.