Caracterização Sócio-Temporal de Conteúdos em Redes Sociais baseada em Processamento em Fluxo*

Nicollas R. de Oliveira[1], Dianne S. V. Medeiros[1], Diogo M. F. Mattos[1]

[1]MídiaCom/TET/PPGEET/UFF
Universidade Federal Fluminense (UFF)
Niterói, RJ – Brasil

Abstract. *The propagation dynamics and speed of subjects published on Twitter characterizes the social network platform as an uninterrupted data source. This paper proposes a distributed approach based on complex network metrics for the socio-temporal characterization of textual data from Twitter. The proposal integrates Apache Kafka for the data ingestion and Apache Spark for the data flow processing to ensure the continuous and efficient capture of content from different sources. The proposal identifies, correlates and monitors the use of hashtags in real-time through a dynamic graph structure, generating an ontology about the topic of interest. Unlike previous works, which use historical data, the proposal is applied to a real use case with great repercussion and engagement of Twitter users. By evaluating metric fluctuations such as centrality, diameter and density for multiple components of the generated graph, the results reveal writing trends and relationship patterns that reinforce the feeling of echo chambers and media opportunism in the logic of using hashtags.*

Resumo. *A velocidade e a dinâmica de propagação de assuntos veiculados no* Twitter *caracterizam a plataforma como uma fonte de dados ininterrupta. Este artigo propõe uma abordagem distribuída baseada em métricas de redes complexas para a caracterização socio-temporal de dados textuais provenientes do* Twitter. *A proposta integra o* Apache Kafka *na ingestão dos dados e o* Apache Spark Streaming *no processamento em fluxo dos dados para garantir a captura contínua e o processamento eficiente do conteúdo de diferentes fontes. A proposta identifica, correlaciona e monitora o uso de* hashtags *em tempo real, através de uma estrutura de grafo dinâmica, gerando uma ontologia sobre o tópico de interesse. Diferente de trabalhos anteriores, que empregam dados históricos, a proposta é aplicada a um caso de uso real com grande repercussão e engajamento dos usuários do* Twitter. *Avaliando as flutuações de métricas como centralidade, diâmetro e densidade para múltiplas componentes do grafo de hashtags, os resultados revelam tendências de escrita e padrões de relacionamento que reforçam a sensação de câmaras de eco e oportunismo midiático na lógica de utilização de* hashtags.

1. Introdução

Incorporadas nas postagens do *Twitter* em 2007, as *hashtags* são palavras, ou frases sem espaçamento, prefixadas com o caractere cardinal "#". Recentemente, as *hashtags* extrapolam sua função original de anotar, categorizar e contextualizar os *tweets*, postagens

*Os autores agradecem ao CNPq, CAPES, FAPERJ, FAPESP (2018/23062-5), RNP e Prefeitura Municipal de Niterói (Edital PDPA PMN/UFF/FEC) pelo financiamento que viabilizou essa pesquisa.

no *Twitter*, e se tornaram uma funcionalidade de indexação multiplataforma usada como termômetro de eventos sociopolíticos, comerciais e culturais. Portanto, a compreensão dos relacionamentos entre *hashtags* revela padrões subjacentes de comunicação de estilo moderno [Zhang, 2019, Yang et al., 2020] e auxilia para o gerenciamento de redes sociais.

Ao considerar que *hashtags*, assim como qualquer outra manifestação textual, expressam os interesses, opiniões e crenças dos seus autores, pode-se assumir a hipótese de que estão igualmente sujeitas à influência de terceiros. Contudo, devido à natureza inerentemente dinâmica e da ausência de padronização no uso, *hashtags* também são susceptíveis a alterações de popularidade e significado em curtos intervalos. Essas variações impõem às abordagens de monitoramento de redes sociais requisitos de responsividade e adaptabilidade equiparáveis aos de aplicações em tempo real [Stilo e Velardi, 2017].

Este artigo propõe uma abordagem distribuída para monitoramento e caracterização temporal do conteúdo relacionado a eventos reais repercutidos em redes sociais. Focando no *Twitter*, a abordagem proposta aplica uma lógica capaz de realizar a captura ininterrupta e dinâmica de *tweets* relacionados ao evento monitorado. Assim, alimenta-se incrementalmente um grafo relacional de *hashtags* extraídas do fluxo de *tweets*, no qual *hashtags* são representadas por nós e uma aresta existe entre dois nós adjacentes quando as *hashtags* são usadas em um mesmo *tweet*. A dinamicidade da coleta é provida pela atualização periódica das *hashtags* que orientam a captura de *tweets*, através de métricas de redes complexas obtidas do grafo relacional. Ao utilizar como caso de uso um evento de grande repercussão e alto engajamento popular em redes sociais, tal como o engajamento provocado pelo programa de televisão *Big Brother Brasil 2021* (*BBB21*), é possível analisar perfis de compartilhamento de *hashtags* característicos de câmaras de eco e de cunho oportunista através das flutuações de métricas como centralidade, diâmetro e densidade para múltiplas componentes do grafo de relacionamento entre *hashtags*.

As *hashtags* são frequentemente exploradas em diversas pesquisas, como detecção [Chen et al., 2018] e monitoramento de eventos [Giridhar et al., 2017], predição de popularidade [Huang et al., 2020], e sistemas de recomendação [Alsini et al., 2020, Yang et al., 2020] incorporação de semântica em vetores [Liu et al., 2018] ou modelagem de tópicos [Wang et al., 2016]. Outros trabalhos concentram-se na análise e caracterização de *hashtags* em torno de temas específicos [Cossard et al., 2020] ou empregam plataformas sociais não usuais [Yang et al., 2020]. Diferentemente dos estudos anteriores, a abordagem incremental proposta neste artigo caracteriza eventos em tempo real, garantindo que quaisquer tendências tipográficas, padrões de conexão e disjunção de domínios do conhecimento sejam detectados.

O restante do artigo está organizado da seguinte forma. A Seção 2 discute o problema das câmaras de eco em redes sociais. A Seção 3 embasa o processamento em fluxo distribuído. A Seção 4 apresenta a proposta da abordagem incremental e descreve a arquitetura e a operação. A Seção 5 discute os resultados, e a Seção 6 examina os trabalhos relacionados. A Seção 7 conclui o artigo.

2. As Câmaras de Eco e o Uso de *Hashtags*

As mídias sociais são um importante meio de interação pública e são capazes tanto de hospedar conteúdos potencialmente úteis, quanto de serem um reduto de discurso de ódio e notícias falsas [de Oliveira et al., 2021]. No entanto, a ampla divulgação desses conteúdos

promove uma significativa mudança de paradigma na criação e consumo de informação. O processo de seleção de informações, tradicionalmente mediado por jornalistas ou editores, atualmente inclui outros atores, os usuários. Tal desintermediação degrada a imparcialidade do processo de seleção, uma vez que os usuários tendem a absorver e compartilhar informações que se aderem ao seu sistema de crenças, fenômeno conhecido como viés de confirmação [Vicario et al., 2016, de Oliveira et al., 2021]. Tão recorrente quanto o viés de confirmação, o efeito "câmara de eco" é um fenômeno social definido pela tendência dos usuários em interagir e ingressar em grupos homogêneos com ideias semelhantes às suas. Tal fenômeno pode ter raízes em vieses comportamentais, como a exposição seletiva, ou em vieses de algoritmo que restringem o acesso a fontes específicas de informação com base em perfis digitais [Cossard et al., 2020, Colleoni et al., 2014]. Assim, existe a relação de simbiose entre redes sociais e eventos do mundo real, como programas de televisão e *shows*, e, portanto, é fundamental (i) desenvolver soluções em tempo real capazes de caracterizar as câmaras de eco em torno das repercussões de eventos externos às redes sociais. Atrelado ao desafio da caracterização está (ii) a dificuldade de fazê-la dinamicamente, acompanhando a velocidade natural de geração e o volume de dados relacionados a eventos reais. Uma maneira de satisfazer esse imediatismo, sem recorrer a técnicas possivelmente mais demoradas e complexas de processamento de linguagem natural e aprendizado de máquina, é através do processamento das palavras-chave indexadoras das mensagens, as *hashtags*.

A menor complexidade ao processar *hashtags* ao invés de processar todo o texto dos *tweets* é acompanhada por três desafios principais. O primeiro é a polissemia, ou seja, uma mesma *hashtag* pode se referir a eventos distintos em diferentes janelas temporais. Como um segundo desafio ao lidar com *hashtags*, a sinonímia está relacionada ao fato de diferentes *hashtags* poderem possuir o mesmo significado. Tal característica está diretamente relacionada à arbitrariedade de sua criação, uma vez que que não há um consenso entre os usuários, ou orientação das plataformas, sobre a padronização da escrita das *hashtags*. O terceiro desafio é a obscuridade que remete à dificuldade de interpretação enfrentada tanto por humanos quanto por algoritmos [Stilo e Velardi, 2017]. Sendo frequentemente compostas por acrônimos, palavras concatenadas, neologismos, abreviações ou combinações das opções anteriores, as *hashtags* podem demorar a serem completamente compreendidas dependendo do conhecimento prévio do leitor sobre o contexto relacionado. Para mitigar esses desafios, a proposta emprega uma estrutura em grafo para prover um entendimento com base no contexto ou tema relacionado as *hashtags*, sem a necessidade de interpretação individual de cada uma. Uma vantagem adicional da estruturação de *hashtags* em grafos é a associação à ontologias para a representação do conhecimento relacionando termos, palavras, expressões ou axiomas pertencentes a um mesmo domínio de interesse [de Oliveira et al., 2020]. Dessa forma, garante-se o monitoramento de discussões *online*, sem comprometer a capacidade das *hashtags* de oferecer metadados valiosos sobre o texto associado de forma compactada.

3. O Processamento em Fluxo Distribuído

Tradicionalmente, há três principais abordagens adotadas no processamento de dados: em lotes (*batch*), em micro-lotes (*micro-batch*) e em fluxos (*streaming*). Quando lidando com grandes conjunto de dados estáticos, ou seja, previamente coletados e represados ao longo do tempo, é comum a adoção do processamento em lotes. Devido ao armaze-

namento e pré-processamento prévios, a abordagem de processamento em lotes tende a apresentar latências significativamente grandes, da ordem de segundos ou minutos. Tal fator inviabiliza sua utilização em aplicações em tempo real, que requerem respostas da ordem de sub-segundo. Para minimizar os efeitos da latência, esta abordagem evoluiu para um processamento baseado em micro-lotes, em que o fluxo é segmentado em pequenos blocos de dados em lotes. Pautado por intervalos de tempo pequenos, ou pelo tamanho em bytes, o fluxo de entrada é aglutinado em blocos curtos de dados e entregue ao sistema de processamento em lotes. Diferentemente das anteriores, a abordagem de processamento em fluxo de dados analisa uma sequência massiva de dados continuamente gerados e fornece resultados incrementais e dinâmicos, a menos que seja explicitamente encerrada. Por essa razão, essa abordagem apresenta duas principais vantagens: (i) a expressividade, uma vez que não há limitação por nenhuma abstração não natural e (ii) o tempo de execução reduzido, devido ao imediatismo do processamento. Em contrapartida, o rendimento mais baixo e o custo para tornar o sistema tolerante a falhas ou capaz de realizar um balanceamento de carga são desafios conhecidos dessa abordagem de processamento [Nasiri et al., 2019, Lopez et al., 2016a].

Sistemas de processamento em fluxo distribuído (DSPS) são caracterizados pelo emprego de uma abstração de fluxo de dados conhecida como grafo acíclico direcionado (GAD) de operadores. Tais operadores interconectados executam funções simples como contagem, filtragem, projeção e agregação, porém de maneira paralela garantindo que o processamento de fluxo ocorra com baixa latência e com alta vazão [Liu e Buyya, 2020].

3.1. Apache Kafka

O Kafka[1] é uma estrutura de ingestão de dados distribuída que coordena de forma confiável a transmissão de dados entre servidores e clientes usando TCP. Os processos de leitura e gravação de dados no Kafka ocorrem como eventos, por exemplo, um registro ou mensagem de documentação associada a três atributos: chave, valor e carimbo de data e hora. Para organizar e armazenar milhares de eventos de diferentes fontes, o Kafka apresenta quatro conceitos-chave: tópico, produtor, consumidor e *broker*. De forma análoga às pastas em um sistema de arquivos, os **tópicos** servem como contêineres de dados intermediários para eventos transmitidos entre aplicativos ou sistemas. Internamente, cada tópico pode ser subdividido em várias partições, permitindo assim um balanceamento de carga, uma recuperação mais rápida de informações e redundância de dados. Dentro da lógica de gerenciamento de eventos, é possível separar os clientes em **produtores**, responsáveis por registrar eventos em tópicos, e **consumidores**, cuja função é ler eventos de uma partição. Um ***broker*** Kafka consiste em um nó de um sistema distribuído que lida com os processos de leitura, escrita e balanceamento de carga [Yadranjiaghdam et al., 2017, Sun et al., 2019].

3.2. Apache Spark

O Apache Spark é uma plataforma de processamento distribuído de propósito geral que estende o modelo *MapReduce* do Hadoop para oferecer suporte de forma eficiente a mais tipos de cálculos, incluindo consultas interativas e processamento em fluxo. O Spark processa tarefas na memória para atingir esses objetivos, evitando gravar e ler resultados

[1]Disponível em https://kafka.apache.org/.

intermediários no disco. Além de combinar diferentes tipos de processamento, o Spark é projetado para ser rápido, altamente acessível e fornecer Interfaces de Programação de Aplicações (*Application Programming Interfaces* – APIs) simples em diferentes linguagens de programação, como Python, Java, Scala e SQL. A incorporação de ferramentas específicas, por exemplo para realização de processamento de dados em fluxo em linha ou manipulação de grafos, é feita na forma de bibliotecas no topo da plataforma [Rao et al., 2019, Lopez et al., 2016b].

4. A Abordagem Incremental Proposta para Caracterização de *Hashtags*

A abordagem incremental proposta visa revelar padrões latentes no comportamento dos usuários do *Twitter* ao compartilharem *hashtags* relacionadas a um evento monitorado. Usando um evento preeminente em tempo real como um caso de uso, a abordagem integra conceitos de processamento em fluxo distribuído e redes complexas para capturar, processar e, finalmente, analisar o conteúdo proveniente da rede social. Para tanto, é proposta uma arquitetura modular, conforme Figura 1. A arquitetura é adaptável a múltiplas fontes de dados diferentes, tendo todos os módulos desenvolvidos em linguagem Python.

O *Twitter* age como fonte de dados para o **Módulo de Captura**, que compreende a API de captura em fluxo (*streaming*) de dados da plataforma de rede social. Ao inicializar o módulo com *hashtags* de entrada, a aplicação garante a captura ininterrupta do fluxo de *tweets* relacionado à entrada. O **Módulo de Ingestão**, implementado usando Apache Kafka, garante que outras fontes de dados possam ser adicionadas ao esquema de monitoramento de fluxo. Tal módulo é essencial para evitar gargalos e manter a organização e entrega dos dados coletados. Posteriormente, cada *tweet* é introduzido no **Módulo de Processamento**, cujo núcleo é o Apache Spark, o arcabouço para processamento em fluxo. Nesse módulo, as *hashtags* são extraídas de *tweets* e incorporadas a um grafo relacional usando a biblioteca `GraphX` [2]. Para trazer dinamismo e realismo à manutenção incremental do grafo, as *hashtags* são atualizadas periodicamente de acordo com o critério de centralidade de proximidade. Por fim, o **Módulo de Caracterização** analisa temporalmente o grafo relacional formado ao final de cada ciclo de monitoramento.

4.1. A Coleta de Dados

Embora as *hashtags* estejam incorporadas em várias plataformas, como *Instagram*, *Facebook* e *LinkedIn*, seu uso como um indicador de evento é mais frequente nas postagens do *Twitter*. Para monitorar um evento em tempo real, o artigo foca no estudo de caso do episódio final do programa de televisão *Big Brother Brasil 2021*, a versão brasileira do famoso *reality show* Big Brother, exibido em 4 de Maio de 2021 pela Rede Globo de Televisão. O foco nesse estudo de caso baseia-se na repercussão do programa quanto à audiência relatada pelo IBOPE (Instituto Brasileiro de Opinião Pública e Estatística) [3] e o alto engajamento alavancado nas redes sociais, alcançando recordes de mensagens no *Twitter* [4]. Nesse contexto, para coletar *hashtags* relacionadas ao evento específico, emprega-se a API *Twitter Streaming*, que captura todos os *tweets* recebidos relacionados

[2]Disponível em https://spark.apache.org/docs/latest/graphx-programming-guide.html.

[3]Disponível em https://www.kantaribopemedia.com/dados-de-audiencia-nas-15-pracas-regulares-com-base-no-ranking-consolidado-26-04-a-02-05-2021/

[4]Disponível em https://www1.folha.uol.com.br/ilustrada/2021/02/bbb-21-e-marco-de-era-em-que-a-internet-domina-e-pauta-o-que-passa-na-tv.shtml

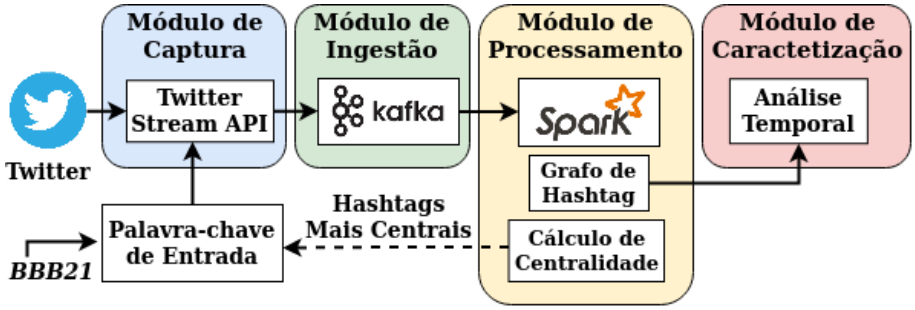


Figura 1. A arquitetura proposta é modular e realiza uma abordagem incremental para a caracterização temporal do conteúdo extraído do *Twitter*. A partir de uma palavra-chave de entrada inicial, os *tweets* relacionados são capturados e injetados no sistema. Após a extração, as *hashtags* são incorporadas a um grafo relacional, que é atualizado periodicamente usando as *hashtags* mais centrais desse grafo como novas palavras-chave de entrada. Finalmente, o grafo de *hashtags* é analisado temporalmente.

à *hashtag* de entrada, por exemplo, *BBB21*. Assim, após um tempo de coleta de 37 horas, é consolidada uma base de dados contendo mais de 1,5 milhão de *tweets*, cujas estatísticas estão expressas na Tabela 1.

4.2. O Grafo Relacional de *Hashtags*

Uma vez capturados e processados pelo módulo de ingestão, os *tweets* têm suas *hashtags* extraídas para compor incrementalmente um grafo não direcionado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. No grafo relacional de *hashtags*, cada nó representa uma *hashtag* distinta e a aresta entre um par de nós indica que as *hashtags* aparecem simultaneamente no mesmo *tweet*. Nós isolados representam quaisquer ocorrências únicas de *hashtags* em um *tweet*. Devido à abordagem incremental, na qual novas capturas são utilizadas para atualizar o grafo relacional, a estrutura cresce continuamente. Para manter o dinamismo e a conformidade com as mudanças, as *hashtags* de entrada são substituídas a cada intervalo de monitoramento de uma (1) hora pelas 20 *hashtags* mais centrais do grafo relacional atual. Tais valores foram usados como exemplo de aplicação. Esta seleção é determinada pelo cálculo da centralidade de proximidade (c_c) de cada nó (v_i), dada por

$$c_c(v_i) = \frac{\mathcal{V} - 1}{\sum_{j \neq i} \delta^*(v_i, v_j)}, \tag{1}$$

em que $\mathcal{V}$ é o número total de nós e $\delta^*(v_i, v_j)$ é a distância mais curta em número de saltos entre o par de nós v_i e v_j. Na prática, quando usada em um grafo de *hashtags*, a centralidade de proximidade mede efetivamente o quão próxima, em média, cada *hashtag* está de todas as outras *hashtags* em uma comunidade, ou seja, dentro de um grupo de *hashtags* semelhantes que estão densamente conectadas. Assim, a *hashtag* mais central é a mais representativa da comunidade.

5. A Avaliação e a Discussão da Proposta

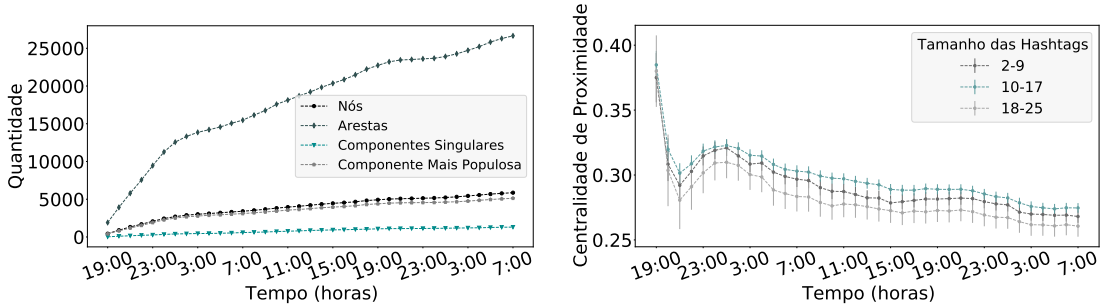
Na caracterização temporal e semântica conduzida neste artigo, são considerados aspectos como (i) o crescimento estrutural do grafo relacional de *hashtags*; (ii) os padrões

Tabela 1. Estatísticas da Captura de Dados Relacionados ao *BBB21*.

Intervalo Total de Captura	18:00 04/05/2021 - 07:00 06/05/2021
Quantidade Total de *Tweets*	1.737.753
Quantidade de *Tweets* Contendo *Hashtags*	443.547
Número de *Hashtags* Distintas	7175
Vazão Média	782 *tweets*/min
Vazão Máxima (Pico)	2998 *tweets*/min

de escrita das *hashtags*; (iii) métricas de redes complexas como centralidade, densidade, diâmetro e assortatividade; e (iv) a distância entre os significados de *hashtags* pertencentes a componente principal do grafo gerado. Vale ressaltar que os resultados apresentados são médias com intervalo de confiança de 95%.

Em uma primeira análise, a Figura 2(a) mostra o progresso temporal do grafo relacional em diferentes perspectivas, uma avaliando o número de arestas totais e outra considerando o número de nós no grafo e seus componentes (subgrafos). Quando comparadas, essas perspectivas divergem no padrão de crescimento, evidenciando que todas as curvas relacionadas a nós apresentam pouca variação. Além disso, constata-se que a componente principal concentra a maioria dos nós e, consequentemente, *hashtags* de todo o grafo relacional. Em contraste, a curva de arestas possui um comportamento ascendente ao longo da captura, demonstrando que usuários estão mais propensos a utilizar novas associações de *hashtags* previamente usadas do que propriamente criar novas. A existência de vários componentes de nó único no grafo gerado, marcada pela presença de nós isolados do componente mais populoso do grafo, pode revelar *hashtags* oportunistas. Tal fenômeno possivelmente ocorre quando o usuário, intencionalmente, ou não, posta *tweets* relacionados ao evento monitorado, mas criam ou usam *hashtags* específicas que não possuem adesão generalizada. Embora componentes desconectados sejam inerentes às primeiras horas de captura, sua permanência até o fim do período de captura reforça a suposição de que essas *hashtags* não pertencem ao domínio de conhecimento do evento, mas representam um comportamento de usuários oportunistas.



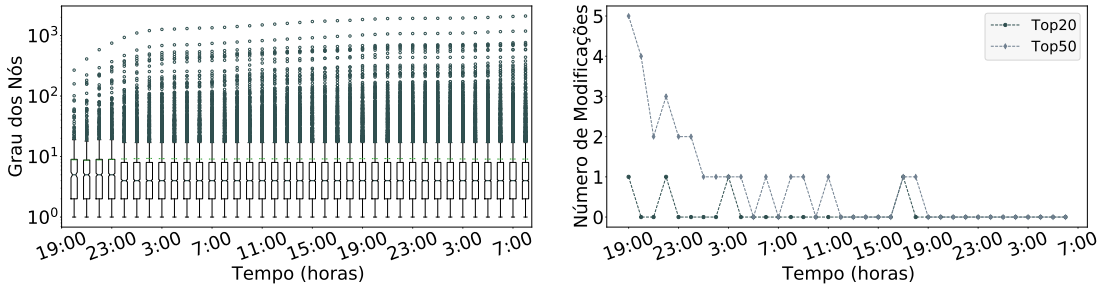
(a) Crescimento temporal do grafo relacional de *hashtags* quando considerando o número de nós e arestas.

(b) Variação temporal da centralidade de proximidade considerando *hashtags* com diferentes quantidades de caracteres.

Figura 2. Caracterização temporal da estrutura do grafo gerado e do padrão de escrita das *hashtags* que o integram.

Em uma caracterização tipográfica, ou seja, a avaliação do padrão de escrita das *hashtags*, a Figura 2(b) representa a média da centralidade de proximidade para diferentes

intervalos de tamanho de *hashtag*. O tamanho é medido em número de caracteres de cada *hashtag*. Conforme esperado, para as primeiras horas de captura, percebe-se que todas as curvas possuem alta variabilidade dos valores de centralidade, marcada pelos grandes intervalos de confiança. Essa variabilidade reflete diretamente o uso de uma única *hashtag* como entrada, o que inicialmente favorece as *hashtags* presentes nos *tweets* relacionados a essa entrada enviesada. Assim, a entrada inicial é a mais representativa da comunidade formada pelo primeiro conjunto de *hashtags* coletadas. Outra conclusão está relacionada à maior centralidade das *hashtags* que não são muito curtas nem muito longas. A adesão dos usuários ao uso de *hashtags* de tamanho médio pode estar relacionada à praticidade de sua escrita, por não serem muito longas, e ao fato de poderem transmitir mais informações do que *hashtags* muito curtas.



(a) Dispersão dos graus dos nós que compõe o grafo de *hashtags*.

(b) Alterações em diferentes grupos de *hashtags* mais centrais ao longo do tempo.

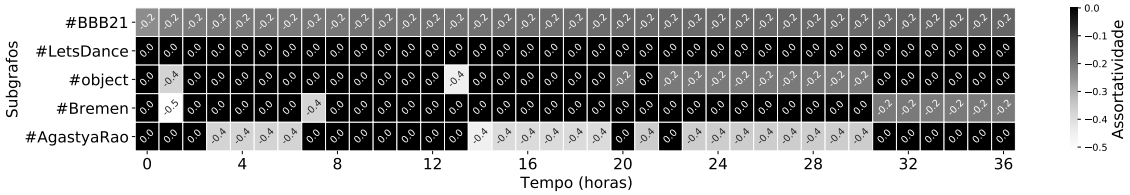
Figura 3. Além da notória discrepância entre graus dos nós, é possível notar a presença de um grupo mais coeso de *hashtags* mais centrais (Top20), uma vez que este apresenta apenas três alterações, cerca de 13%, de todas as permutações durante toda a captura do fluxo de dados.

Considerando todo o grafo relacional, analisa-se também a distribuição do grau dos nós, aspecto fundamental para compreender a dinâmica das associações entre *hashtags*. O diagrama de caixa na Figura 3(a) torna evidente a disparidade entre os graus dos nós do grafo. Uma vez que o grau de um nó em um grafo não orientado refere-se ao número de arestas que se conectam a ele, é possível afirmar que o valor médio do grau é constante e próximo a 10. No entanto, alguns nós estão muito distantes desse padrão de conectividade, variando entre 11 e mais de 1.000 conexões. Em particular, ao observar o nó com o grau mais discrepante em relação à média (*#BBB21*), corrobora-se a hipótese de viés no monitoramento.

A Figura 3(b) fornece uma visão discretizada do número de mudanças no grupo mais central de *hashtags* entre cada intervalo de captura de 1 hora. Analisando o grupo das 50 *hashtags* mais centrais (Top50), nota-se um declínio perceptível no número de alterações ao longo do tempo, enfatizado pelo período a partir da 24^a hora de captura, na qual não ocorrem mais alterações. Em contraste, a curva Top20 expressa um comportamento mais estável, com apenas quatro mudanças em todo o tempo de captura. Comparando as duas curvas, é possível inferir que (i) como o maior número de mudanças (87%) ocorre fora do Top20, há um núcleo quase imutável de *hashtags* mais centrais. Esse fenômeno fortalece a hipótese de que para entender um evento no *Twitter* de forma plena, basta monitorar um grupo restrito de *hashtags* relacionadas àquele evento, evitando desperdício de tempo e recursos; (ii) a estabilização do número de mudanças observadas

(a) Densidade.

(b) Diâmetro.



(c) Assortatividade.

Figura 4. Discretização temporal das métricas de redes complexas referentes aos cinco subgrafos mais populosos.

nas duas curvas pode ser interpretada como uma saturação do tema monitorado.

A existência de componentes desconexas entre si no grafo de *hashtags*, como discutida na análise da Figura 2(a), motiva a execução de uma análise temporal individualizada para cada subgrafo. Ao considerar a análise de métricas de redes complexas, como densidade, diâmetro e assortatividade, torna-se possível quantificar diversas características latentes que integram um grafo. Diante de eventuais uniões de subgrafos ao longo da captura, essa análise foca nos cinco subgrafos mais populosos que se mantêm disjuntos durante todo o processo. Em uma primeira perspectiva, a Figura 4(a) apresenta os resultados da análise para a métrica de densidade (d_e) dada pela Equação 2,

$$d_e = \frac{2m}{n(n-1)} \tag{2}$$

em que m e n são o número total de arestas e nós do subgrafo, respectivamente. Definida entre $[0, 1]$, torna-se evidente a disparidade entre as densidades dos subgrafos. O indicativo de baixa densidade, dado pela tonalidade clara observada no subgrafo associado à $\#BBB21$, evidencia uma constante prevalência da quantidade de nós em detrimento da quantidade de arestas. Isso reitera a suposição feita na análise da Figura 2(a), de que usuários ao debaterem um determinado tema preferem publicar *tweets* com novas associações de *hashtags* previamente usadas a inventar novas *hashtags*. Ressalta-se ainda que a baixa densidade do subgrafo principal está relacionada ao uso de combinações prevalentes de *hashtags*. Em uma segunda perspectiva comparativa, foram dispostos tem-

poralmente os diâmetros (d_{ia}) de cada subgrafo, calculando-os com base na Equação 3. O diâmetro de um grafo é expresso pela sua máxima excentricidade (ϵ), Equação 4, ou seja, a maior distância mínima, em números de saltos, entre dois nós v_i e v_j. Por essa razão, o diâmetro de um grafo pode assumir valores inteiros no intervalo de $[1, \infty[$.

$$d_{ia} = \max_{v_i \in \mathcal{V}} \epsilon(v_i) \qquad (3) \qquad \epsilon(n_i) = \max_{v_j \in \mathcal{V}} d(v_i, v_j) \qquad (4)$$

A Figura 4(b) mostra que há uma diferença significativa entre os diâmetros de cada subgrafo avaliado ao longo do tempo. Enquanto que outros subgrafos ($\#LetsDance$, $\#object$, $\#Bremen$ e $\#AgastyaRao$) apresentam diâmetros que não ultrapassam três saltos ($d_{ia} < 3$), o subgrafo principal ($\#BBB21$), relacionado ao evento monitorado, apresenta diâmetros variando entre 7 e 8. Mesmo sendo medida em saltos, essa distância entre nós pertencentes ao mesmo subgrafo revela *hashtags* relacionadas a assuntos ou nichos semânticos completamente dispares entre si.

Em uma terceira avaliação, observou-se o comportamento temporal dos subgrafos sob a perspectiva da assortatividade, métrica definida entre $[-1, 1]$ que quantifica a tendência dos nós de um grafo se conectarem a outros nós com características semelhantes. Ao considerar o grau como característica a ser avaliada, valores positivos de assortatividade indicam uma correlação entre nós de grau semelhante, enquanto valores negativos indicam relações entre nós de grau diferente. Valores nulos traduzem a completa conexão entre todos os nós em um grafo. Casos extremos, positivos ou negativos, mostram que o grafo exibe padrões de mistura entre ordenamentos perfeitos ou padrões não ordenados, respectivamente. Dessa maneira, pode-se constatar o recorrente padrão disassortativo do subgrafo ($\#BBB21$), evidenciando que a maioria das conexões das *hashtags* mais centrais provêm de *hashtags* com grau baixo. Em contraponto, subgrafos menos populosos exibem predominadamente valores nulos de assortatividade, uma consequência direta do tipo de estrutura formada por poucos nós completamente conectados.

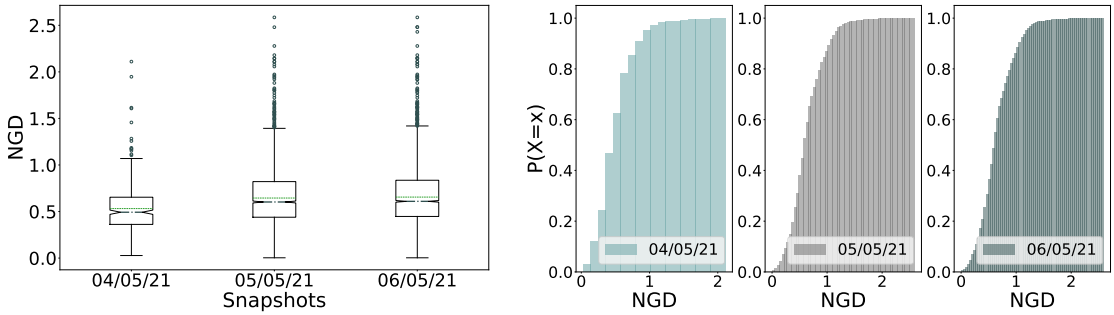
Diante da possível existência de domínios de conhecimento distintos dentro do mesmo subgrafo principal, torna-se importante realizar uma caracterização semântica para medir a distância semântica entre a *hashtag* mais central ($\#BBB21$) e as demais *hashtags* que o compõem. Contudo, a ausência do caractere espaço, informando o início e o término de uma palavra, dificulta a compreensão do significado de algumas *hashtags*. Uma vez que métodos computacionais tradicionalmente usados nessa medição dependem da aplicação de técnicas de tokenização ou incorporação de palavras (*word embeddings*), opta-se pelo cálculo da NGD (*Normalized Google Distance*) [Cilibrasi e Vitanyi, 2007], uma métrica assimétrica de distância semântica entre dois termos não vetorizados (x e y). Definida entre $[0, \infty[$, a NGD é expressa pela Equação 5 dada por

$$NGD(x, y) = \frac{max\{log\, f(x), log\, f(y)\} - log\, f(x, y)}{log\, \mathcal{N} - min\{log\, f(x), log\, f(y)\}}, \qquad (5)$$

em que f_x e f_y são o número de resultados retornados pelo *Google* para os termos pesquisados x e y, respectivamente, e $f_{x,y}$ é o número de páginas da web em que x e y estão presentes. Embora a escolha do parâmetro $\mathcal{N}$ não seja pré-estabelecido, utiliza-se um valor grande [5]. Usualmente dois termos podem ser considerados semanticamente des-

[5]Assim como outros trabalhos, este artigo adota $\mathcal{N} = 2,527 \times 10^{10}$, o número de resultados encontrados

correlacionados quando a $NGD > 1$. Sendo uma métrica assimétrica, o resultado de similaridade é consideravelmente impactado pela ordem de pesquisa dos termos. Por esta razão, o conjunto de valores utilizados nas Figuras 5(a) e 5(b) advém da média de ambas as configurações.



(a) Dispersão dos valores de distância semântica entre *hashtags*.

(b) Função de distribuição acumulada dos valores de distância semântica entre *hashtags*.

Figura 5. Caracterização semântica do subgrafo mais populoso em três momentos da captura.

A Figura 5(a) retrata a dispersão dos valores da distância semântica entre a ($\#BBB21$) e as demais *hashtags* do mesmo subgrafo em três momentos distintos: após a 1^a, 18^a e 37^a hora de monitoramento. Embora a média e a mediana dos valores da NGD, nos três momentos, circundem próximas a 0.55, percebe-se que nos dois últimos dias de monitoramento houve um aumento da ocorrência de *hashtags* com valores de NGD superiores aos demais.

Nesse sentido, cogita-se a hipótese de que *hashtags* cuja distância semântica até a *hashtag* central é maior $0,6$, possuem uma centralidade menor. Aplica-se o t-test de Welch, uma adaptação do Student's t-test indicada para conjuntos de amostra com diferentes variâncias ou tamanhos, para provar a suposição, refutando a hipótese nula, com significância estatística superior a 95% (valor $p << 0,05$). Suportado por essa verificação estatística e correlacionando-a com Figura 5(b), pode-se inferir o percentagem total de *hashtags* com significados suficientemente distantes do tema monitorado. Dependendo do momento observado, constata-se que o grafo relacional de *hashtags* obteve até 78% de coesão semântica.

6. Trabalhos Relacionados

As *hashtags* são usadas em estudos realizados para diferentes fins, sejam aqueles dedicados à detecção [Chen et al., 2018] e monitoramento de eventos [Giridhar et al., 2017], predição de popularidade [Huang et al., 2020] ou sistemas de recomendação [Alsini et al., 2020, Yang et al., 2020]. Em contrapartida, Liu *et al.* que desenvolvem um modelo de vetorização de *hashtags* para preservar a semântica, incorporam informações extraídas do conteúdo textual e das múltiplas relações estruturais dentro de um *tweet* [Liu et al., 2018]. Abordando a modelagem de tópicos, Wang *et al.* recorrem a diferentes tipos de grafos de *hashtags* para elaborar um arcabouço de modelagem semântica de *tweets* capaz de lidar com a natureza esparsa e ruidosa dos

ao pesquisar o artigo "a" no *Google*.

textos curtos [Wang et al., 2016]. Cossard *et al.* analisam no *Twitter* o debate sobre vacinação, tendo como premissa a existência de câmaras de eco [Cossard et al., 2020]. A partir da aplicação de técnicas de detecção de comunidade e de particionamento de grafos, é possível caracterizar e distinguir usuários em grupos de acordo com seus posicionamentos, defensivo ou cético, sobre vacinas. Por fim, as características relacionais e textuais de grupo são empregadas no treinamento de dois algoritmos de classificação. Zhang propõe uma análise empírica tridimensional sobre as *hashtags* compartilhadas no *Instagram* [Zhang, 2019]. Na dimensão espaço-temporal, os autores empregam um algoritmo não supervisionado para agrupar *hashtags* de acordo com a trajetória temporal desses termos. Explorando a dimensão semântica, os autores incorporam o significado de cada *hashtag* ao longo dos anos em vetores distintos e os comparam aplicando a similaridade do cosseno. Na dimensão social, investiga-se a atuação das *hashtags* como características válidas em uma abordagem não supervisionada para inferir relações sociais. Resultados mostram não somente uma maior tendência dos usuários em compartilhar *hashtags* em locais específicos, como também que aproximadamente 10% das *hashtags* sofrem deslocamento semântico. Cui *et al.* propõem uma abordagem para investigar as propriedades das *hashtags* para detecção de eventos repentinos, englobando tanto a caracterização quanto a aplicação de um algoritmo não supervisionado [Cui et al., 2012]. Para associar cada *hashtag* à ocorrência de um evento específico, os autores estabelecem níveis de intensidade em uma caracterização tridimensional contendo atributos como instabilidade temporal, possibilidade de ser conteúdo humorístico e entropia de autoria. Kowald *et al.* exploram a caracterização temporal para compreender o processo de reutilização de *hashtags* no *Twitter* [Kowald et al., 2017]. Modelando a contribuição temporal de duas perspectivas, uma considerando *hashtags* de usuários antigos e a outra contendo *hashtags* postadas por seguidores do usuário, os autores desenvolvem um algoritmo preditivo para recomendação de *hashtags* inspirado na cognição humana.

Ao contrário dos trabalhos anteriores de caracterização do uso de *hashtags*, que empregam dados represados, este artigo prioriza os dados em fluxo uma vez que se concentra na captura contínua de *hashtags* relacionadas a eventos em tempo real. Esta condição introduz dinamismo à abordagem, possibilitando monitorar tendências tipográficas, padrões de conexão, além de detectar qualquer disjunção de domínios do conhecimento.

7. Conclusão

Este artigo propôs uma abordagem incremental capaz de capturar, processar e caracterizar temporalmente e semanticamente as *hashtags* compartilhadas por usuários no *Twitter* durante um evento em tempo real. A proposta foi desenvolvida na linguagem Python, integrando o Apache Kafka e o Apache Spark na ingestão e processamento dos dados coletados do *Twitter*. A proposta emprega a API do *Twitter* para realizar a coleta do fluxo de *tweets* relacionado a uma *hashtag* de entrada. As *hashtags* em cada *tweet* capturado são extraídas e inseridas em um grafo relacional incrementado em tempo de execução, em que cada nó representa uma *hashtag* distinta e uma aresta representa a ocorrência simultânea de duas *hashtags* em um mesmo *tweet*. Visando garantir dinamicidade e adaptabilidade na manutenção do grafo, as *hashtags* de entrada são periodicamente substituídas pelas *hashtags* mais centrais através da métrica de centralidade de proximidade. Após 37 horas de monitoramento de um evento de grande engajamento em redes sociais, foi possível re-

velar padrões latentes no comportamento dos usuários. Em postagens, usuários do *Twitter* são mais adeptos a empregarem novas combinações de *hashtags* previamente usadas do que criar suas próprias *hashtags*. As *hashtags* publicadas costumam possuir um tamanho mediano, entre 10 e 17 caracteres. Mostrou-se também que mesmo realizando um monitoramento direcionado a um evento específico, o grafo gerado pode conter *hashtags* relacionadas a assuntos ou nichos semânticos completamente dispares entre si. Além disso, a presença de componentes desconexas entre si durante todo período de captura e manutenção do grafo, revela indícios de *hashtags* oportunistas presentes em *tweets* relacionados ao tema monitorado. Como trabalhos futuros, pretende-se agregar algoritmos de detecção de comunidade na análise temporal, bem como incorporar outros tipos de grafos e lógicas de atualização do monitoramento.

Referências

Alsini, A., Datta, A. e Huynh, D. Q. (2020). On utilizing communities detected from social networks in hashtag recommendation. *IEEE Transactions on Computational Social Systems*, 7(4):971–982.

Chen, X., Zhou, X., Sellis, T. e Li, X. (2018). Social event detection with retweeting behavior correlation. *Expert Systems with Applications*, 114:516–523.

Cilibrasi, R. L. e Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3):370–383.

Colleoni, E., Rozza, A. e Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332.

Cossard, A., De Francisci Morales, G., Kalimeri, K., Mejova, Y., Paolotti, D. e Starnini, M. (2020). Falling into the echo chamber: The italian vaccination debate on twitter. Em *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, p. 130–140.

Cui, A., Zhang, M., Liu, Y., Ma, S. e Zhang, K. (2012). Discover breaking events with popular hashtags in twitter. Em *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, p. 1794–1798, New York, NY, USA. Association for Computing Machinery.

de Oliveira, N. R., Medeiros, D. S. V. e Mattos, D. M. (2020). A syntactic-relationship approach to construct well-informative knowledge graphs representation. Em *2020 4th Conference on Cloud and Internet of Things (CIoT)*, p. 75–82.

de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V. e Mattos, D. M. F. (2021). Identifying fake news on social networks based on natural language processing: Trends and challenges. *Information*, 12(1).

Giridhar, P., Wang, S., Abdelzaher, T., Amin, T. A. e Kaplan, L. (2017). Social fusion: Integrating twitter and instagram for event monitoring. Em *2017 IEEE International Conference on Autonomic Computing (ICAC)*, p. 1–10.

Huang, J., Tang, Y., Hu, Y., Li, J. e Hu, C. (2020). Predicting the active period of popularity evolution: A case study on twitter hashtags. *Information Sciences*, 512:315–326.

Kowald, D., Pujari, S. C. e Lex, E. (2017). Temporal effects on hashtag reuse in twitter: A cognitive-inspired hashtag recommendation approach. Em *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, p. 1401–1410. International World Wide Web Conferences Steering Committee.

Liu, J., He, Z. e Huang, Y. (2018). Hashtag2vec: Learning hashtag representation with relational hierarchical embedding model. Em *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, p. 3456–3462. International Joint Conferences on Artificial Intelligence Organization.

Liu, X. e Buyya, R. (2020). Resource management and scheduling in distributed stream processing systems: A taxonomy, review, and future directions. *ACM Comput. Surv.*, 53(3).

Lopez, M. A., Lobato, A. G. P. e Duarte, O. C. M. B. (2016a). Monitoramento de tráfego e detecção de ameaças por sistemas distribuıdos de processamento de fluxos: uma análise de desempenho. Em *Anais do XXI Workshop de Gerência e Operação de Redes e Serviços*. SBC.

Lopez, M. A., Lobato, A. G. P. e Duarte, O. C. M. B. (2016b). A performance comparison of open-source stream processing platforms. Em *2016 IEEE Global Communications Conference (GLOBECOM)*, p. 1–6.

Nasiri, H., Nasehi, S. e Goudarzi, M. (2019). Evaluation of distributed stream processing frameworks for iot applications in smart cities. *Journal of Big Data*, 6(1):52.

Rao, T. R., Mitra, P., Bhatt, R. e Goswami, A. (2019). The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems*, 60(3):1165–1245.

Stilo, G. e Velardi, P. (2017). Hashtag sense clustering based on temporal similarity. *Comput. Linguist.*, 43(1):181–200.

Sun, A. Y., Zhong, Z., Jeong, H. e Yang, Q. (2019). Building complex event processing capability for intelligent environmental monitoring. *Environmental Modelling & Software*, 116:1–6.

Vicario, M. D., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G. e Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports*, 6.

Wang, Y., Liu, J., Huang, Y. e Feng, X. (2016). Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1919–1933.

Yadranjiaghdam, B., Yasrobi, S. e Tabrizi, N. (2017). Developing a real-time data analytics framework for twitter streaming data. Em *2017 IEEE International Congress on Big Data (BigData Congress)*, p. 329–336.

Yang, C., Wang, X. e Jiang, B. (2020). Sentiment enhanced multi-modal hashtag recommendation for micro-videos. *IEEE Access*, 8:78252–78264.

Zhang, Y. (2019). Language in our time: An empirical analysis of hashtags. Em *The World Wide Web Conference*, WWW '19, p. 2378–2389, New York, NY, USA. Association for Computing Machinery.