Análise Temporal de Coesão de Discurso em Mídia Social Durante Grandes Eventos

João Matheus N. Gonçalves[1], Jonice Oliveira[1], Fabio Porto[2], Tiago C. França[3]

[1]Instituto de Computação – UFRJ – Rio de Janeiro – RJ – Brasil

[2]Departamento de Ciência da Computação – LNCC – Petrópolis – RJ – Brasil

[3]Departamento de Computação – UFRRJ – Rio de Janeiro – RJ – Brasil

{joaomng, jonice}@dcc.ufrj.br, fporto@lncc.br, tcruzfranca@ufrrj.br

Abstract. *Events like COVID-19 lead to a large number of publications on social media. Different publications and sub-events are discussed in such a way that the discourse may or may not be aligned, leading to greater or lesser textual cohesion between publications. In this work, a method is proposed for the analysis of textual cohesion and its variation over time. The method was used and evaluated with synthetic databases built with known levels of cohesion and subsequently applied to a database of tweets published during the pandemic. With the results, it was possible to understand the evolution of cohesion over time in tweets written in portuguese related to COVID-19.*

Resumo. *Uma grande quantidade de publicações ocorre nas mídias sociais relacionadas a eventos como a COVID-19. Diferentes publicações e subeventos são tratados, fazendo com que o discurso possa ou não estar alinhado, levando a uma maior ou menor coesão textual entre as publicações. Neste trabalho, é apresentado o método VERSATILE para análise de coesão textual e sua variação ao longo do tempo. O método foi avaliado com bases sintéticas construídas com níveis de coesão conhecidos, e posteriormente aplicada a uma base de tuítes publicados durante a pandemia. Com os resultados, foi possível compreender a evolução da coesão ao longo do tempo em tuítes em português sobre a COVID-19.*

1. Introdução

Coletivamente, os usuários das mídias sociais (MS) geram, rapidamente, um grande volume de dados sobre eventos que acontecem no mundo. Em grandes eventos (como a pandemia da COVID-19), o volume de interações sobre o assunto é muito grande e, devido à longa duração do evento, o conteúdo pode variar, por exemplo com o surgimento de subtemas relacionados a um mesmo grande evento. A análise do volume de dados requer soluções computacionais que viabilizem a identificação e a observação de aspectos do discurso.

Este trabalho se volta à análise da coesão textual do discurso em MS. A coesão diz respeito à manutenção da continuidade semântica, interpretabilidade e sentido de um texto, mantendo conectadas as partes que o compõem [Antunes 2005]. Para [Halliday e Hasan 1976], a coesão diferencia-se em duas categorias: a gramatical, que conecta componentes de um texto através de sua estrutura, usando mecanismos como conjunções, elipse e substituição; e a lexical, que está relacionada à escolha dos termos num texto, e se dá, sobretudo, através da reiteração. A reiteração pode ser obtida pela repetição de um mesmo termo ou pelo uso sinônimos ou palavras muito próximas que tenham a mesma forma base.

O objetivo deste trabalho é colaborar com a área de análise de discurso construindo um método para investigação de coesão textual de publicações relacionadas a um grande evento de longa duração (e com diferentes ocorrências/subeventos). Foi realizada uma análise do método proposto (o VERSATILE) com bases de dados sintéticas, criadas de forma que

possuíssem características desejadas. Em seguida, realizou-se a análise de uma base de tuítes sobre a COVID-19. A principal contribuição deste trabalho é o método VERSATILE para análise de coesão de bases textuais de documentos que representam discursos em MS.

Existem algumas ferramentas automáticas de análise de coesão textual, utilizando métodos como vetorização [Crossley *et al.* 2019] ou outras representações em grafos [Lachner e Neuburg 2019], e analisando diferentes tipos de coesão. O presente trabalho analisa a coesão lexical, uma vez que busca reiteração entre múltiplos documentos de texto. O método proposto utiliza e combina a análise de redes de cliques [Fadigas e Pereira 2013] com métodos para análise de grafos que variam no tempo conforme proposto em [França 2019]. O VERSATILE expande o que foi proposto em [Fadigas e Pereira 2013] ao aplicar e analisar as métricas para redes de cliques que não são minimamente conectadas. Além disso, aplica a análise de coesão a textos de MS e analisa a variação da coesão ao longo do tempo.

2. O Método VERSATILE

O VERSATILE foi definido em etapas (representadas no diagrama da Figura 1) que se completam para realizar a análise da coesão textual e a variação do discurso no tempo de publicações sobre um tema em uma mídia social. As análises para identificação da coesão no texto são realizadas usando redes de cliques (grafos completos).

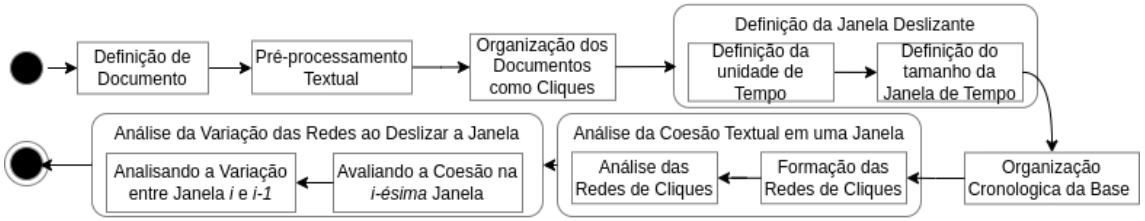

Figura 1. Etapas do método VERSATILE

Para abordar a questão da variação no tempo, foi elaborada uma estratégia de janelas (intervalos de tempo) de análise que deslizam em unidades de tempo. Os documentos utilizados para a análise precisam estar associados a estampas de tempo do momento da sua publicação. O grafo que representa a rede será construído e analisado usando uma abordagem de Grafo Variando no Tempo (*Time Varying Graph* ou TVG). O TVG permite a representação e análise da variação da rede, possibilitando a avaliação da topologia dessa rede no tempo. A variação ocorre quando novos nós e conexões surgem (ou se intensificam) na rede, enquanto outros deixam de existir (ou enfraquecem) [Casteigts *et al*. 2012; Petter e Jari 2013]. A representação de TVG adotada neste trabalho será uma sequência ordenada de grafos.

Na etapa de **definição do documento**, um documento de texto (d) será utilizado como unidade a ser analisada. Todo d pertence a uma base textual (B) ($d \in B$) composta por N publicações coletadas de uma mídia social. O documento pode ser todo texto da publicação, uma frase ou um parágrafo de uma publicação. A definição e a granularidade do documento dependem da análise a ser realizada.

No **pré-processamento textual** foi feita: a conversão de letras para minúsculas, a remoção da pontuação, a tokenização, a remoção de *stopwords* (palavras como artigos ou conectivos textuais), e a extração de radicais com técnicas como *lemmatization* ou *stemming* [Khyani *et al.* 2021]. Cada palavra será um token (n-gram igual 1). Os tokens foram agrupados e identificados como pertencendo a um documento.

A **organização dos documentos como cliques** sucede o pré-processamento, e é feita a partir da criação de cliques (grafos isolados), conforme proposto por [Fadigas e Pereira

2013]. Um clique é um grafo G = {V,E}, tal que, para todo par $v_i, v_j \in V$, onde $v_i \neq v_j$, temos $(v_i, v_j) \in E$. Os cliques representam os documentos pré-processados, sendo os tokens os nós do grafo, sendo as arestas definidas pela coexistência dos tokens em um documento.

Para a **definição da janela deslizante**, são estabelecidos os intervalos a serem analisados (janelas) como divisões no tempo de duração do evento estudado. O termo "deslizante" faz referência ao intervalo entre o começo de uma janela e o começo da próxima, de modo que as janelas "deslizam" uma certa quantidade de tempo. Define-se a **unidade de tempo** (*t*) para organizar e analisar os dados em uma granularidade (meses, dias, horas, minutos, etc.) que possibilite a melhor investigação sobre o evento estudado. Uma **janela de tempo** (T) corresponde a um período de duração *p* múltiplo de *t*. Se *t* corresponde a 1 dia, pode-se definir janelas de 5*t*, então *p*=5*t*. Assim, temos $T_1 = \{t_1, t_2, t_3, t_4, t_5\}$ que corresponderá aos documentos publicados nesse intervalo.

O **deslizamento das janelas** (*s*) leva à criação de conjuntos com interseções entre si. Assim, uma janela T_i desliza em um intervalo *s* (múltiplo de *t*) para a próxima janela determinada por T_{i+1}. As janelas T_i e T_{i+1} possuirão publicações em comum, sendo o tamanho da interseção dado por *p* - *s*. Em janelas com duração *p*=5 dias e deslizamento *s*=2 dias, haverá interseção de 3 dias entre duas janelas contíguas. A Figura 2 apresenta um exemplo do uso da janela deslizante. No exemplo, *t*=1 dia, e os dados estão distribuídos em um intervalo de 10 dias. A janela é de *p*=5*t* e devem deslizar 1 dia (*s*=1), havendo 4 dias de interseção.

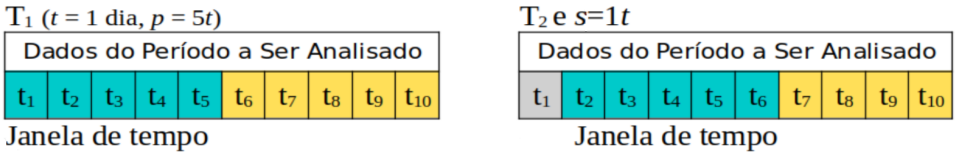

Figura 2 - Exemplo da janela deslizante com duração de 5*t* e deslizamento *s*=1*t*.

Na **organização cronológica**, os dados são ordenados de acordo com a estampa de tempo das mensagens. Essa etapa trata da preparação para que a análise seja realizada nas janelas de acordo com a unidade de tempo definida.

A **análise da coesão textual em uma janela** se baseou em [Fadigas e Pereira 2013]. Ela se inicia após o processo da formação da rede de cliques. A conexão dos cliques acontecerá de acordo com as palavras (tokens) iguais entre eles. As conexões ocorrem por justaposição (dois ou mais cliques possuem um nó em comum) ou sobreposição (mais de um nó de dois ou mais cliques são iguais). Uma rede com poucas ocorrências desses processos seria pouco coesa, uma vez que isso indicaria menor ocorrência de formas de reiteração dos termos entre os documentos. As métricas utilizadas para análise da coesão foram: variação de densidade - $v(\Delta)$; variação do grau médio - $v(\langle k \rangle)$; coeficiente de clusterização - *C*; fragmentação - *F;* e fragmentação de cliques - F_{cliques} [Fadigas e Pereira 2013].

A **análise da variação das redes ao deslizar a janela** possibilita que se perceba a diferença entre os índices de coesão ao longo do tempo. Numa base textual relacionada a um longo período, haverá a oportunidade de observar a variação das métricas de coesão no tempo. Cada métrica possibilitará, em algum grau, a inferência de aspectos do discurso.

4. Aplicação do VERSATILE e Análise dos Resultados

O VERSATILE foi aplicado em bases sintéticas (com diferentes níveis de coesão preestabelecidos) que simulavam tuítes publicados ao longo do tempo; e numa base textual do

Twitter, extraída de [Neves *et. al* 2022]. Para todas as bases, o documento usado foi o texto integral do tuíte.

Foram construídas três bases sintéticas: uma "pouco conexa", uma "conexa" e uma "muito conexa". Cada mensagem nas bases recebeu uma estampa de tempo com data e hora da publicação (como acontece com os tuítes). Para a construção dessas bases, foi definido que as palavras nas mensagens se repetiriam em uma porcentagem x do total de palavras usadas. Para a base "pouco conexa", o valor de x foi definido empiricamente para estar entre 0% e 12%; para a base "conexa", x variou de 15% a 33%; e para a base "muito conexa", este valor variou entre 70% e 95%. Estas diferenças na construção das bases influenciam diretamente as redes de cliques geradas ao se aplicar o VERSATILE, como mostra a Figura 4.

A base do Twitter foi obtida de [Neves *et. al* 2022]. As mensagens foram publicadas por um conjunto de perfis e estão relacionadas ao período da COVID-19 no Brasil. Os dados são do período de março a maio de 2020. A coleta foi realizada usando a API do Twitter dentro das permissões vigentes até 2022. Foram analisados 159 tuítes com o propósito de demonstrar a aplicação do VERSATILE em mensagens relacionadas a um evento real.

4.1. Análise das Bases Sintéticas

Foram analisados os índices de coesão das janelas de tempo correspondendo a uma transição gradual entre: i) uma base pouco conexa (B1) e uma base conexa (B2); e ii) duas bases conexas diferentes (B2 e B3). A unidade de tempo (t) foi de um dia (t=1), a janela p=10 dias, e o deslize s=1 dia. O propósito foi simular cenários de mudança na coesão do discurso em MS, variando de discursos mais ou menos coesos sobre diferentes assuntos ao longo do tempo. As métricas de coesão também foram calculadas para cada uma das bases sintéticas, usando uma janela de 10 dias que correspondia à integralidade da base (sem deslizar).

As bases simulavam mensagens ao longo de dez dias. Para a transição entre bases B1 e B2, considerou-se que o primeiro dia de B2 estaria logo após o último dia de B1. Então, aplicou-se o VERSATILE a partir do segundo dia de B1, com a última janela, T9, tendo início no último dia desta base. Como $p=10$ e $s=1$, a janela T_1 correspondia aos nove últimos dias de B1 e o primeiro dia de B2; T_2 aos oito últimos dias de B1 e dois primeiros de B2, e a última janela, T_9, começava no último dia de B1 e terminava no penúltimo dia de B2.

A Figura 4 (a) apresenta o resultado das análises da transição da base pouco conexa para a base conexa, onde observou-se aumento gradual de índices de coesão como $v(\Delta)$ e $v(\langle k \rangle)$. A Figura 4 (b) apresenta os resultados das análises quando duas bases conexas são colocadas adjacentes uma à outra. Mesmo sendo as duas bases conexas, as bases têm "assuntos" diferentes (por serem formadas de bancos de palavras diferentes). Assim, a conectividade na transição não é tão grande quanto em cada base individualmente.

4.2. Base de Tuítes da COVID-19

Para análise dos tuítes da COVID-19, definiu-se t=1 dia, p=6 dias e s=1 dia. Após explorar os dados, percebeu-se que a janela de 6 dias proporcionaria um volume considerável de dados textuais, mas reduziria a chance de que os textos nas janelas ficassem muito abrangentes. Tais janelas podem diferir estruturalmente. Ou seja, enquanto as janelas das bases sintéticas eram de 10 dias e tinham invariavelmente 200 tuítes (todos com 20 palavras), nas janelas da base do Twitter, os valores podem variar.

Nas bases de dados com os textos dos tuítes (Figura 4 (c)), pode-se observar um $v(\Delta)$ majoritariamente entre 0,5 e 1,6, sendo o índice mais baixo próximo daquele observado em

bases sintéticas “conexas”. O mais alto não chega a ser a variação observada nas bases “muito conexas”, pois estas tinham variação de densidade próxima de 8. Isto aponta para um nível de coesão médio de acordo com a análise das bases sintéticas.

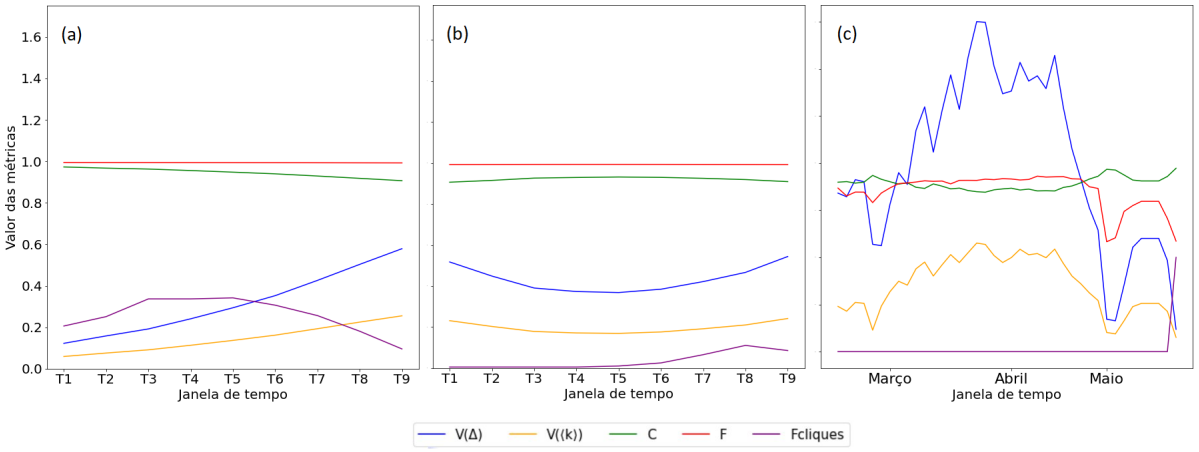

Figura 4. Variação de janelas em bases: a)Transição da base pouco conexa para a conexa; b) Transição entre duas bases conexas diferentes; c) Base do Twitter.

O $v(\langle k \rangle)$ de 18/03 a 13/04 esteve, em geral, maior do que o valor observado para as bases sintéticas “conexas”, mas sem chegar aos valores da base “muito conexa”, sendo um indício de uma quantidade considerável reiterações entre os tuítes, colaborando para sua coesão. Foram poucas as janelas nas quais os valores estiveram abaixo daqueles encontrados para bases sintéticas conexas. Em 17/03 aconteceu a primeira morte por COVID-19 no Brasil. Nessa data, passou a ser crime desrespeitar as medidas de isolamento [Sanarmed 2020].

O valor de C das janelas ficou majoritariamente abaixo daquele em bases sintéticas conexas, mas com algumas ocorrências próximas de 1 (valor máximo) e mantendo-se estável ao longo do período estudado. Os valores mais baixos de C provavelmente se relacionam a uma maior coesão, visto que indicam um “distanciamento” da rede de cliques para o estado inicial de cliques isolados, no qual C sempre é igual a 1.

Com os textos do Twitter houve diferença significativa de fragmentação entre janelas diferentes. Esta diminuiu consideravelmente nas últimas 3 janelas, se aproximando do índice das bases sintéticas “muito conexas”. A fragmentação de cliques foi nula exceto por uma janela, o que significa que o estado final da rede de cliques, em quase todos os casos, apresentou um número de componentes igual a 1. Disto segue que o estado final das redes de cliques foi, nestes casos, um grafo conexo, o que é um indicador positivo de coesão conforme a análise das bases sintéticas.

Ressalta-se a variação brusca das métricas em 5 janelas de tempo contíguas a partir de 13/04. Este período está relacionado a subeventos muito importantes (causando uma variação e diversidade de assuntos), como a demissão do ministro da saúde (16/04), o Brasil fica de fora da ACT Accelerator (05/05), chegou-se a 10.000 mortos (09/05), a contratação e queda do 2º Ministro da Saúde (15/05). Porém, não foi possível investigar tal suposição, porque a coleta teve problemas nas janelas começando em 14/04 até aquelas começando em 7/05.

6. Conclusão

Neste trabalho foi apresentado e aplicado o método VERSATILE para avaliação da coesão textual lexical entre mensagens publicadas por diferentes usuários de uma MS que estão relacionadas a um assunto em comum. Foram realizados testes com bases sintéticas e com

uma base da COVID-19 de tuítes escritos em português brasileiro. Foi possível perceber o nível de coesão da base, o que demonstrou a diferença entre os termos usados nas publicações na época. A partir dos resultados, percebeu-se o potencial do método para análise automática da coesão de conteúdo textual publicado nas MS.

Como trabalho futuro, vislumbra-se a análise de bases maiores, bem como o aperfeiçoamento do método, levando em conta a presença de sinônimos ou a função sintagmática dos termos em orações ao analisar um texto, como feito em [Lachner e Neuburg 2019]. Também é possível explorar outras características do grafo formado para identificar o conteúdo do discurso. Um exemplo seria encontrar *clusters* nos grafos, indicando padrões e estruturas neles presentes.

Agradecimentos

Este trabalho foi apoiado em parte por créditos e recursos da Oracle Cloud, providos pelo programa Oracle for Research (award number CPQ-2160239). Agradecemos também ao CNPq, CAPES e LNCC por todo o suporte.

Referências

Antunes, C. (2005). “Lutar com as palavras: coesão e coerência”. São Paulo: Parábola Editorial.

Crossley, S.A., Kyle, K. e Dascalu, M. (2019) “The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap”. *Behav Res* 51, 14–2. https://doi.org/10.3758/s13428-018-1142-4.

Fadigas, I.S. e Pereira, H.B.B. (2013) “A network approach based on cliques”. Physica A: Statistical Mechanics and its Applications.

França, T. C.. "ANDARE: um *framework* para inclusão da análise de dados de mídias sociais no contexto da preparação e resposta à emergência em situações de manifestações de massa", 2019, Tese (Doutorado) - Curso de Pós-graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019, https://tinyurl.com/tmaydae4. Acesso em: 08 Mar. 2023.

Halliday, M. e Hasan, R. (1976) “Cohesion in English”. London: Longman Group Ltd.

Khyani, D., Siddhartha, B. S., Niveditha, N. M., e Divya, B. M. (2021) “An interpretation of lemmatization and stemming in natural language processing”. Journal of University of Shanghai for Science and Technology, *22*(10), P. 350-357.

Lachner, A. e Neuburg, C. (2019) “Learning by writing explanations: computer-based feedback about the explanatory cohesion enhances students’ transfer”. *Instr Sci* 47, 19–37, https://doi.org/10.1007/s11251-018-9470-4.

Neves, J. C. B. ; França, Tiago Cruz ; Bastos, M. P.; Carvalho, P. V. R.; Gomes, J. O. Analysis of government agencies and stakeholders? twitter communications during the first surge of COVID-19 in Brazil. WORK-A Journal of Prevention Assessment & Rehabilitation, v. 73, p. 1-13, 2022.

Sanarmed (2020). “Linha do tempo do coronavirus no Brasil”, https://www.sanarmed.com/linha-do-tempo-do-coronavirus-no-brasil.