sbbd:39

Análise da Variação da Coesão do Discurso sobre a COVID-19 em Mídia Social

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]DEXL Lab – 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. *During the course of extreme events, such as the COVID-19 pandemic, a large volume of publications on the topic tends to occur on social media. The public reports and shares opinions about the event and its sub-events, leading to a variation of the discourse over time, which requires computational solutions for its analysis, mainly due to the volume of data and duration of the analyzed period. In this work, we apply the VERSATILE method, for the analysis of textual cohesion over time, in a database with a large volume of tweets in Portuguese published in the first semester of the pandemic. It was possible to correlate variations in textual cohesion with sub-events related to COVID-19 in Brazil and around the world, in addition to better understanding the metrics used for the analysis.*

Resumo. *Durante o curso de eventos extremos, como a pandemia de COVID-19, um grande volume de publicações acerca do tema tende a ocorrer em mídias sociais. O público relata e opina sobre o evento e também subeventos, levando a uma variação do discurso ao longo do tempo, que necessita de soluções computacionais para sua análise, devido sobretudo ao volume de dados e duração do período analisado. Neste trabalho, aplicamos o método VERSATILE, para análise de coesão textual ao longo do tempo, numa base com um grande volume de tuítes em português publicados no primeiro semestre da pandemia. Foi possível correlacionar as variações de coesão textual com os subeventos relacionados à COVID-19 no Brasil e no mundo, além de compreender melhor as métricas utilizadas para a análise.*

1. Introdução

A população gera muita informação nas mídias sociais (MS), comentando acontecimentos e compartilhando opiniões. Durante eventos extremos, as pessoas fazem publicações relacionadas às ocorrências em seus perfis nessas mídias [Lin *et. al* 2016; França 2019]. O conteúdo das publicações pode variar de acordo com os interesses do usuário ou por causa da mudança dos acontecimentos relacionados ao evento.

A pandemia da COVID-19 foi um evento prolongado e bastante comentado nas MS [Tsao *et. al* 2021]. Foi um evento extremo que motivou um grande número de publicações nas MS, que se relacionavam com os acontecimentos e mudaram à medida que novidades foram surgindo. Por exemplo, no Brasil, o primeiro tema era o isolamento social. Depois, a demissão do ministro da saúde. Outros temas como uso de máscara, vacinas, entre outros [Neves et al. 2022] surgiram, ganharam atenção e deram espaço para outros acontecimentos. Ou seja, os acontecimentos durante a pandemia motivaram as publicações (e, consequentemente, o discurso) nas MS. Diferentes análises sobre a COVID-19 foram e estão sendo realizadas usando dados de MS [Fernandes *et. al* 2020; Tsao *et. al* 2021; Recuero e Soares 2021; Neves et al. 2022].

A atenção coletiva dada pelos usuários a determinado acontecimento relacionado ao evento afetará a coesão do discurso presente nas publicações. [Antunes 2005] definiu a

coesão textual como sendo aquilo que confere ao texto uma manutenção da continuidade semântica, interpretabilidade e sentido, mantendo as partes do texto conectadas. A coesão pode ser gramatical e lexical. A primeira está relacionada à conexão de componentes de um texto através de sua estrutura por meio do emprego de mecanismos como conjunções, elipse e substituição; a segunda está relacionada à escolha dos termos num texto, e se dá, principalmente, por meio da reiteração com a repetição do mesmo termo, uso de sinônimos ou emprego de palavras muito próximas que tenham a mesma forma base [Halliday e Hasan 1976]. Porém, analisar a coesão em bases textuais sobre a COVID-19 extraídas de mídia social e avaliar a mudança de temas é uma tarefa que requer meios apropriados capazes de lidar com grandes volumes de dados.

A avaliação da coesão textual de bases de mídias sociais possibilitará a identificação da existência (ou não) de um foco comum (independente de discordâncias) entre as publicações dos diferentes usuários, através das ocorrências de reiteração ao longo do texto de diferentes publicações. A análise da variação da coesão do discurso no tempo pode proporcionar compreensões relevantes sobre a mudança de foco dos usuários de uma mídia social tomando por base seus discursos. Isto é, um assunto pode deixar de ser o alvo das publicações, por causa de um novo acontecimento associado ao evento. A baixa coesão pode indicar mensagens abordando diferentes assuntos (foco variado); enquanto uma maior coesão indica que os usuários estão abordando temas comuns. Já a variação da coesão no tempo pode indicar que os usuários mudaram o tema das suas publicações indo para uma maior ou menor coesão. Finalmente, a variação da coesão pode também indicar a mudança dos assuntos e estar associada ao surgimento de novos acontecimentos relacionados a um evento.

O objetivo deste trabalho é analisar o discurso em mídia sociais durante o primeiro semestre da pandemia através da análise de coesão textual lexical em um grande volume de dados textuais. Para tanto, utilizamos uma base com tuítes de 01/01/2020 a 31/05/2020 [Melo e Figueiredo 2020] que versam sobre a COVID-19, e aplicamos o método VERSATILE [Gonçalves *et. al* 2023], analisando a variação da coesão ao longo de janelas de tempo no período mencionado, e tomando como base a análise de coesão para bases textuais sintéticas em [Gonçalves *et. al* 2023].

2. Trabalhos Correlatos

[Recuero e Soares 2021] analisaram o discurso sobre COVID-19 no Twitter empregando técnicas de análise de redes sociais, os nós na rede os *retweets* e *menções* as arestas. Eles também analisaram o conteúdo das publicações focando em desinformação relacionada a cura para a COVID-19. [Fernandes *et. al* 2020] também analisaram o discurso sobre a pandemia, mas observando apenas um perfil no Instagram. [Neves et al. 2022] avaliaram os tipos de conteúdo das mensagens de perfis oficiais do governo e de órgãos de comunicação classificando manualmente o conteúdo como orientação ou informação desnecessária, por exemplo. Outros trabalhos analisaram o discurso sobre a COVID-19, ainda que de forma não-automatizada e não exclusivamente em MS: [Saldanha 2020] analisou o discurso acerca de ensino remoto durante a pandemia; e [Neto *et. al* 2020] buscou identificar e analisar *fake news* acerca da COVID-19. Esses trabalhos não analisaram a coesão textual.

Sobre a coesão textual, [Crossley *et. al* 2019] apresentaram uma ferramenta automatizada que, entre outras coisas, procura elos (*links*) entre diferentes documentos textuais. A análise é realizada com emprego de métodos de vetorização de palavras e análise semântica, além de incluir a análise do ponto de vista lexical, buscando palavras-chave iguais em documentos textuais distintos. Esta ferramenta foi criada com um propósito de uso geral,

podendo ser aplicada em diversos contextos além de MS. Diferente do método adotado neste trabalho, a proposta desses autores não se aplica a textos escritos em idiomas diferentes do inglês e não analisam a variação temporal da coesão no tempo, além de não usar redes de cliques.

[Muttaqien 2019] analisou a coesão textual em MS. O autor focou no que ele chamou de "coesão sistêmica" que é aquela que depende de mecanismos específicos da mídia social utilizada (no caso do Twitter, características como menções e hashtags). A análise de coesão foi feita de maneira manual.

3. Descrição da Base e Método de Análise

A base de tuítes de [Melo e Figueiredo 2020] foi utilizada. As mensagens foram coletadas do Twitter usando palavras-chave relacionadas à pandemia da COVID-19 no Brasil. Foram analisados 700.000 tuítes publicados entre 1/01/2020 a 31/05/2020.

Para a análise de discurso, utilizou-se o método VERSATILE [Gonçalves *et. al* 2023]. Dada uma base textual, define-se o documento a ser analisado; faz-se o pré-processamento do texto; organizam-se os documentos como cliques (grafos completos); definem-se os parâmetros para a análise temporal da coesão; faz-se a organização cronológica da base textual; para cada janela de tempo, forma-se a rede de cliques através de sobreposição e justaposição [Fadigas e Perera 2013]; e analisam-se as redes de cliques e como elas mudam ao longo do tempo. Para tal análise, são calculadas algumas métricas de coesão propostas em [Fadigas e Pereira 2013] nas redes de cliques obtidas para cada janela de tempo: variação de densidade ($v(\Delta)$), variação do grau médio ($v(\langle k \rangle)$), coeficiente de clusterização (C), fragmentação (F) e fragmentação de cliques ($F_{cliques}$). As definições destas métricas constam em [Fadigas e Pereira 2013].

O documento textual utilizado para a análise foi o texto completo de cada tuíte e a estampa de tempo relacionada ao momento da sua publicação. Para definição da janela deslizante, a unidade de tempo *t* foi de um dia, o tamanho *p* de cada janela foi de 6 dias, e o deslize *s* foi de um dia. O pré-processamento incluiu a remoção de pontuação, links, caracteres especiais, *stopwords* e lematização dos tokens.

Ao analisar as métricas de coesão para as janelas de tempo da base de tuítes, utilizamos também os resultados da análise de coesão em bases sintéticas [Gonçalves *et. al* 2023] para fundamentar a interpretação dos valores das métricas de coesão. As bases sintéticas foram construídas para serem "pouco conexa", "conexa" e "muito conexa". Uma base menos coesa gera uma rede de cliques menos conectada [Gonçalves *et. al* 2023].

4. Resultados e Discussão

Os resultados das análises foram interpretados à luz dos resultados de [Gonçalves *et. al* 2023] para bases com características conhecidas: pouco conexas (baixa coesão), conexas (coesa) e muito conexas (alta coesão).

A Figura 1(a) apresenta a variação de densidade, a partir do dia 16/01. Por volta de 9/03 a densidade esteve mais alta do que o índice para as bases muito conexas [Gonçalves *et. al* 2023], que foi próximo de 9. Esse resultado da rede é um índice que aponta para a alta coesão textual nesses períodos, mas pode estar relacionado ao número de tuítes. O total de tokens nessas janelas é muito grande, de modo que o denominador no cálculo da densidade, dado por n(n-1), é ainda maior no estado inicial. Mas os tuítes em geral são pequenos, de

modo que a quantidade de arestas (no numerador) é relativamente pequena. Assim, a densidade no estado inicial é muito pequena. Já a densidade do estado final fica muito maior: numa base muito coesa e grande, no estado final existe uma redução considerável do número de vértices, o que gera uma redução ainda maior no denominador da fórmula para Δ. Além disso, o número de arestas aumenta consideravelmente, também pelo fato da base ser coesa e grande. Com isso, temos uma densidade final muito maior que a densidade inicial, aumentando muito a variação da densidade.

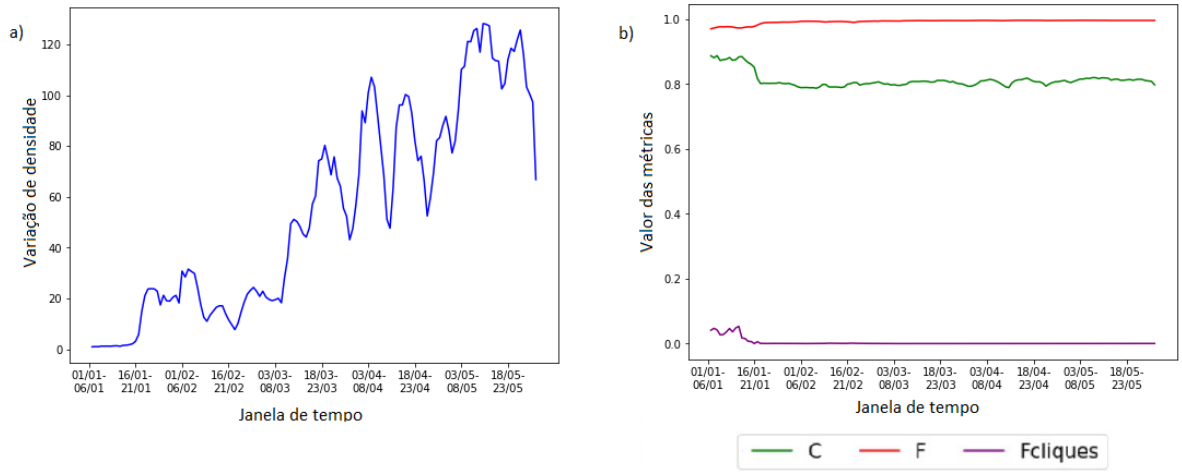

Figura 1. (a)Variação de densidade, (b) Coeficiente de clusterização (C), Fragmentação (F) e Fcliques nas janelas de tempo de 01/01/2020 a 31/05/2020.

O índice C ficou majoritariamente abaixo de 0,8 nos períodos onde a variação de densidade e de grau médio (Figura 2) foram muito altas, um resultado um pouco menor do que o valor deste índice para as bases sintéticas muito conexas. Para as bases sintéticas, conforme a coesão aumentava, observa-se que esse valor de C menor do que 0,8 nos períodos mencionados condiz com uma alta coesão textual, próxima daquela nas bases muito conexas.

Na Figura 1(b), observa-se que a fragmentação esteve sempre próxima de 1 (exceto nas janelas anteriores ao dia 16/01). Isso se deve, provavelmente, ao grande número de tuítes: numa base muito grande, o denominador no segundo termo no cálculo da fragmentação [Fadigas e Pereira 2013], dado por $n \cdot (n-1)$, é muito grande, mesmo após muitas sobreposições e justaposições; mas os cliques no estado inicial são muito pequenos, por se tratarem de tuítes. Com isso, o valor no numerador - que é dado por $\sum_{i=1}^{k}(n_i(n_i-1))$, onde k é o número de cliques no estado inicial e n_i é o número de vértices do clique i - fica relativamente pequeno. Então, o segundo termo fica muito próximo de 0 e a fragmentação tende a ser próxima de 1. Assim, ainda que a base apresentasse alta coesão textual, a fragmentação ainda seria próxima de 1. Esta só seria próxima de 0 se o número de vértices no estado final diminuísse mais ainda, numa base ainda mais conexa que se aproxima de ser um único grafo completo com uma quantidade relativamente pequena de vértices.

A fragmentação de cliques foi majoritariamente nula ou muito próxima de zero, o que significa que, na maioria dos casos, o estado final da rede de cliques apresentou um número de componentes conexos muito menor do que o número de cliques do estado inicial, o que é um indicador positivo de coesão.

Assim como $v(\Delta)$, a variação do grau médio (Figura 2) esteve muito maior do que aquela nas bases muito conexas, em períodos similares aos observados na variação de

densidade. Os picos observados foram em datas próximas a alguns eventos importantes relacionados à COVID-19 [Sanarmed 2020]: em 21/01 foi confirmada a primeira transmissão do novo coronavírus entre humanos; em 26/02 foi confirmado o primeiro caso de COVID-19 no Brasil; em 11/03, a OMS declarou oficialmente o início da pandemia da COVID-19; em 17/03 foi confirmada a primeira morte no Brasil; em 03/05, o Brasil passou de 100.000 mortes por COVID-19.

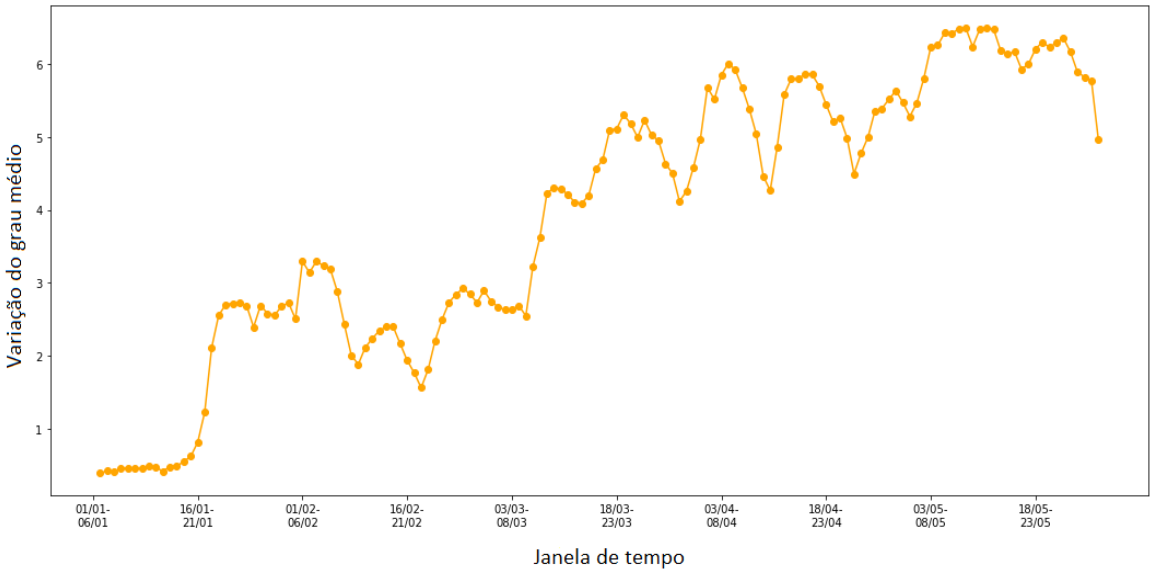

Figura 2. Variação do grau médio na base do Twitter analisada.

5. Considerações Finais

Este trabalho analisou a coesão textual lexical de tuítes relacionados a pandemia da COVID-19 publicados em português no primeiro semestre de 2020. Para tanto, foi aplicado o método VERSATILE [Gonçalves *et. al* 2023]. Os resultados proporcionaram uma compreensão de aspectos do discurso ao longo do período analisado e sua relação com os acontecimentos relacionados à COVID-19 no Brasil e no resto do mundo. Foi possível observar a progressão da coesão textual durante o período analisado, refletindo mudanças nos assuntos comentados e uma possível maior popularidade de certos assuntos do que de outros, a julgar pelos índices de coesão. A análise também se beneficiou das diversas métricas usadas, agregando na informação sobre o discurso ao longo do tempo.

Uma limitação do trabalho está relacionada à coleta de tuítes. Ainda que a base contenha um grande número de mensagens, ela está limitada a tuítes que continham algumas palavras-chave relacionadas diretamente à COVID-19 [de Melo e Figueiredo 2020]. Além disso, a base contém tuítes até 31/05, uma fração do período total de emergência em decorrência da pandemia de COVID-19. Como trabalhos futuros, pode-se analisar mensagens publicadas durante todo o período de duração da pandemia enquanto emergência global.

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, Volume 392, 10ª edição. https://doi.org/10.1016/j.physa.2013.01.055.

Fernandes, C. M.; de Oliveira, L. A.; de Campos, M. M.; Coimbra, M. R. (2020) “A Pós-verdade em tempos de Covid 19: o negacionismo no discurso de Jair Bolsonaro no Instagram”. *Liinc Em Revista*, *16*(2), e5317-e5317.

França, Tiago Cruz de. "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.

Gonçalves, J. M. N.; Oliveira, J.; Porto, F.; França, T. C. (2023) “Análise Temporal de Coesão de Discurso em Mídia Social Durante Grandes Eventos”. Anais do XII Brazilian Workshop on Social Network Analysis and Mining (pp. 234-239). SBC.

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

Lin, X.; Lachlan, K.A.; Spence, P.R. (2016) “Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on Twitter and Weibo”. Computers in Human Behavior, Volume 65, Pgs. 576-581, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2016.04.032.

Melo, T.; Figueiredo C.M.S. (2020) “A first public dataset from Brazilian twitter and news on COVID-19 in Portuguese”, Data in Brief, Volume 32, 106179, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2020.106179.

Muttaqien, M. Z. (2019) “Systemic cohesion in social media conversations: Cases on Facebook and Twitter”. Indonesian Journal of Applied Linguistics, 9, 413-423. doi: 10.17509/ijal.v9i2.20239

Neves, J. C. B. ; França, Tiago Cruz ; Bastos, M. P.; Carvalho, P. V. R.; Gomes, J. O. (2022) “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.

Neto, M.; de Oliveira Gomes, T.; Porto, F. R.; Rafael, R. D. M. R.; Fonseca, M. H. S.; Nascimento, J. (2020) “Fake news no cenário da pandemia de Covid-19”. *C*ogitare enfermagem, *25*.

Recuero, R. D. C. e Soares, F. B. (2021) “O Discurso Desinformativo sobre a Cura da covid-19 no Twitter: Estudo de caso”. *E-Compós: Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação. Brasília, DF. Vol. 24 (2021), p. 1-29.*

Saldanha, L. C. D. (2020) “O discurso do ensino remoto durante a pandemia de COVID-19”, Revista educação e cultura contemporânea, *17*(50), 124-144.

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

Tsao, S. F.; Chen, H.; Tisseverasinghe, T.; Yang, Y.; Li, L.; Butt, Z. A. (2021) “What social media told us in the time of COVID-19: a scoping review”. *The Lancet Digital Health*, *3*(3), e175-e194.