Mineração de textos aplicada a postagens do Twitter sobre Coronavírus: uma análise na linha do tempo

Text mining applied to Twitter posts on Coronavirus: an analysis in the timeline

Alexandre Ribeiro Afonso[a, *]

Cláudio Gottschalg Duque[b]

RESUMO: Este artigo descreve uma pesquisa sobre a mineração de postagens coletadas do Twitter, contendo duas palavras-chave: "Coronavírus" e "Brasil". O enfoque é a listagem das frequências dos substantivos (nouns), e a verificação de tais frequências como indicadores dos interesses de discussão, em quatro períodos de tempo: de fevereiro a junho de 2020. O método de pesquisa é quantitativo e envolve a coleta, filtragem, mineração dos textos e análise de resultados. Para a mineração de textos utiliza-se o algoritmo de clustering K-Means e, posteriormente, o software para análise de corpus AntConc. Conclui-se que o método aplicado sinaliza sobre os principais pontos de discussão e suas mudanças ao longo do tempo. Tais sinalizações poderiam contribuir para a criação de categorias de postagens mais detalhadas em uma posterior Análise de Conteúdo.

Palavras-chave: Mineração de Textos. Corpus. Twitter. Coronavírus. Brasil.

ABSTRACT: This article describes a research about the mining of posts collected from Twitter, containing two keywords: "Coronavírus" and "Brasil". The focus is on listing the frequencies of nouns, and verifying those frequencies as indicators about the interests of discussion, in four time periods: from February to June 2020. The research method is quantitative and involves the collection, filtering, text mining and analysis of results. In text mining, the K-Means clustering algorithm is used and, subsequently, AntConc corpus analysis software. It is concluded that the applied method signals about the main points of discussion and their changes over time. Such signs could contribute to the creation of more detailed categories of posts in a later Content Analysis.

Keywords: Text Mining. Corpus. Twitter. Coronavírus. Brasil.

[a] Research Expert Group for Intelligent Information in Multimodal Environment using Natural language Technologies and Ontologies.

[b] Programa de Pós-Graduação em Ciência da Informação, Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, Brasil.

[*] Correspondência para/*Correspondence to*: Alexandre Ribeiro Afonso. E-mail: rafonso.alex@gmail.com.

Recebido em/*Received*: 15/08/2020; Aprovado em/*Approved*: 06/11/2020.

INTRODUÇÃO

A extração de conhecimento (COSTA e GOTTSCHALG-DUQUE, 2007; DUQUE e LOBIN, 2004), partindo-se de grandes quantidades de dados textuais em mídias sociais digitais, tem sido realizada através da aplicação de algoritmos de coleta, filtragem, mineração e análise de dados, com a meta da percepção de padrões no mar de informação (e desinformação) onde se encontra atualmente a web.

Para a realização de tal extração de conhecimento, uma das etapas do processo é a mineração de dados, que consiste na aplicação de algoritmos de inteligência artificial para a exploração de quantidades massivas de dados. Quando os dados estão na forma textual (ou seja, dados não estruturados) utiliza-se a nomenclatura: mineração de textos. A mineração de textos consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente para objetivos específicos, tal como a tomada de decisões (ARANHA; PASSOS, 2006).

Sobre a coleta e uso de dados nas ciências humanas e sociais, os Métodos Digitais envolvem a exploração e pesquisa sobre dados digitais, originários da web, com a intenção de se chegar a conclusões em estudos típicos das humanidades, onde práticas similares já ocorriam antes do surgimento da web como fonte de dados de pesquisa: estudos de uso e usuários da informação, levantamento de opiniões de públicos, mapeamento de rumores e desinformação, entre vários outros tópicos (ROGERS, 2016). Paralelamente, os algoritmos de mineração de dados e inteligência artificial estão cada vez mais elaborados e acessíveis nas áreas de computação e engenharia. O fato é que a interação dessas vertentes, no Brasil, ainda é recente, e as primeiras discussões sobre o tema ainda ocorrem. O primeiro congresso internacional em humanidades digitais ocorreu em 2018 na cidade do Rio de Janeiro, e desde então as pesquisas e experimentações de junção das duas vertentes têm sido divulgadas mais fortemente e com um nome específico.

De acordo com Afonso e Duque (2019), observa-se que cada campo de conhecimento aborda a pesquisa em mídias sociais digitais sob um enfoque específico, de acordo com o interesse da linha de pesquisa do pesquisador (informacional, computacional, linguístico, etc.). É visível que um dos temas em ampla discussão atualmente no Twitter, abrange a propagação do novo Coronavírus e tópicos relacionados à pandemia mundial. Em diferentes línguas, o assunto é abordado em diversas perspectivas e formatos: encontram-se comentários opinativos, *retweets* de textos que almejam ser informativos (porém, muitos são de fontes desconhecidas), *retweets* de jornalistas e profissionais de comunicação, textos humorísticos, tabelas de dados sobre a pandemia, entre vários outros formatos.

Considerando a diversidade de postagens sobre o Coronavírus no Twitter, sem fazer qualquer distinção ou filtragem aos formatos existentes, mas selecionando postagens apenas em português, a pesquisa aqui descrita procura capturar os pontos mais comentados em quatro intervalos de datas diferentes: de fevereiro a junho de 2020. Foram utilizados pacotes de software conhecidos e aplicados em outras línguas, considerando como base principal a mineração de textos e a captura de frequências de palavras, que revelariam pontos de interesse por parte dos internautas em relação à pandemia.

Sabe-se que essa massa de dados poderia ser minerada e analisada sob diversas perspectivas, com diversos objetivos de levantamento e utilizando algoritmos e sistemas diversos, além dos aqui aplicados. Neste primeiro estudo, entretanto, os pontos de interesse sobre as postagens em cada intervalo de tempo e suas mudanças ao longo do tempo são o enfoque. A caracterização aprofundada dos discursos, suas

comparações e a análise de conteúdo sobre tais postagens seriam melhor trabalhadas futuramente.

TRABALHOS RELACIONADOS

Sobre trabalhos envolvendo a mineração de textos do tipo *clustering* e a análise de frequências em corpus, que são as duas técnicas utilizadas na pesquisa aqui descrita para textos em português brasileiro, alguns trabalhos similares podem ser listados. Procurou-se listar trabalhos relacionados aos procedimentos aqui relatados, com maior enfoque nos resultados da análise da informação e não na descrição detalhada do mecanismo computacional.

Nesse sentido, o trabalho de Afonso e Duque (2014) traz uma série de análises para os resultados de três algoritmos de *clustering,* sobre textos de artigos científicos e jornalísticos em português. O trabalho relata uma série de técnicas de filtragem de dados e a aplicação dos conhecidos algoritmos (*Simple K-Means*, *sIB* e *EM*). Os autores concluem que a aplicação do algoritmo *sIB* é a melhor escolha para ambos os formatos de textos analisados, com maior acerto no agrupamento, sendo 68,9% de acertos no agrupamento para textos jornalísticos e 77,8% para textos científicos.

Com o objetivo de demonstrar uma abordagem para caracterização de informações relevantes de eventos, o trabalho de Souza (2017) utiliza a extração de tópicos em dados compartilhados no Twitter. Avalia-se o desempenho de três métodos de aprendizagem de máquina (*K-Means*, *Latent Dirichlet Allocation*-LDA e *Non-Negative Matrix Factorization*-NMF) usados para extrair tópicos sobre as bases de dados da Operação Lava Jato e do processo de *impeachmen*t da então presidente do Brasil, em duas arquiteturas de pré-processamento diferentes (tradicional e com reconhecimento de entidade). Observa-se que as técnicas de pré-processamento têm influência direta sobre o resultado da extração de tópicos. O autor coloca que a técnica *Silhouette* ajuda a encontrar o melhor valor de *clusters* para uma determinada amostra de dados. Nos resultados, o NMF apresentou o melhor desempenho nas duas bases de dados, tanto na tarefa de extração de tópicos quanto no tempo de execução.

A descoberta de temas-chave em uma discussão de redes sociais é descrita por Klinczak e Kaestner (2017) em um trabalho sobre mineração de textos. Tal descoberta de conhecimento é caracterizada por grupos de termos relevantes restritos a um contexto, e o estudo de sua evolução ao longo do tempo. Para isso, utilizam-se procedimentos baseados em mineração de dados e processamento de texto. No início, as técnicas de processamento de texto são usadas para identificar os termos mais relevantes que aparecem nas mensagens de texto da rede social. Em seguida, esses termos são agrupados usando os algoritmos *K-Means* e *K-Medoids* clássicos, e também o recente algoritmo NMF (*Non-negative Matrix Factorization*). Finalmente, associam-se os termos mais relevantes dos agrupamentos de documentos para caracterizar os principais temas das mensagens consideradas.

Antunes et al. (2014) apresenta os resultados preliminares da pesquisa "Monitoramento de informação sobre doenças negligenciadas: o e-Monitor Dengue". Define o e-Monitor Dengue como um sistema de monitoramento de informação na Internet feito por meio de um mecanismo robô, software ou agente inteligente que vasculha os sites sobre dengue, disponíveis na Internet. Segundo o autor, dentre as mídias sociais, considera-se que o Twitter pode desempenhar um papel na gestão da informação ao permitir identificar usuários que podem atuar como filtro de informação, sendo possível acessar diretamente a informação mais relevante para uma

determinada área de interesse. Assim, a partir do monitoramento do Twitter, a primeira pergunta a ser respondida no âmbito da pesquisa foi "Quem fala sobre dengue?", outra pergunta a ser respondida foi "Quando se fala de dengue?". Observa-se que o número de *tweets* acompanha o crescimento do número de casos de dengue. Conclui-se que há indícios de uma relação entre os rumores sobre dengue e o aumento de número de casos notificados.

Ainda sobre a pesquisa em mídias sociais, Borba, Marinho e Caregnato (2017) apresentam uma análise altmétrica do termo "Repositório Institucional" no Twitter no período de 2009 a 2015. O estudo é de caráter quali-quantitativo, do tipo descritivo e utilizou a altmetria buscando auferir as postagens relacionadas ao termo, bem como realizar uma análise de conteúdo das mesmas, a fim de identificar as relações estabelecidas entre o termo e assuntos associados. Os resultados mostraram que as primeiras publicações sobre o termo "Repositório Institucional" em língua portuguesa no Twitter aparecem em 2009, sendo que a maior quantidade de *tweets* sobre o tema aconteceu em 2013. Destacou-se a prevalência da categoria "Citações sociais" e "Divulgação de Repositório Institucional", que, na maioria das vezes, menciona a implantação de repositórios. A *hashtag* mais encontrada foi #opendoar, que é o diretório oficial de repositórios acadêmicos de acesso a aberto, desenvolvido pela Universidade de Nottingham, no Reino Unido.

Um trabalho similar ao aqui relatado é sobre a descrição do fenômeno da referenciação, partindo-se de um corpus contendo postagens opinativas coletadas do YouTube. Na pesquisa realizada por Afonso e Té (2017), os vídeos selecionados descrevem ou comentam sobre o processo de *impeachment* da ex-presidente do Brasil Dilma Rousseff, iniciado no ano de 2015. Os autores descrevem como o objeto discursivo *impeachment* é colocado em formas ou expressões nominais pela composição de unidades morfossintáticas variadas. Este estudo estabelece relações com a informática, no que diz respeito à análise automatizada de sentimentos, partindo-se dos dados em mídias sociais.

Ao efetuar uma comparação com os três primeiros trabalhos e os três últimos, pode-se afirmar que este trabalho é similar aos três primeiros, os quais descrevem a aplicação dos algoritmos de *clustering* em grandes quantidades de textos de mídias sociais, e também é similar aos três últimos que fazem uma análise em relação à ocorrência de termos nas mídias sociais, ou seja, procura-se fazer a junção das duas metodologias de forma complementar.

PROBLEMA E QUESTÕES DE PESQUISA

O objetivo principal do trabalho foi localizar os pontos/tópicos de interesse em postagens sobre Coronavírus no Twitter, baseando-se para isso na frequência das palavras em tais postagens. Para localizar tais pontos de interesse, as ferramentas de análise e mineração de dados conhecidas e de livre uso são utilizadas, sendo aplicadas em postagens de quatro intervalos de tempo distintos:

a) 30 de janeiro de 2020 a 05 de fevereiro de 2020 (30.669 postagens)

b) 06 de fevereiro de 2020 a 04 de abril de 2020 (30.239 postagens)

c) 05 de abril de 2020 a 04 de maio de 2020 (26.235 postagens)

d) 05 de maio de 2020 a 19 de junho de 2020 (31.709 postagens)

Observe que as quatro bases de dados possuem quantidades de dias diferentes, isso ocorre pelo fato da divisão inicial ser por quantidade de postagens, ou seja, procurou-se balancear as bases por quantidades de postagens aproximadas, e posteriormente verificar as datas de limite de cada base. É perceptível que a quantidade de *tweets* na base de janeiro a fevereiro é um pouco maior que as duas bases seguintes, e também abrange menos dias.

Para cada corpus, em cada um dos intervalos de tempo descrito, procurou-se descobrir os tópicos de maior interesse, baseando-se, para isso, na frequência dos substantivos (*nouns*) presentes nos *tweets*. Para tal tarefa, aplicam-se algoritmos específicos de aprendizagem de máquinas não supervisionada e calcula-se a frequência de determinados substantivos. Além de verificar os pontos de interesse em cada intervalo de tempo pela frequência dos substantivos, também verificou-se a mudança dos substantivos de maior frequência ao longo do tempo. Nessa perspectiva, foram elaboradas três questões de pesquisa:

a) Quais os principais substantivos (etiqueta *noun*), considerando-se a frequência de uso, em cada intervalo de tempo registrado?

b) O que seria possível concluir sobre as frequências de substantivos coletadas e analisadas para cada período citado, além de suas modificações e manutenções ao longo dos períodos considerados?

c) A partir da metodologia de: filtragem, mineração e análise de dados aplicada, seria possível afirmar que tal metodologia revela padrões significativos nos corpora de estudo?

METODOLOGIA E FERRAMENTAS

Utilizando-se a API (*Application Programming Interface*) do Twitter foram coletadas as postagens de 30 de janeiro de 2020 à 19 de junho de 2020 que continham o padrão "Coronavírus" junto ao padrão "Brasil", ambos deveriam estar na mesma postagem, incluindo a possibilidade dos padrões conterem letras em maiúsculo ou minúsculo, e grafado sem acentuação. Foram coletados 118.852 *tweets* com os padrões de coleta citados. Feito isso, os *tweets* foram divididos em quatro períodos de data, como especificado anteriormente, e as seguintes ações executadas para cada corpus, de cada período:

1) Exclusão de postagens em outros idiomas e *retweets* do corpus de análise;

2) Exclusão das palavras-chave originais de coleta ("Coronavírus" e "Brasil");

3) Etiquetagem morfossintática das postagens e extração apenas de substantivos (etiqueta *noun*);

4) Pesagem *tf-idf* sobre cada postagem;

5) Aplicação do algoritmo *K-Means* com número de grupos igual a vinte;

6) Coleta dos vinte substantivos mais frequentes no grupo gerado mais concentrado;

7) Coleta dos substantivos mais frequentes dentre os vinte grupos gerados pelo *K-Means*;

8) Escolha de alguns substantivos para análise de frequência lateral utilizando o sistema *AntConc*.

Sobre as ações 1, 2 e 3, para cada corpus, os *retweets* foram desconsiderados, pois o objetivo seria quantificar pontos de interesse e não quantificar as propagações de uma mesma postagem. As palavras-chave de coleta "Coronavírus" e "Brasil" foram retiradas de cada *tweet* em cada *corpus*, obviamente, para não serem consideradas no processo de mineração de textos (*clustering*), já que estariam presentes em todas as postagens coletadas.

A etiquetagem morfossintática foi realizada por um etiquetador (*PoS Tagger*) que coloca etiquetas como (*noun*) para substantivos ou (*verb*) para verbos, considerando o português brasileiro no processo. Os substantivos foram considerados como os elementos que guiam a temática da postagem, pois o internauta que comenta ou argue sobre algo (um objeto, um evento, um fato) precisa, de alguma maneira no discurso, identificar ou nomear esse algo. Ou seja, a entidade sobre quem se escreve no texto de postagem (o referente) em algum momento do texto será nomeada ou descoberta na leitura, o substantivo é uma maneira (comum) de nomear, porém, nem sempre o referente será identificado por apenas um substantivo comum, como: *governo*, *pandemia* ou *vírus*. A identificação do referente poderia ocorrer na forma composta, como um sintagma (por exemplo, *vírus da nova gripe* ou *vírus da pandemia*), ou ainda ele só poderá ser identificado pela interpretação de constituintes mais complexos, onde algum tipo de inferência é necessário para identificação, por exemplo: "O microrganismo que adoece, mata e desde fevereiro estaria no Brasil". Apesar do substantivo, em única palavra, nem sempre identificar completamente o referente, em muitos casos ele estará presente na composição de constituintes maiores: nos exemplos citados, *vírus* e *microrganismo* fazem parte do identificador do referente, mas não o identificam completamente. Por essa razão, em uma investigação primária, considerou-se o substantivo (etiqueta *noun*) como elemento de identificação dos pontos de interesse no corpus. Observe também, que para o etiquetador utilizado, substantivos (etiqueta *noun*) são diferentes de nomes próprios (etiqueta *propn*) e esta última classe não foi coletada e analisada, neste trabalho.

Para o entendimento inicial do fenômeno da referenciação e a relação com as expressões nominais do texto, na perspectiva da Linguística Textual, veja o trabalho de Koch (2008). Para uma abordagem sobre o fenômeno da referenciação em postagens de mídias sociais e com vistas à análise de sentimentos, os estudos descritos em Afonso (2017) e Afonso e Duque (2019) abordam a referenciação e as (re)nomeações em comentários de vídeos do YouTube.

Durante as ações 1, 2 e 3, foi utilizada a linguagem de programação *Python* e suas bibliotecas de Processamento de Linguagem Natural, como a biblioteca *spaCy*. Para as ações 4 e 5, foi utilizada a biblioteca *sklearn*, também da linguagem de programação *Python*. A pesagem *tf-idf* é realizada para cada substantivo de cada postagem, em seguida o algoritmo de *clustering K-Means* é aplicado para cada corpus. Testou-se o algoritmo para cada corpus, com três números de grupos (10, 15 e 20), o que gerou agrupamentos de postagens muito semelhantes: todos eles com altas concentrações em três ou quatro grupos e com frequências de palavras aproximadas nesses grupos mais concentrados. Optou-se então por adotar o número de grupos *k* igual a vinte, por permitir maior distribuição dos dados. A alta concentração dos dados em poucos grupos já era esperada, uma vez que em cada período específico as postagens sobre o Coronavírus possuem palavras similares.

O modelo treinado para etiquetagem morfossintática, utilizado com o *spaCy*, é chamado *pt_core_news_lg*. O algoritmo *K-Means* foi escolhido por ser conhecido e muito utilizado em mineração de dados e textos, o mesmo ocorre com a pesagem *tf-idf*, mas outros algoritmos e pesagens podem ser aplicados e testados.

A coleta de frequências de *tokens* ao lado dos substantivos mais frequentes (ação 8), e encontrados nos grupos gerados pelo *K-Means,* foi realizada pelo software de análise de corpora *AntConc*, sistema computacional descrito por Kader e Richter (2013), como uma ferramenta de análise para estudos linguísticos sob uma perspectiva empírica.

Neste trabalho, o objetivo de uso do *AntConc* foi verificar em que vizinhança de outras palavras esses substantivos mais frequentes ocorrem, procurando obter um melhor entendimento do uso de tais substantivos em postagens. O *AntConc* é conhecido nas pesquisas em linguística de corpus, tanto para busca de padrões quanto para a compilação de corpora. Seu uso é eficaz na identificação e quantificação de frequências de *tokens* em grandes bases textuais, e as técnicas e possibilidades de resultados deste campo têm também ampla possibilidade de aplicação e uso na Ciência da Informação, como relatado em Bowker (2018).

É importante salientar que apesar das técnicas e ferramentas aqui utilizadas serem já conhecidas e utilizadas em outras situações, a maneira como foram aplicadas suas junções é experimental, e uma das questões de pesquisa é justamente verificar se essas junções geram percepções, por uma abordagem quantitativa, sobre os pontos de interesse em discussão dos internautas.

RESULTADOS E ANÁLISE DE RESULTADOS

Os resultados estão listados a seguir, na forma de respostas às questões de pesquisa levantadas.

1. Quais os principais substantivos (etiqueta *noun*), considerando-se a frequência de uso, em cada intervalo de tempo registrado?

Para levantar essas frequências, aplicou-se a filtragem, que é a etapa onde os *retweets* foram excluídos, e somente a etiqueta *noun* é considerada como item de agrupamento. Em sequência, o algoritmo de *clustering K-Means* é executado com vinte grupos.

A tabela 01, a seguir, lista os vinte substantivos mais frequentes do grupo mais concentrado. Somente um grupo foi escolhido, pelo fato da concentração ocorrer de maneira bastante desproporcional em apenas três ou quatro grupos. Considerando que o objetivo é capturar os substantivos mais frequentes, a coleta descrita na tabela 01 analisa o grupo mais concentrado, com mais postagens. Como os pontos de discussão por parte dos internautas são similares em um intervalo de tempo, isso explicaria a concentração em poucos grupos.

Tabela 01 – Ranking de ocorrências de substantivos do grupo com maior número de postagens, após execução do algoritmo de *clustering*.

	***Corpus* 01:** **30/01 a 05/02**	***Corpus* 02:** **06/02 a 04/04**	***Corpus* 03:** **05/04 a 04/05**	***Corpus* 04:** **05/05 a 19/06**
1	(casos, 483)	(casos, 407)	(mortes, 513)	(mortes, 914)
2	(saúde, 192)	(mortes, 302)	(casos, 463)	(casos, 675)
3	(governo, 165)	(saúde, 252)	(pandemia, 292)	(pandemia, 526)
4	(brasileiros, 162)	(pandemia, 215)	(mundo, 228)	(país, 458)

5	(caso, 152)	(novo, 203)	(saúde, 218)	(covid19, 448)
6	(carnaval, 137)	(país, 178)	(país, 211)	(mundo, 360)
7	(mundo, 128)	(pessoas, 163)	(pessoas, 211)	(saúde, 359)
8	(pessoas, 123)	(mundo, 159)	(novo, 192)	(novo, 292)
9	(país, 118)	(governo, 157)	(número, 182)	(número, 290)
10	(novo, 108)	(dia, 154)	(covid19, 174)	(pessoas, 282)
11	(surto, 107)	(presidente, 146)	(dia, 159)	(presidente, 276)
12	(quarentena, 105)	(bolsonaro, 135)	(presidente, 159)	(governo, 247)
13	(dias, 94)	(combate, 132)	(mortos, 157)	(dia, 233)
14	(emergência, 90)	(número, 129)	(bolsonaro, 156)	(óbitos, 227)
15	(número, 83)	(covid19, 114)	(governo, 152)	(mortos, 226)
16	(epidemia, 79)	(países, 99)	(combate, 149)	(bolsonaro, 217)
17	(doença, 79)	(testes, 95)	(crise, 136)	(combate, 197)
18	(hospital, 79)	(crise, 91)	(óbitos, 131)	(horas, 175)
19	(dengue, 62)	(caso, 90)	(países, 109)	(dados, 162)
20	(gente, 62)	(mortos, 87)	(testes, 107)	(países, 153)

Para capturarmos as distribuições dos substantivos nos vinte grupos existentes, e não relatar somente a frequência em um único grupo, foram coletados nos corpora de cada intervalo de tempo a frequência destes substantivos em grupos, ou seja, quantificou-se a presença dos substantivos nos vinte grupos para cada corpus. A tabela 02, a seguir, mostra tais resultados.

Tabela 02 – Resultado da coleta dos substantivos mais frequentes dentre os vinte grupos gerados pelo *K-Means*.

	***Corpus* 01: 30/01 a 05/02**	***Corpus* 02: 06/02 a 04/04**	***Corpus* 03: 05/04 a 04/05**	***Corpus* 04: 05/05 a 19/06**
1	(casos, 18)	(casos, 19)	(pandemia, 18)	(mortes, 19)
2	(saúde, 18)	(saúde, 19)	(casos, 18)	(covid19, 18)
3	(quarentena, 16)	(país, 19)	(saúde, 17)	(país, 18)
4	(governo, 16)	(mortes, 19)	(mortes, 17)	(novo, 18)
5	(país, 16)	(pandemia, 17)	(país, 17)	(mundo, 18)
6	(brasileiros, 16)	(novo, 15)	(covid19, 16)	(casos, 18)
7	(carnaval, 15)	(mundo, 14)	(mundo, 16)	(pandemia, 17)
8	(caso, 15)	(mortos, 13)	(número, 16)	(número, 17)
9	(surto, 14)	(pessoas, 13)	(pessoas, 15)	(saúde, 17)
10	(hospital, 13)	(presidente, 13)	(novo, 15)	(óbitos, 17)

11	(emergência, 12)	(dia, 13)	(dia, 14)	(horas, 16)
12	(mundo, 12)	(governo, 13)	(governo, 14)	(pessoas, 16)
13	(pessoas, 12)	(bolsonaro, 12)	(presidente, 14)	(mortos, 15)
14	(dias, 12)	(número, 11)	(óbitos, 13)	(dia, 15)
15	(gente, 11)	(covid19, 11)	(mortos, 13)	(presidente, 14)
16	(novo, 11)	(caso, 10)	(bolsonaro, 12)	(bolsonaro, 14)
17	(número, 10)	(combate, 9)	(crise, 12)	(governo, 12)
18	(doença, 10)	(ministério, 8)	(combate, 10)	(dados, 10)
19	(epidemia, 9)	(crise, 8)	(testes, 9)	(combate, 10)
20	(ministério, 6)	(janeiro, 7)	(horas, 9)	(países, 9)

Para um melhor entendimento do uso de alguns substantivos citados anteriormente, o software de análise de corpora *AntConc* foi aplicado nos textos de cada corpus, sem os *retweets.*

Não sendo possível descrever a vizinhança de palavras de todos os substantivos coletados nas tabelas 01 e 02 em um único artigo, selecionou-se então alguns deles que, para os pesquisadores que atuaram neste trabalho, são de maior interesse. Por exemplo, para o substantivo *carnaval* do corpus 01, as frequências de *tokens* ao lado deste substantivo seriam descobertas pelo uso do sistema *AntConc,* especificamente através do seu módulo *Collocates,* o qual permite visualizar as frequências de *tokens* laterais para um substantivo, considerando um número máximo de *tokens* à esquerda e à direita do substantivo *carnaval.* Tal intervalo, o sistema chama de *Window Span*, e foram pesquisados os cinco *tokens* mais à esquerda e mais à direita do substantivo de interesse. O *AntConc* também permite medir o grau de associação de substantivos, como *carnaval,* com outros *tokens*, o que em linguística de corpus chama-se de *Stat*, porém, essa medida estatística não baseia-se somente na frequência e tais medidas não foram o foco deste trabalho.

Obviamente, classes gramaticais como os artigos definidos (*o*, *a*, *os*, *as*), preposições e contrações de preposições, como (*de*, *do*), apareceriam com alta frequência ao lado do substantivo de pesquisa considerado: *carnaval.* Portanto, na tabela seguinte descrevemos, somente para cada substantivo de interesse, as frequências que são significativas.

Tabela 03 – Ocorrências de *tokens* ao lado de alguns substantivos em destaque nas tabelas 01 e 02 anteriores.

Substantivo e *corpus* de origem.	*Token* e soma de ocorrências à esquerda e à direita do substantivo considerado.
carnaval (corpus 01)	cancelar (30); risco (13); cancelado (11);cancelamento (10)
emergência (corpus 01)	estado (45); decretar(29); declara(28)
dengue (corpus 01)	febre (13); amarela(13); grave(11)
hospital (corpus 01)	china(80); construiu(52); leitos(51)
ministério (corpus 02)	saúde (262); casos(48); confirmados (36)
governo (corpus 02)	federal (42); bolsonaro(34); congresso(13)
crise (corpus 02)	econômica (12); enfrentar (10); bolsonaro (10)
janeiro (corpus 02)	em(165); de(102); brasil(127); chegou(45)
governo (corpus 03)	federal (71); bolsonaro (52); isolamento (12)
crise (corpus 03)	bolsonaro (26); política (22); econômica (17)
horas (corpus 03)	mortes (139); últimas(114); óbitos(31); total(30)
governo (corpus 04)	federal (90); bolsonaro (70); estado (13)
dados (corpus 04)	saúde(60); ministério(38); segundo (28)
países (corpus 04)	mundo (23); eua (18); américa (14); sul (14)

2. O que seria possível concluir sobre as frequências de substantivos coletadas e analisadas para cada período citado, além de suas modificações e manutenções ao longo dos períodos considerados?

Através da análise das tabelas 01 e 02 pode-se verificar que o corpus 01 traz substantivos que mais se diferenciam dos substantivos dos demais corpora: *carnaval*, *emergência*, *dengue* e *hospital* são alguns desses substantivos diferenciais. Os corpora 02, 03 e 04 registram muitos substantivos em comum. Algumas constatações sobre as coletas são as seguintes:

- O substantivo *mortes* não aparece nas listagens de frequência do corpus 01
- O nome *Covid19* não aparece na listagem do corpus 01 mas aparece nos outros corpora
- Substantivos e nomes relacionados ao governo federal, como: *governo*, *presidente* e *Bolsonaro* aparecem sempre com alta frequência nos quatro corpora
- Substantivos relacionados a levantamentos como *número* e *dados* aparecem também com alta frequência em todos os corpora
- Apesar da alta semelhança entre os substantivos de maior frequência nos corpora 02, 03 e 04, observa-se que o substantivo *ministério* aparece apenas nos corpora 01 e 02
- O substantivo *crise* aparece somente nos corpora 02 e 03
- O nome *janeiro* e o substantivo *dados* aparecem somente nos corpora 02 e 04, respectivamente.

Em relação à tabela 03, pode-se observar a vizinhança em que alguns substantivos aparecem com mais frequência, esses *tokens* laterais situam o uso do substantivo na postagem. Por exemplo, o substantivo *carnaval* tem como *tokens* laterais de alta frequência: *cancelar* e *risco.* Da mesma maneira, o substantivo *governo* se relaciona principalmente com o *token: federal,* que tem alta frequência, mas o *token: estado,* com menor frequência, também é presente. O substantivo *crise* aparece juntamente aos *tokens: econômica* e *política*, como elementos laterais.

O nome *janeiro* no corpus 02 tem a preposição "*em*" lateral e à esquerda com alta frequência, e ocorre também a preposição "*de*" na mesma posição e com alta frequência. A coleta pode estar indicando tanto que *janeiro* refere-se ao mês de *janeiro* ("em janeiro"), como a cidade ou estado do *Rio de Janeiro*. A princípio, verifica-se que a preposição "*em*" tem maior frequência e, logo, indicaria *janeiro* (mês do ano) como mais frequente. Considera-se que outros tipos de análise são necessários para uma investigação mais detalhada a essa situação, para verificar qual dos três nomes é mais frequente (mês, cidade ou estado) e em que situação de uso ocorrem.

3. A partir da metodologia de: filtragem, mineração e análise de dados aplicada, seria possível afirmar que tal metodologia revela padrões significativos nos corpora de estudo?

Como observado na prática realizada, os sistemas computacionais contribuíram na limpeza, filtragens, localização de padrões, agrupamento de postagens, descoberta de similaridades em postagens e cálculo de frequências. Os sistemas computacionais realizaram tais tarefas em um tempo muito curto, se comparado a um processo manual que levaria meses para realização. Contudo, o trabalho de agrupamento de postagens similares e o cálculo de frequências em postagens, para descoberta de pontos de discussão, é apenas uma parte do trabalho. O entendimento do que se escreve sobre uma entidade como "governo federal", e a captura das várias possibilidades de discursos opinativos e informativos, talvez só possam ser mapeados e categorizados com uma análise de conteúdo realizada manualmente, por um humano.

A língua portuguesa permite que um mesmo conteúdo seja descrito em formas totalmente diferentes, utilizando o amplo léxico da língua e suas estruturas gramaticais em combinações muito variadas, e nesse sentido, a percepção humana ao julgar equivalências e similaridades nos discursos é bem mais capacitada. Ainda, a descoberta de grupos (ou categorias) de postagens similares talvez só seja precisa quando ocorre a leitura e releitura das postagens e a interpretação dos textos na forma integral, não considerando somente substantivos e outras classes de palavras.

Em contrapartida, não há dúvidas que a técnica de *clustering* e a análise de frequências apoiada em software conseguem efetuar uma pré-análise e capturar uma série de pistas sobre o conteúdo das postagens, e através desse resultado primário da busca quantitativa, muito do que foi coletado e analisado já se pode considerar como extração de conhecimento.

Deve-se observar também, que os sistemas computacionais não são infalíveis. Por exemplo, o nome *Bolsonaro* deveria ter uma *PoS Tag* do tipo *proper name,* e não deveria estar com a etiqueta *noun* como detectado pelo sistema utilizado. Considere também, que os dados linguísticos de mídias sociais são difíceis de serem processados, uma vez que nesse ambiente a utilização do vocabulário informal é constante.

CONSIDERAÇÕES FINAIS

A pesquisa descrita aplicou a metodologia quantitativa de coleta de frequências de substantivos e apoiada em software, com a meta de verificar se tal coleta revelaria os principais pontos de interesse dos usuários do Twitter, ao postarem utilizando as palavras "Coronavírus" e "Brasil".

O trabalho realizado procurou ter um entendimento global dos dados, sem entrar em detalhes sobre formatos, discursos e tipos específicos de postagens. Este objetivo inicial tem a intenção de colocar uma ordem nos dados, de maneira a contribuir com um entendimento geral e direcionar em recortes de pesquisa futuros, e nesse sentido, a metodologia adotada gerou os resultados esperados, possibilitando que novas questões sejam levantadas.

Os resultados mostram que métodos puramente quantitativos revelam padrões, que por si só podem gerar conhecimento sobre os dados coletados. Porém, uma análise de conteúdo mais detalhada sobre os dados poderia gerar afirmações mais consistentes, já que um pesquisador humano é capaz de propor relações mais complexas entre os discursos.

O método quantitativo permite também um ganho de tempo considerável, por realizar filtragens nos dados e gerar apontamentos ou indicadores, para uma posterior análise qualitativa. Deve-se considerar também que existem várias técnicas e mecanismos computacionais com características específicas que ainda podem ser aplicados.

A junção destas vertentes (quantitativo e qualitativo) é uma tendência nas pesquisas em mídias sociais, e como observado em outros trabalhos referenciados, a união dos mecanismos computacionais com alguns métodos e interesses das humanidades tem sido cada vez mais forte, campo que tem sido denominado Humanidades Digitais.

Em trabalhos futuros, pesquisas com outros algoritmos e capturando postagens com características mais específicas devem ser realizadas, e espera-se obter novas visualizações a partir de outras perspectivas.

REFERÊNCIAS

AFONSO, A. R. A referenciação em textos do YouTube: um estudo com vistas à análise de sentimentos. *Liinc em Revista,* v. 13, n. 2, 2017.

AFONSO, A. R.; DUQUE, C. G. Análise de sentimentos em comentários de vídeos do YouTube utilizando aprendizagem de máquinas supervisionada. *Ciência da Informação,* v. 48, n. 3, 2019.

AFONSO, A. R.; DUQUE, C. G. Automated text clustering of newspaper and scientific texts in brazilian portuguese: analysis and comparison of methods. *JISTEM,* São Paulo, v.11, n.2, p. 415-436, ago. 2014.

AFONSO, A. R.; TÉ, J. Um estudo sobre referenciação e a construção da opinião a partir de um corpus textual extraído do YouTube. *Domínios de Lingu@gem,* v. 11, n. 2, p. 339-350, 27 mar. 2017.

ANTUNES, M. N. *et al.* Monitoramento de informação em mídias sociais: o e-Monitor Dengue. *TransInformação,* Campinas, v. 26, n. 1, p. 9-18, 2014.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação,* v. 5, n. 2, 2006.

BORBA, V. R.; MARINHO, A. C. M.; CAREGNATO, S. Análise do termo "Repositório Institucional" no twitter: um estudo altmétrico. *Em Questão,* v. 23, n. 5, p. 290-308, 2017.

BOWKER, L. Corpus linguistics is not just for linguists: considering the potential of computer-based corpus methods for library and information science research. *Library Hi Tech,* v.36, n.2, 2018.

COSTA, S. M. S.; GOTTSCHALG-DUQUE, C. Towards an ontology of ElPub/SciX: a proposal. *In:* INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, 11., 2007, Viena. *Proceedings…* Viena: ÖKK-Editions, 2007. V. 1. P. 249-256.

DUQUE, C. G.; LOBIN, H. Ontology extraction for index generation. *In:* ICCC - INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, 8., 2004, Brasília. *Proceedings...* Brasília: ELPUB, 2004. p. 111-120.

KADER, C. C. C.; RICHTER, M. G. Linguística de corpus: possibilidades e avanços. *Instrumento,* v. 15, n. 1, p. 13-23, jan./jun. 2013.

KLINCZAK, M. N. M.; KAESTNER, C. A. Identificação de temas em redes sociais por meio de técnicas de agrupamento. *Anais do Computer on the Beach,* p. 090-099, 2017.

KOCH, I. V. Como se constroem e se reconstroem os objetos-de-discurso. *Investigações,* Recife, v. 21, n. 2, p. 99-114, 2008.

ROGERS, R. O fim do virtual: os métodos digitais. *Lumina,* v. 10, n. 3, 2016.

SOUZA, B. A. *Uma abordagem para seleção de tópicos relevantes em redes sociais online.* 2017. Dissertação de Mestrado (Programa de Pós-Graduação em Informática do Instituto de Computação) - Universidade Federal do Amazonas, Manaus, 2017.