IDENTIFICAÇÃO DE PADRÕES DE TWEETS: UM ESTUDO DE CASO SOBRE A COVID-19 NO BRASIL

Bruna Pereira Paz[1], Leila Cristina C. Bergamasco[2]
[1,2]Departamento de Ciência da Computação, Centro Universitário FEI
unifbpaz@fei.edu.br leila.cristina@fei.edu.br

Resumo: A partir de março de 2020, a população passou a conviver com a pandemia do Coronavírus (COVID-19) e tem tentado, por meio dos dados gerados encontrar soluções. Um dos aspectos estudados tange à reação social e psicológica da sociedade perante tal fenômeno, que pode ser expressa por meio de redes sociais. Este projeto de pesquisa busca criar a primeira base de dados referente as interações no Twitter sobre a doença do COVID-19 e identificar associações dos conteúdos publicados antes e após o início da vacinação da população brasileira.

1. Introdução

A doença Coronavírus (COVID-19) causada pelo vírus de síndrome severa aguda coronavírus 2 (*severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)*) foi anunciada oficialmente pela Organização Mundial de Saúde (OMS) em fevereiro de 2020 e em março de 2020, e declarou estado de pandemia devido à doença. Desde então mais de 180 milhões de pessoas foram infectadas e mais de 4 milhões de pessoas vieram à óbito [1]. O Brasil ocupa a 7a posição entre países com maior mortalidade, com 2542 mortes por milhão de habitante [2].

COVID-19 é a primeira pandemia após o início da transformação digital da sociedade, a partir do século XX, e dessa forma, uma variedade de dados relacionados à doença está sendo disponibilizados ao público e pesquisadores, fomentando diversas aplicações incluindo previsão, planejamento, gestão, tomada de decisão e rastreabilidade da doença [3]. Além de pesquisas focadas em aspectos diretamente associadas à COVID-19, nota-se um interesse sobre as consequências psicológicas e sociais geradas pela doença. Nesse sentido, redes sociais como Facebook, Twitter e Instagram são uma fonte rica de dados. O Twitter possui mais de 350 milhões de usuários ativos e gera por dia mais de 200 bilhões de *tweets* por dia, sendo a 7ª rede social mais utilizada no Brasil.

Observa-se, portanto, um ao alto volume de dados textuais gerados diariamente e oriundo de diversas fontes, incluindo as redes sociais. Sendo necessário o desenvolvimento de soluções computacionais para extrair informações e padrões relevantes de tal massa de dados. Tais técnicas pertencem ao domínio de Mineração de Texto (MT) que, por sua vez, é uma das estratégias de mineração dentro do contexto de Mineração de Dados.

2. Metodologia

Para a construção do projeto serão utilizados conceitos da mineração de Texto, que pode ser definida como sendo a descoberta não-trivial de padrões ocultos ou desconhecidos e potencialmente úteis em textos. Tais padrões podem oferecer informações uteis em contextos específicos e, diferentemente de outros tipos de dados, os textos muitas vezes são não-estruturados, ambíguos e de difícil processamento [4]. Para trabalhar com os dados serão divididos alguns aspectos da MT em blocos a serem trabalhados durante o projeto (Figura 1).

Figura 1 – Integração da área de MT com diferentes domínios de conhecimento.

Dado que o objetivo do presente projeto de pesquisa é encontrar padrões sobre o comportamento social dos usuários do Twitter referente a pandemia do COVID-19 e, consequentemente, fornecer para a comunidade científica uma base de dados inédita, com dados a partir do início da vacinação da população, grande parte dos esforços envolvidos nesse trabalho se darão na etapa de pré-processamento dos *tweets* recuperados. Dessa forma durante o projeto, serão avaliadas as técnicas computacionais pertinentes a cada uma das etapas de pré-processamento como a *tokenização* e remoção de *stopwords.* Adicionalmente o algoritmo Apriori será utilizado para descobrir associações frequentes e válidas no conteúdo analisado.

Será utilizada API do Twitter para recuperar os *tweets* relacionados ao tema do presente projeto de pesquisa. Adicionalmente será utilizada a linguagem Python e suas bibliotecas *numpy, nltk e apyori* para o desenvolvimento e análise dos resultados.

$$Suporte(A) = \frac{NA}{NT}$$
$$Confiança(A \rightarrow B) = \frac{Suporte(A \cup B)}{Suporte(A)}$$

Figura 2 – Equações de suporte e confiança.

Para avaliar a qualidade das regras encontradas, serão utilizadas as métricas de suporte e confiança, descritas nas equações acima (Figura 2), respectivamente. O termo *NA* se refere ao número de

transações que *A* ocorre e *NT* é o número total de transações

3. Resultados Parciais

Até o momento foram realizadas revisões na literatura a respeito de métodos de coleta e armazenamento de massa de dados, envolvendo majoritariamente linguagem Python.

Para realizar a coleta de dados de sua base, o Twitter fornece duas aplicações: a *Representational State Transfer* (REST) API e a Streaming API. As API's, permitem que desenvolvedores acessem e coletem amostras do que foi publicado na plataforma. Para a extração da massa foi utilizada a linguagem Python com a biblioteca Tweepy, que fornece acesso a todos os métodos da API RESTful do Twitter.

As etapas para coleta consistem em:

1. Conexão com a API do Twitter pela conta de desenvolvedor previamente criada;
2. Busca de dados por meio de palavras-chaves utilizadas no corpo do tweet publicado que correlacionam com a COVID-19. Exemplos de palavras – chaves utilizadas foram: "covid", "pandemia" e "coronavírus". Adicionalmente as palavras-chaves, a busca considerou o intervalo de tempo pré vacina e pós vacina, entre 03/2020 e 02/2021.
3. Conversão dos dados do formato JSON em formato .csv (*Comma Separated Values*).
4. Retirada das *stopwords (exemplo: de, para, da, as..)* dos tweets extraídos para melhor analise das palavras.

Após a extração dos dados foi criado um *word cloud (nuvem de palavras),* que consiste em reunir as palavras mais utilizadas na base de dados formada pelos *tweets* extraídos. A Figura 3 apresenta os termos mais frequentes utilizados entre março/2021 e fevereiro/2021, um período no qual havia muita incerteza sobre as vacinas.

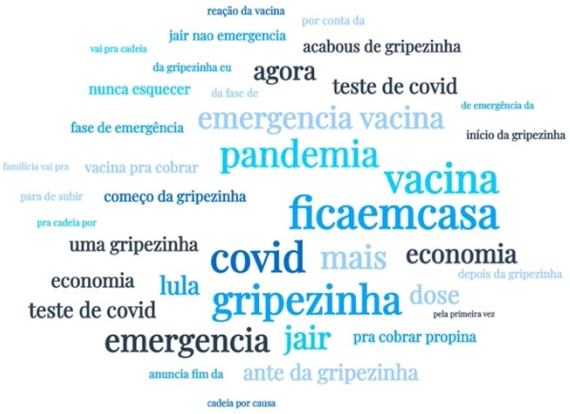

Figura 3 – Nuvem de palavras mais encontradas entre março/2020 e fevereiro/2021

Nota-se em uma análise exploratória do resultado obtido que houve uma grande repercussão sobre as aplicações dos testes de covid e a criação das vacinas, seguidas das preocupações sobre o *lockdown,* políticas públicas e a iniciativas como o *#FicaEmCasa*.

A Figura 4 apresenta uma *word cloud* criada no intervalo pós vacina (após fevereiro de 2021) e nota-se uma mudança nos termos encontrados.

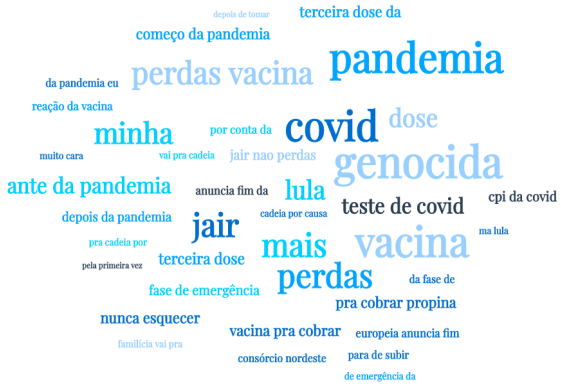

Figura 4 – Nuvem de palavras mais encontradas após inicio da vacinação.

A vacina ainda é muito citada, mas o teor político domina nos assuntos discutidos no momento pós vacinação, pode-se notar também uma grande preocupação sobre as perdas causadas pela covid e os efeitos que a gestão do presidente Jair Bolsonaro teve em relação à pandemia.

4. Etapas Futuras

As próximas etapas a serem percorridas no projeto serão focadas no desenvolvimento das regras de associação utilizando o algoritmo Apriori, assim como a análise e documentação dos resultados obtidos durante todo o processo de construção do projeto de iniciação cientifica.

5. Referências

[1] SAUDE, Organização Mundial de. Coronavirus (COVID-19) Dashboard. [S.l.: s.n.], 2020.

[2] DATA, Our World in. Covid Deaths. [S.l.: s.n.], 2020.

[3] ALSUNAIDI, Shikah J et al. Applications of Big Data Analytics to Control COVID-19 Pandemic. Sensors, Multidisciplinary Digital Publishing Institute, v. 21, n. 7, p. 2282, 2021

[4] WAEGEL, Daniel B. The Development of Text-Mining Tools and Algorithms. In.

Agradecimentos

À instituição FEI, seu corpo docente, direção e administração que oportunizaram a construção deste projeto.

[1] Aluno de IC do Centro Universitário FEI. Projeto com vigência de 02/2022 a 02/2023.