CONSTRUÇÃO DO CORPORA CALENDÁRIO BRASILEIRO DE SAÚDE

LEONARDO GULARTE COELHO[1];LARISSA ASTROGILDO DE FREITAS[2]

[1]*Universidade Federal de Pelotas – lgcoelho@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*

1. INTRODUÇÃO

Atualmente existem diversas campanhas de conscientização na área da saúde. Tendo em vista este contexto, foram criados calendários nacionais da saúde com o objetivo de reunir e organizar tais eventos. Cada país possui um calendário próprio, podendo compartilhar datas semelhantes a outros e, seus meses são definidos por uma ou mais cores, representando suas principais datas comemorativas, a fim de facilitar tanto o estudo nas áreas, quanto campanhas de marketing dirigidas a eles. O calendário brasileiro de saúde pode ser encontrado no site do Ministério da Saúde (OMS, 2020).

As campanhas de marketing podem ser realizadas via redes sociais, meio extremamente eficiente e rápido de divulgação. Porém, a coleta e o tratamento das postagens pode ser uma tarefa difícil e demorada quando realizada manualmente. Com essa ideia, foi proposta a criação de um corpora destinado às campanhas do calendário brasileiro de saúde.

Com base no estudo de SARDINHA (2004), foi usada a ideia de Corpus Linguísticos para a representação desse conjunto, que se trata de uma coletânea de textos escritos coletados criteriosamente para serem uma amostra de uma língua ou variedade linguística (SARDINHA, 2004). Este trabalho, propõe a criação de um corpus para cada mês e no final um corpora (plural de corpus) sobre o calendário brasileiro de saúde.

Inicialmente as postagens foram extraídas da rede social Twitter, por conta da facilidade de busca por palavras-chaves através do uso das *Hashtags*. Após, os dados serem extraídos, eles passaram por uma análise linguística e estatística (por exemplo: frequência de palavras e expressões), com o intuito de encontrar padrões. Além disso, os *tweets* foram coletados no período de 10 anos (de 2010 a 2020), possibilitando uma análise temporal.

2. METODOLOGIA

O projeto foi totalmente desenvolvido em Python (VAN ROSSUM and DRAKE, 2011), utilizando as bibliotecas Pandas (PANDAS, 2020), NLTK (NATURAL LANGUAGE TOOLKIT, 2020), RegEx (PYTHON, 2020) e GetOldTweets versão 3 (MOTTL, 2020). Ele baseia-se em três scripts que, sequencialmente: extraem os *tweets* em grupos; realiza a união desses grupos em um *dataframe*; e, finalmente, analisa este *dataframe*.

Para a extração de dados, foi pensada na possibilidade de realizar o processo de escavação de dados (ou *Web Scraping* em inglês), que é a prática de coletar dados por qualquer outro meio além do programa interagindo com uma API (MITCHELL, 2015).

Pensando nisso foram estudadas duas bibliotecas do Python (VAN ROSSUM and DRAKE, 2011): a Tweepy (TWEEPY, 2020) e a GetOldTweets (MOTTL, 2020). Ambas possuem a mesma finalidade de extrair *tweets*, no entanto, com

diferenças que poderiam vir a afetar nos resultados. O Tweepy permite a captura de postagens no momento que são publicadas, garantindo que mesmo que o usuário apague a informação, ela ainda estará no corpus. Porém, não consegue obter *tweets* antigos. O GetOldTweets (MOTTL, 2020) faz a captura através de um *crawler* dentro do Twitter, podendo extrair postagens de um determinado tempo. Porém sem poder obter dados deletados.

Cada API tem suas vantagens e desvantagens, foi escolhida a GetOldTweets (MOTTL, 2020) por mostrar maior versatilidade e dinamismo na captura dos dados. A extração acontece ao final de cada mês, para manter um cronograma com dados antigo e atuais.

Para cada mês, é gerado um corpus, contendo as mesmas especificações, são elas: nome do usuário; link para a postagem original; data de publicação; *hashtags* usadas; número de *retweets* e favoritos obtidos; idioma; menções; localização; marcação com o mês equivalente no calendário (por exemplo: Maio Amarelo); e o texto da postagem.

É necessário tratar alguns dados dos conjuntos, como eliminar a substring "*RT @nome_do_usuario:*" do campo de texto das postagens, e *tweets* duplicados. Através do uso de expressões regulares, foram analisados cada texto, e, para um fim mais genérico, casos especiais que ocorriam na string foram ignorados.

Com os dados tratados, foi construído um corpora, unindo o corpus dos meses de maio, junho, julho e agosto, até o momento.

O corpora desenvolvido nesse trabalho encontra-se disponível nos formatos normalizado e não normalizado. A normalização de textos é o processo de transformações em uma string com o objetivo de remover distinções que são irrelevantes para determinadas aplicações (SPROAT et al., 2001).

3. RESULTADOS E DISCUSSÃO

Após extrair os dados eles foram analisados, em busca de padrões de comportamento linguístico e estatístico. Os dados finais são incertos, podendo ser alterados com o passar do tempo e influência de outras extrações.

Até o momento, foi construído um corpora de tamanho Médio-Grandes (de 1 milhão a 10 milhões de palavras), segundo a definição de SARDINHA (2004), com 129928 *tweets* extraídos. Na Tabela 1 é possível ver as *hashtags* mais utilizadas pelos usuários nos meses coletados, sendo a mais usada, até o presente momento, a "#maioamarelo" com 7821 *tweets*.

Tabela 1. Frequência de *Hashtags* dos *Tweets* Obtidos.

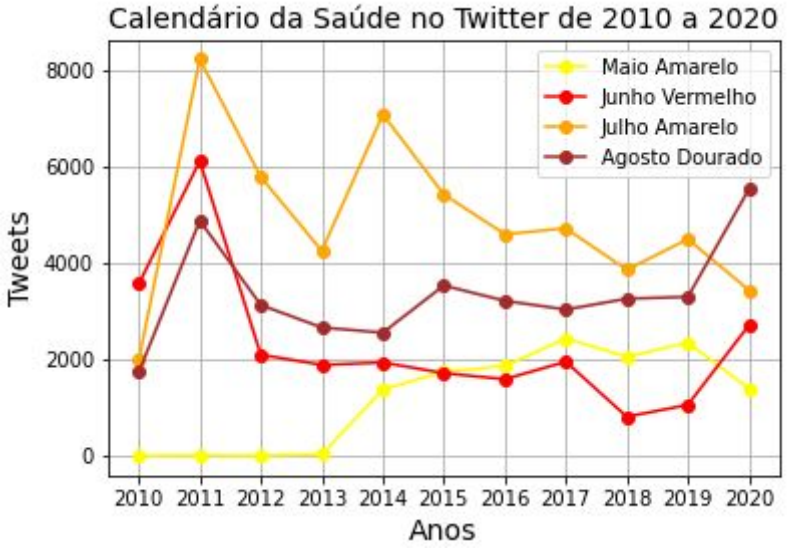
Maio Amarelo		Junho Vermelho		Julho Amarelo		Agosto Dourado	
#	Qtd	#	Qtd	#	Qtd	#	Qtd
#MaioAmarelo	7821	#doesange	5745	#hepatite	518	#amamenta	1557
#maioamarelo	2674	#DoeSangue	3851	#JulhoAmarelo	494	#AgostoDourado	813
#segurança	447	#JunhoVermelho	2091	#hepatites	452	#Amamenta	636

#MinhaEscolhaFazADiferença	289	#doevida	634	#Hepatite	298	#agostodourado	615
#transitoseguro	236	#junhovermelho	599	#WorldHepatitisDay	278	#aleitamentomaterno	559

Uma das métricas de análise foi a comparação do uso das *hashtags* durante o período de 10 anos escolhido para a coleta, e foi notado que cada mês teve seu pico por vários motivos, muitas vezes não ligados ao pós-surgimento da campanha, mas, provavelmente, ajudaram a impulsionar a criação dela. Esse é o caso, por exemplo, da campanha Junho Vermelho, onde o pico de postagens que usaram alguma das *hashtags* do mês aconteceu em 2011, enquanto sua criação foi no ano de 2015.

Já em contraparte, a campanha Julho Amarelo foi criada pela OMS instituída em 2010 e em 2011 houve a maior utilização das *hashtags* do evento. No Gráfico 1 é apresentada a evolução das postagens direcionadas às campanhas de 2010 a 2020, onde é possível observar que, de forma geral, houve mais postagens no ano de 2011.

Gráfico 1. Relação do número de *tweets* com o ano de postagem.



4. CONCLUSÕES

Como não haviam corpora especializados nessa área específica, deu-se início a construção de um corpora destinado às campanhas do calendário brasileiro de saúde. Esse corpora faz parte de um projeto maior que está sendo desenvolvido em paralelo, na área de estudo de fake news na área da saúde.

Futuramente, a intenção é expandir o corpora do calendário brasileiro de saúde nos demais meses do ano (setembro, outubro, novembro, dezembro, janeiro, fevereiro, março e abril) e com postagens de outras redes sociais, como Facebook e Instagram. Além disso, deseja-se disponibilizá-lo para uso da comunidade.

5. REFERÊNCIAS BIBLIOGRÁFICAS

SARDINHA, T.B. **Linguística de Corpus**. Barueri: Editora Manole Ltda, 2004.

VAN ROSSUM, G.; DRAKE, F.L. **The Python Language Reference Manual**. Network Theory Limited, 2011.

MITCHELL, M. **Web Scraping With Python**. Sebastopol: O'Reilly Media, 2015.

SPROAT, R.; BLACK, A.W.; Chen, S.; KUMAR, S.; OSTENDORF, M.; RICHARDS, C. Normalization of Non-Standard Words. **Computer Speech & Language**. v.15, n.3, p.287-333, 2001.

OMS. **Calendário da Saúde**. Ministério da Saúde. 16 ago. 2019. Acessado em 20 set. 2020. Online. Disponível em: http://www.saude.gov.br/calendario-da-saude

TWITTER. **Tudo sobre o Twitter**. Acessado em 20 set. 2020. Online. Disponível em: https://about.twitter.com/pt/company.html

FACEBOOK. **Facebook - Sobre**. Acessado em 20 set. 2020. Online. Disponível em: https://www.facebook.com/facebook/about/

INSTAGRAM. **Sobre o Instagram**. Acessado em 20 set. 2020. Online. Disponível em: https://www.twitter.com

PYTHON. **Expressões Regulares HOWTO**. Acessado em 20 set. 2020. Online. Disponível em: https://docs.python.org/pt-br/3.8/howto/regex.html

PANDAS. **Pandas Documentation**. Acessado em 20 set. 2020. Online. Disponível em: https://pandas.pydata.org/docs/

NATURAL LANGUAGE TOOLKIT. **NLTK 3.5 Documentation**. Acessado em 20 set. 2020. Online. Disponível em: https://www.nltk.org/

MOTTL, Dmitry. **GetOldTweets3**. PYPI. Acessado em 20 set. 2020. Online. Disponível em: https://pypi.org/project/GetOldTweets3

TWEEPY. **Tweepy Documentation**. Acessado em 20 set. 2020. Online. Disponível em: http://docs.tweepy.org/en/latest/

MELLO, Heloisa C. **Junho Vermelho: Conheça a Campanha que Movimenta a Doação de Sangue no Brasil**. Estadão, São Paulo, 16 jul. 2020.Acessado em 20 set. 2020. Online. Disponível em: https://blog.medicalway.com.br/junho-vermelho-doacao-de-sangue

MALAR, João Pedro. **Julho Amarelo 2020: campanha faz alerta sobre hepatites virais**. Medicalway, 11 ago. 2019.Acessado em 20 set. 2020. Online. Disponível em: https://emais.estadao.com.br/noticias/bem-estar,julho-amarelo-2020-campanha-faz-alerta-sobre-hepatites-virais,70003365047.

EISENSTEIN, J. **Natural Language Processing**. 13 nov. 2018. Acessado em 20 set. 2020. Online. Disponível em:https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf