IV SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R
R & PYTHON E AS TENDÊNCIAS DE COLABORAÇÃO
NITERÓI, 21 A 23 DE MAIO DE 2019

MINERAÇÃO DE TEXTOS: UMA ANÁLISE DE SENTIMENTO DOS TWEETS NA COPA DO MUNDO FIFA DE 2018

Suzana de Lima Santos da Silva[1], Anderson Luiz Ara Souza[2], Paulo Henrique Ferreira da Silva[3]

Introdução

A Copa do Mundo FIFA, ocorrida de quatro em quatro anos, é um dos maiores eventos esportivos internacionais, nela 32 seleções disputam o título de campeã mundial de futebol. Mesmo sendo uma competição, seus objetivos permeiam a celebração do esporte e do futebol, bem como a união de todas as nações. Foi idealizada em 1928 por Jules Rimet, o então presidente daquela época (BIBAS; TEIXEIRA, 1982).

No ano de 2018 ocorreu a 21ª edição na Rússia contando com 11 cidades sede e disputado em 12 estádios onde ao final da competição a seleção que obteve o título de melhor seleção do mundo foi a francesa.

Ao mesmo tempo, as mídias sociais fazem parte da rotina das pessoas. A comunicação que antes demorava para acontecer está cada vez mais rápida e instantânea. Na Copa do Mundo FIFA 2018, as redes sociais também foram destaque, famosas *tags* tomavam proporções mundiais e diariamente havia uma notícia.

Neste contexto, o Twitter é uma rede social que permite aos usuários enviar e receber mensagens de texto, fotos, *gif* e vídeos, onde aparecem em um perfil de cada usuário e seus seguidores recebem as atualizações feitas. Criado em 2006 por Jack Dorsey, Evan Williams, Biz Stone e Noah Glass, onde Jack Dorsey é o atual CEO. Toda publicação no twitter recebe o nome de *tweet*.

A fim de analisar os *tweets* da Copa do Mundo de 2018 foi feita uma coleta de dados direto da internet, conhecida com *web scraping*, para fazer as análises e a estruturação dos dados foi utilizado técnicas de *text mining* – mineração de texto.

Objetivos

Assim, este trabalho tem como objetivo realizar um estudo de análise de sentimentos através das técnicas de mineração de textos para os *tweets* em relação à Copa do mundo de 2018, em especial para o último dia, 15 de julho de 2018. O trabalho é inteiramente feito no *software R* (R Core Team, 2018), versão 3.5.1.

Material e Método

Com o crescimento das mídias digitais, volumosos conjuntos de dados não estruturados ou semiestruturado surgiram e com isso a urgência de se encontrar técnicas para analisar esses dados. Sabe-se que 80% do conteúdo online está em formato digital (CHEN, 2001) daí vem a grande importância de técnicas específicas para realizar análises textuais.

Mineração de texto diz respeito ao processo de obter informações de dados em texto cujo principal objetivo é identificar padrões e tendências de dados não estruturados ou semiestruturados, envolvendo assim várias técnicas criando um campo multidisciplinar.
Segundo Aranha e Passos (ARANHA; PASSOS, 2006) os desafios que esse método traz são: Grande volume de dado; Super parametrização; Estruturas dinâmicas; Dados ruidosos; Ambiguidade e Alta dimensionalidade.

[1] Universidade Federal da Bahia (UFBA), suzilima81@gmail.com

[2] Universidade Federal da Bahia (UFBA), anderson.ara@ufba.br

[3] Universidade Federal da Bahia (UFBA), paulohenri@ufba.br

A mineração de texto, segundo Dixon (1997), é dividida em processos: Recuperação da Informação/Indexação; Extração da Informação/Coleta; Mineração da Informação e Interpretação/Análise.

As informações textuais coletadas podem ser caracterizadas em dois tipos principais: fatos e opinião (LIU, 2010). Fato é aquilo que aconteceu já opinião é a interpretação dos fatos, ou seja, a opinião é subjetiva.

A análise de sentimento, também conhecida como mineração de opinião, surgiu a partir da necessidade de interpretar o subjetivo de uma maneira automática, sendo assim criar conhecimento estruturado.

Foi utilizada a base de dados do último dia da Copa do Mundo de 2018 com 10.000 observações no âmbito mundial, para fazer a mineração de texto foram utilizados os pacotes, *tm* (INGO FEINERER AND KURT HORNIK, 2018)*, tidytext* (SILGE J, ROBINSON D, 2016)*, dplyr* (HADLEY WICKHAM, ROMAIN FRANÇOIS, LIONEL HENRY AND KIRILL MÜLLER, 2019) e *tidyr (*HADLEY WICKHAM AND LIONEL HENRY, 2018) para análise de sentimentos foi utilizado o pacote *syuzhet* (JOCKERS ML, 2015) e para a visualização de dados os pacotes *ggplot2* (H. WICKHAM, 2016) e *worldcloud2* (DAWEI LANG AND GUAN-TIN CHIEN, 2018).

Resultados e Discussão

A priori os *tweets* seriam extraídos por meio do *software R,* há muitos pacotes que oferecem suporte para tal modalidade. Para este trabalho foi gerada uma API (*Application Programming Interface*, Application Programming Interface), que serve para facilitar o acesso a informações na web, mas posteriormente ocorreu um impasse pois o Twitter modificou suas leis de acesso e a API precisava ser paga e o projeto não tinha recurso, foi então utilizado o Software Sprinklr para a extração dos *tweets* de somente dois dias da Copa do Mundo de 2018. No demais foi utilizado o Software R para a estruturação e análise dos dados.

O pré-procesamento da base foi dado em algumas etapas, primeiro foi verificado se havia duplicidade na base de dados, atestada a unicidade foi realizado um filtro para selecionar apenas os *tweets* que estavam na língua inglesa, restando 7.316 tweets dos 10.000 coletados.

Nos *tweets* coletados foi necessário fazer a conversão para o padrão internacional, após a conversão foi extraída a parte textual com o comando *Corpus*, para a construção da corpora, a coleção de texto.

As palavras mais frequentes na base de dados foram *worldcup, france e 2018,* sendo que a primeira foi citada mais de 12000 vezes nos *tweets* coletados. A fim de facilitar a visualização foi feito o gráfico com auxílio do pacote *ggplot2*.

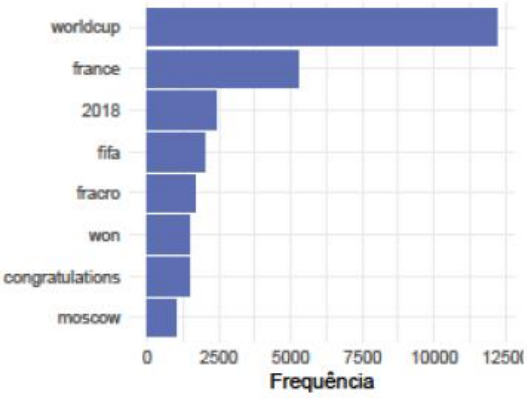

Figura 1 – Frequência de palavras nos *tweets*

Figura 2 – Nuvem de palavras nos *tweets*

As palavras da base de dados, indicado na Figura 2, aparentemente tem uma tendência de torcida, palavras como won e $france$ sendo a de maior fonte mostra que vários $tweets$ corroboram para a França como vencedora da Copa do Mundo de 2018. A nuvem de palavras foi feita com o auxílio do pacote *wordcloud2.*

Figura 3 – *Tweets* pelo território mundial

Há um acúmulo de tweets no hemisfério Norte, mais precisamente em todo continente Europeu e na América do Norte, como já era esperado, países como a Índia aparenta ter grande frequência, a Rússia é o país que é dividido em dois continentes como já era esperada a parte da Europa possui mais notificações que a parte Asiática.

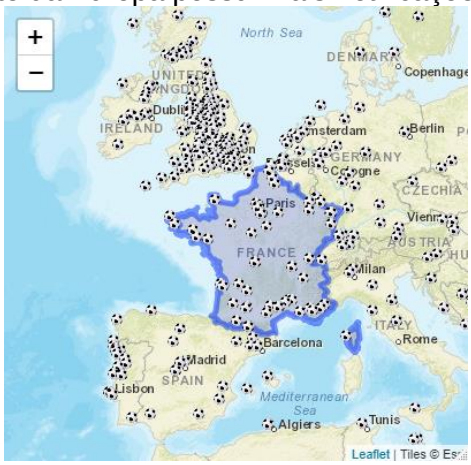

Figura 4 - Tweets pelo território francês

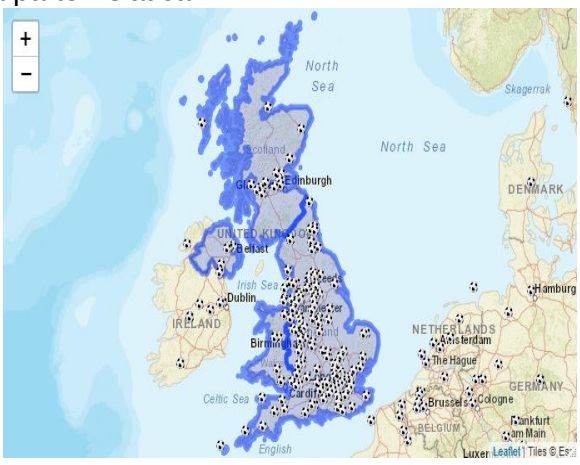

Figura 5 – Tweets pelo território do Reino Unido

A fim de verificar mais afundo o continente Europeu, foram elaborados mais dois gráficos com as notificações da França, por ser a campeã, e do Reino Unido, escolhido pela língua materna que foi usada de filtro na base de dados, pode-se perceber que o Reino Unido tem mais notificações, mas isso pode ser explicado pelo idioma que *tweet* foi extraído.

Para fazer a parte de análise de sentimento foi utilizado o pacote *syuzhet,* o dicionário de sentimento utilizado foi o *'nrc'* desenvolvido por Mohammad Saif M. e Turney, Peter D. (MOHAMMAD; TURNEY, 2010), nele está implementado 8 emoções e 2 sentimentos. Foram encontrados todas as oito emoções e os dois sentimentos.

Para elaboração da variável *score* foi feita como a combinação da variável *positivo* e *negativo* subtraindo uma pela outra, esta manipulação foi feita com auxílio do pacote *dplyr.* Com a variável score foi construída outra variável, *sentimento*, nela foi feita uma condição se valor observado dela para o tweet fosse $score$ < 0 seria atribuída à *string "*negativo", se fosse $score$ > 0 seria atribuída a *string "*positivo" e, por fim, se $score$ == 0 atribuída a *string "*neutro".

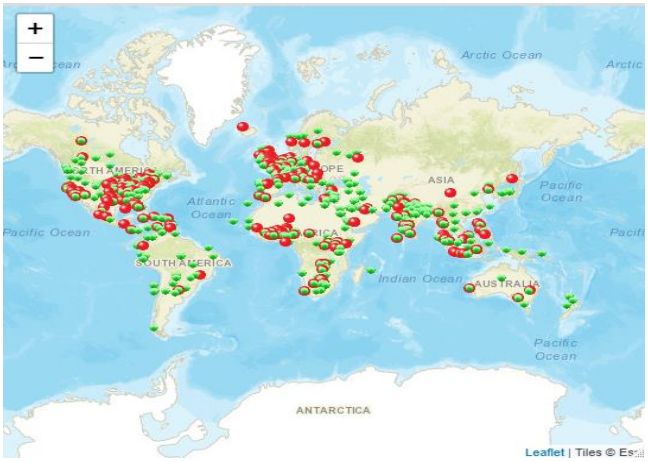

Figura 6- Sentimento pelo Mundo

Grande quantidade de *tweets* negativos na América do Norte e na Europa, na América do Sul em sua maioria foram de *tweets* positivos igualmente na Ásia.

Olhando a França percebe-se que não existe muitas observações já no Reino Unido a maior frequência é de *tweets* negativos.

Figura 7 – Sentimento pela França

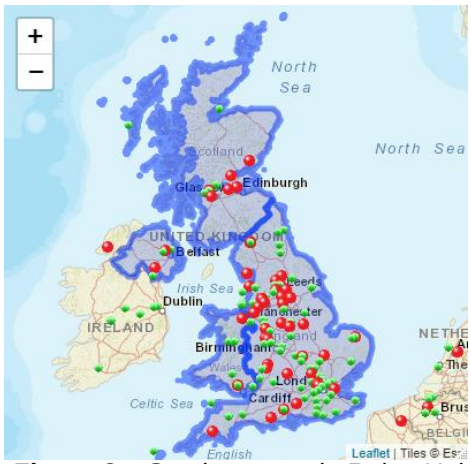

Figura 8 – Sentimento pelo Reino Unido

Conclusão

Neste trabalho não podemos usufruir de todos os dados da Copa do Mundo de 2018 pois as regras de extração de dados do twitter mudou recentemente. O software R mostrou-se bastante proveitoso para Mineração de texto e para as análises gráficas uma ressalva para a subjetividade dos métodos de Análise de Sentimentos.

Referências

Aranha, C.; passos, E. A tecnologia de mineração de textos. Revista Eletrônica de Sistemas de Informação, v. 5, n. 2, 2006.

Autarch. Twitter. 2008. Disponível em: <https://en.wikipedia.org/wiki/Twitter>.

BIBAS, S.; TEIXEIRA, J. de S. As Copas que ninguém viu: Histórias e bastidores. [S.l.]: Catavento, 1982.

Bhaskar Karambelkar and Barret Schloerke (2018). leaflet.extras: Extra Functionality for 'leaflet' Package. R package version 1.0.0. https://CRAN.R-project.org/package=leaflet.extras

Chen, H. Knowledge management systems: a text mining perspective. [S.l.]: Knowledge Computing Corporation, 2001.

Dawei Lang and Guan-tin Chien (2018). wordcloud2: Create Word Cloud by 'htmlwidget'. R package version 0.2.1. https://CRAN.R-project.org/package=wordcloud2

Dixon, M. An overview of document mining technology. Unpublished paper, 1997.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham and Lionel Henry (2018). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.2. https://CRAN.R-project.org/package=tidyr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.8. https://CRAN.R-project.org/package=dplyr

Ingo Feinerer and Kurt Hornik (2018). tm: Text Mining Package. R package version 0.7-5. https://CRAN.R-project.org/package=tm

Jockers ML (2015). _Syuzhet: Extract Sentiment and Plot Arcs from Text_. <URL: https://github.com/mjockers/syuzhet>.

Joe Cheng, Bhaskar Karambelkar and Yihui Xie (2018). leaf let: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.2. https://CRAN.R-project.org/package=leaflet

Liu, B. Sentiment analysis and subjectivity. Handbook of natural language processing, v. 2, p. 627–666, 2010.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.