Modelos estatísticos para classificação de mensagens em redes sociais: uma análise das menções à Unicamp no Twitter

Palavras-Chave: API, Modelos Estatísticos, Twitter

Autores:

Antonio Vítor Ribeiro [IMECC]

Prof. Dr. Guilherme Vieira Nunes Ludwig (orientador) [IMECC]

INTRODUÇÃO:

Redes sociais são um dos meios de comunicação modernos mais utilizados entre a população mundial. O uso de redes sociais se estende inclusive para interação e comunicação política (Zhang *et al.*, 2017, entre outros). De uma maneira geral, as opiniões compartilhadas em redes sociais tipicamente são orientadas por poucos agentes representativos que tendem a influenciar seu círculo de seguidores (Webster, 2014; Zhang *et al.*, 2017).

Há diversos aspectos de análise de dados em redes sociais que precisam ser considerados, como o processamento de texto (veja, e.g., Silva e Ribeiro, 2009) e critérios adequados de classificação (Schölkopf e Smola, 2002) e agrupamento (Bouveyron *et al.*, 2019). Por outro lado, a estrutura da rede social também pode ser estudada, identificando seus atores principais, agrupamentos e *cliques* (Wasserman e Faust, 1994).

O Twitter é uma rede social que utiliza um sistema de *microblog* e está em 8º lugar entre as redes sociais mais utilizadas no Brasil (Hostmidia, 2021). Sua característica de *posts* com no máximo 280 caracteres fazem com que a mensagem seja transmitida de forma mais direta, auxiliando na disseminação, sendo assim referência no quesito tendências e assuntos de pauta (Abrantes, 2019).

METODOLOGIA:

2.1 Coleta dos dados

Para a realização da coleta dos dados, foi utilizada duas linguagens de programação, Python e Rstudio. A linguagem Python foi utilizada para coletar os *tweets*, pois apresenta uma facilidade muito maior para a realização dessa tarefa, quando comparada a outras linguagens. Para a coleta dos dados, que é feita semanalmente, foi utilizado o pacote *TwitterSearch*, e passando o parâmetro para coletar todos os *tweets* com a palavra “Unicamp” no corpo do texto, em português. As informações extraídas eram data de postagem do *tweet*, o usuário que *tweetou* e o texto do *tweet*. Após a coleta, os dados são salvos em um arquivo .csv (comma-separated-values), para que a análise seja realizada de forma prática posteriormente, com a utilização do Rstudio.

A linguagem R foi utilizada para tratar os dados e realizar algumas análises inicias. Para isso, foram carregados alguns pacotes cruciais, como *rtweet*, *readr*, *worldcloud*, *tm*, *cluster* e *syuzhet*. Com isso, as análises realizadas foram: um gráfico de frequência dos *tweets* a cada 1 hora; uma *wordcloud* (nuvem de palavras); gráfico

de barras com os termos mais frequentes; uma análise de *clustering* hierárquico, dendrograma; um gráfico *K-means*; e um gráfico com análise de sentimento dos *tweets*. O livro do Bouveyron et al. foi tomado como referência cientifica para a análise dos modelos de *clustering*. A base de dados dispõe, atualmente, de 28 semanas de dados coletados, dando mais de 120 mil *tweets* coletados.

2.2 Dificuldades

Alguns dados da primeira semana foram perdidos, por conta da permissão para usar a API do twitter não permitir a coleta de dados de dias anteriores, desse modo, não foi possível coletar todos os dados. De mesmo modo, os dias para a coleta de dados foram alterados, para que a coleta fosse padronizada. No início, os dados eram coletados as terças, depois mudou para as sextas-feiras, no período noturno, e nas últimas semanas passaram para as segundas-feiras, também na parte da noite.

Devido as bibliotecas de análise de dados em R serem criadas para a língua inglesa, quando tentamos aplicar na língua portuguesa, não há a mesma eficácia, e nesse sentindo, algumas *stopwords* (palavras sem relevância para a análise de texto) não foram removidas dos textos, tendo de serem removidas de forma manual. Ainda de forma manual, foi retirado todos os tweets realizados por *bot's* que mencionam a Unicamp, pois só inflam os termos e não trazem nenhuma análise.

Os dados foram coletados através da biblioteca *TwitterSearch*, em Python, fazendo com que os dados de texto coletados ficassem limitados a 140 caracteres, que era o limite de caracteres antigo da rede social. Desse modo, todos os *tweets* estão truncados, sendo necessária a realização de uma análise de censura nos dados. Assim, uma limpeza manual nos termos foi necessária, pois as palavras estavam cortadas, poluindo o dicionário de termos que foi criado.

RESULTADOS E DISCUSSÃO:

A pesquisa teve início no começo de outubro de 2021 e conta com 28 semanas de coleta de dados, com um banco de dados de mais de 120 mil tweets, e análises preliminares dos mesmos. As análises foram realizadas em linguagem R, e o conteúdo foi acessado através do API do Twitter, usando o pacote twitteR (Gentry, 2015; R Core Team, 2020). Após os dados serem recolhidos, eles são processados e os atributos são extraídos. Entre os atributos, temos o usuário que fez o tweet, a data em que o tweet foi coletado, hyperlinks e contagem de interações e retweets. Potencialmente informações a respeito de seus seguidores e usuários que seguem; mas não coletamos neste estágio.

A Figura 1 e 2 mostram um gráfico de frequência, entre as semanas 25/09/2021 a 01/10/2021 e 25/01/2022 a 31/01/2022, respectivamente, exibindo o número de tweets que mencionam "unicamp", a cada hora. É possível notar que na figura 1 houveram muito menos tweets em relação a figura 2, sendo a data da figura um em periodo letivo e da figura 2 não. Também é evidente que houve um dia, em ambas as semanas, com um pico de atividade na rede social, indicando algum possível acontecimento que movimentou a comunidade. Tais eventos foram, de forma respectiva a revogação do titulo de Doutor Honoris Causa do Jarbas Passarinho pelo Conselho Universitário da Unicamp (Consu) e uma usúaria do twitter falando sobre a necessidade de doação de sangue para o Hemocentro da Unicamp.

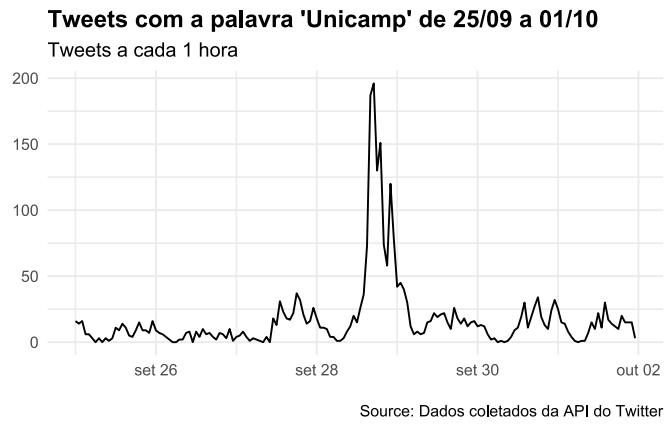

Figura 1: *Gráfico de frequência por hora*

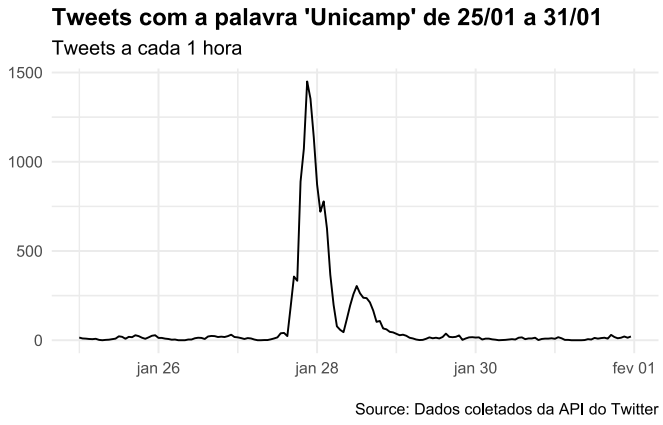

Figura 2: *Gráfico de frequência por hora*

A Figura 3 e 4 exibem o dendrograma para um modelo de cluster hierárquico, ainda das mesmas semanas que a análise anterior, construído com base em distância calculada através das frequências de ocorrências discordantes de palavras no dicionário. Com base nesta análise, podemos buscar grupos de palavras em comum, e termos relacionados entre os tweets. Outra possibilidade de aglomeração dos dados é utilizar a análise dos sentimentos (Jockers, 2015).

Na figura 3 pode-se identificar 4 grupos bem construídos, com todos em torno do mesmo assunto, porém sem grandes relações entre os termos, de modo geral. Isso acontece porque houve uma grande discussão sobre esse assunto por parte dos usuários, não limitando apenas aos retweets sobre a notícia. Já na figura 4 acontece algo estranho, devido a quase 93% dos tweets serem *retweets* sobre a necessidade da doação de sangue ao Hemocentro. As palavras “pra”, “hemocentro” e “sangue” aconteceram com uma frequência um pouco maior, pois como era esse o assunto, outras pessoas *tweetaram* por conta sobre isso, não entrando assim na grande massa de *retweets*. Assim, podemos perceber que houve uma comoção gingante por parte da comunidade, não apenas pelas *retweets*, mas também por levarem o apela adiante.

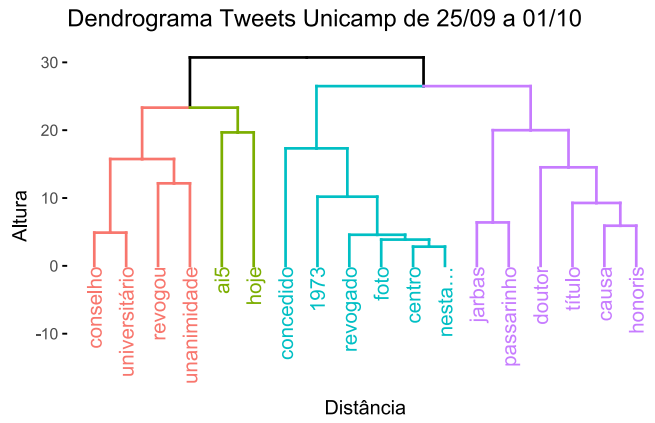

Figura 3: *Dendrograma dos termos mais frequentes*

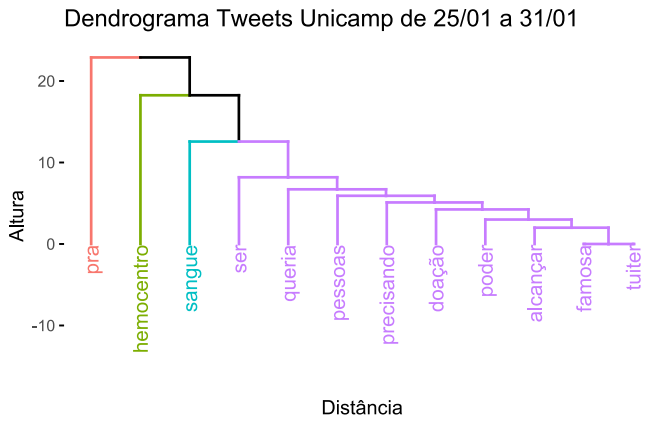

Figura 4: *Dendrograma dos termos mais frequentes*

A Figura 5 e 6 exibem a frequência de sentimentos predominantes em tweets das mesmas semanas já ditas. A *clusterização* é feita com base no cálculo da distância entre as probabilidades a posteriori dos sentimentos encontrados. Foi observado que a análise de sentimentos é particularmente sensível à lematização – por exemplo,

é incapaz de detectar ironia, e atribui sentimento negativos a palavras cotidianas da UNICAMP como “discussão”. Os resultados obtidos nesta abordagem ainda são incipientes e requerem mais trabalho.

Na figura 5, referente aos tweets sobre Jarbas Passarinho, há uma pontuação máxima de apenas 100 pontos de sentimentos, isso se deve ao fato de poucos tweets, ou poucas palavras, transmitirem algum sentimento do qual a máquina possa capitar. Assim, por mais que o assunto da semana fosse sobre a cassação de título, quase todos os tweets com sentimentos capitáveis foram de alunos falando sobre a universidade.

Algo similar acontece na figura 6, onde aproximadamente 93% dos tweets foram sobre a necessidade da doação de sangue, mas o sentimento predominante foi o de negatividade. Esse efeito foi causado por todos esses tweets também não transmitirem sentimentos, e os sentimentos que foram possíveis de armazenar foram, em maioria, de alunos, já em período de recesso acadêmico, comentando sobre cancelamento de matriculas, saudades de casa e sofrimento antecipado pela volta do semestre.

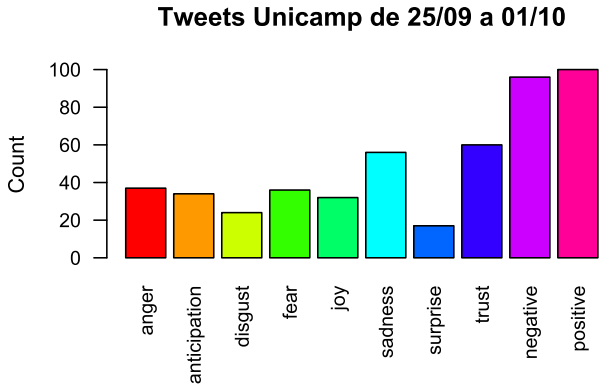

Figura 5: *Gráfico de pontuação de sentimento do tweets*

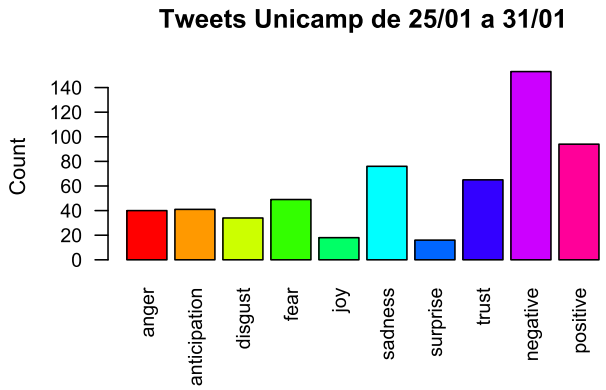

Figura 6: *Gráfico de pontuação de sentimento do tweets*

CONCLUSÕES:

Todos os dados relevantes para as análises foram extraídos e organizados em um banco de dados, juntamente com os termos mais frequentes em cada semana, concebendo assim um dicionário com tais termos, com o parâmetro de esparsidade em 98%. Com isso, foi possível a análise de agrupamento entre as semanas. Assim, podemos identificar grupos em comum que realizam os tweets baseado nessa análise, como por exemplo na 9º semana da análise, um grupo que fica evidente é o de pessoas que são de fora da unicamp, já que quase todas as menções foram sobre o vestibular ou a questões do mesmo, mas em um número muito mais elevado quando comparado a outras semanas, tendo 15x mais menções, em média.

BIBLIOGRAFIA

Bouveyron, C., Cleux, G., Brendan Murphy, T. e Raftery, A. E. (2019). Model-based clustering and classification for data science. Cambridge University Pr.

Feinerer, I. e Hornik, K. (2020). tm: Text Mining Package. R package version 0.7-8. URL: https://CRAN.R-project.org/package=tm.

Gentry, J. (2015). twitteR: R based Twitter client. R package version 1.1.9. URL: https://CRAN.R-project.org/package=twitteR.

HOSTMIDIA. As 10 Redes Sociais mais usadas no Brasil em 2021. URL: https://www.hostmidia.com.br/blog/redes-sociais-mais-usadas/. Acesso em 20, julho de 2022.

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

Kaufman, L. e Rousseeuw, P. J. (2005). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New Jersey.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2021). cluster: Cluster Analysis Basics and Extensions. R package. URL: https://CRAN.R-project.org/package=cluster

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

Schölkopf, B. e Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, Cambridge, Mass.

Silva, C. e Ribeiro, B. (2009). Inductive inference for large scale text classification: kernel approaches and techniques. Springer, New York.

Wasserman, S. e Faust, K. (1994). Social network analysis: methods and applications. Cambridge University Pr.

Webster, J. G. (2014). The marketplace of attention: how audiences take shape in a digital age. MIT Press, Cambridge, Mass.

Zhang, Y., Wells, C., Wang, S. e Rohe, K. (2017) Attention and amplification in the hybrid media system: The composition and activity of Donald Trump's Twitter following during the 2016 presidential election. New Media & Society, v. 20, no. 9, pp. 3161-3182.