



# Análise da subjetividade em língua portuguesa no Twitter relacionada a pandemia da COVID-19

Lucas Alexandre Malakin<sup>1</sup>, Solange Oliveira Rezende<sup>2</sup>  
ICMC-USP

## 1 Introdução

O Twitter é um serviço de micro *blogging* onde os usuários criam e compartilham mensagens de texto com suas visões e opiniões, que são chamados de *tweets*. Esses *tweets* são limitados a 280 caracteres e podem conter aprovação e desaprovação através de textos, emojis e outras mídias sobre diversos assuntos, como filmes, eventos e leis. Dessa maneira, essas mensagens possibilitam analisar o sentimento da população em diversos assuntos, sobretudo dos relacionados com a pandemia de COVID-19 em seu estágio inicial, entre novembro de 2019 e junho de 2020, este trabalho, será focado no estudo do sentimento da população brasileira através de *tweets* relacionados ao coronavírus no início da pandemia. Para tal, foram escolhidos como ferramentas o Python 3 com suas bibliotecas padrão[4][5] e LeIA (Léxico para Inferência Adaptada), que é uma biblioteca baseada no léxico e ferramenta para análise de sentimentos VADER (Valence Aware Dictionary and sEntiment Reasoner)[2] adaptado para textos em português, com suporte para emojis e foco na análise de sentimentos de textos expressos em mídias sociais[1].

## 2 Coleta dos dados

Para a coleta de dados foi utilizado a API de procura do Twitter com licença acadêmica, que é utilizada para buscas relacionadas a dados históricos limitados a dez milhões de *tweets* mensais[8]. Com isso, através de palavras chave, na qual as utilizadas para busca foram covid, covid-19, corona, corona vírus, coronavírus, coronga, vírus e sars-cov-2; com seleção para idioma em português, em um período mais amplo do citado anteriormente, entre setembro de 2019 e fevereiro de 2022; houve a coleta de informações como o conteúdo dos *tweets* (texto, emojis, quantidade de compartilhamentos, comentários, dentre outros) e o perfil dos usuários (localização, descrição do

---

<sup>1</sup>lucas.malakin@usp.br

<sup>2</sup>solange@icmc.usp.br

perfil, quantidade de seguidores, dentre outros). Dessa mensagens, foram coletadas as mais relevantes da plataforma, com a extração diária limitada a 10.000 mensagens. É comum que os dados adquiridos para análise não estejam em um formato adequado para a extração de conhecimento, com isso, fazem-se necessários os usos de métodos de tratamento, limpeza e redução do volume de dados antes de iniciar a etapa do enriquecimento de informação[6], contabilizando mais de 8.000.000 de mensagens.

### 3 Metodologia

Após a etapa de coleta e processamento dos dados, podemos extrair padrões para enriquecimento da informação, como por exemplo, a análise de sentimento, que está relacionada à extração da subjetividade e da polarização em um texto. Ela pode ser realizada a partir de algoritmos de aprendizado de máquinas supervisionados e não supervisionados, possibilitando a compreensão da polarização das sentenças[3]. Através de uma abordagem léxica que classifica as palavras do texto entre negativas, neutras e positivas, não necessitando de um conjunto de dados para treinamento de modelo[9], foi selecionado a LeIA, que é uma adaptação do VADER para português capaz de lidar com palavras, abreviações, gírias e emojis, amplamente utilizados em redes sociais. Através desse enriquecimento dos dados podemos analisar os resultados procurando entender a variação da subjetividade das palavras chave relacionadas ao novo coronavírus.

### 4 Resultados

Através da coleta de dados, pré-processamento e enriquecimento da informação, foi possível gerar a Figura 1. Faz-se notório a presença de dois patamares, um primeiro anterior a janeiro de 2020 e o segundo após janeiro de 2020. Isso causado pela disseminação da doença, em dezembro de 2019 vieram a público os primeiros casos confirmados da COVID-19[7].

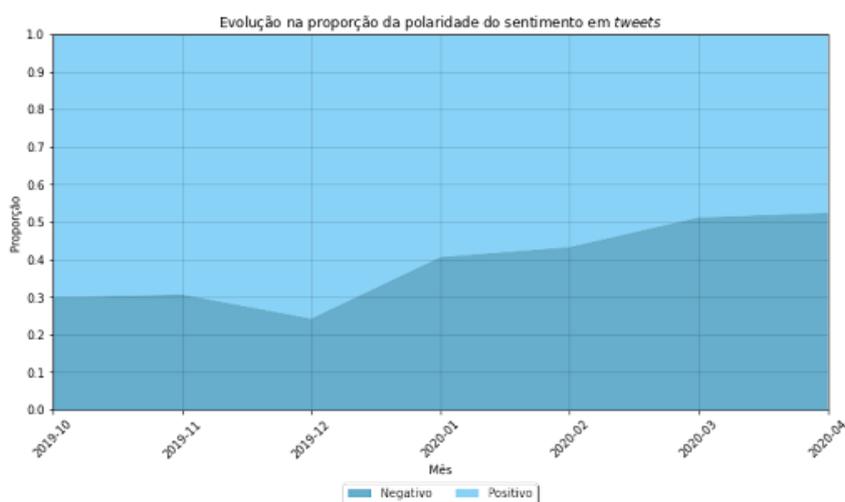


Figura 1: Proporção da polaridade do sentimento em *tweets* entre outubro de 2019 e abril de 2020.



## Referências

- [1] Rafael J. A. Almeida. *LeIA - Léxico para Inferência Adaptada*. <https://github.com/rafjaa/LeIA>. 2018.
- [2] CJ Hutto Eric Gilbert. “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. Em: *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) [http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf](http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf). 2014.
- [3] Hongfang Liu et al. “Towards a semantic lexicon for clinical natural language processing”. Em: *AMIA Annual Symposium Proceedings*. Vol. 2012. American Medical Informatics Association. 2012, p. 568.
- [4] Wes McKinney. “Pandas, python data analysis library”. Em: *URL [http://pandas. pydata. org](http://pandas.pydata.org)* (2015).
- [5] Adil Moujahid. “An introduction to text mining using twitter streaming api and python”. Em: *Adilmoujahid. com* (2014).
- [6] Solange Oliveira Rezende. *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003.
- [7] Alfonso J Rodriguez-Morales et al. “COVID-19 in Latin America: The implications of the first confirmed case in Brazil”. Em: *Travel medicine and infectious disease* 35 (2020), p. 101613.
- [8] *Twitter Developer*, 2022. URL: [https://developer. twitter. com/en](https://developer.twitter.com/en).
- [9] Lei Zhang et al. “Combining lexicon-based and learning-based methods for Twitter sentiment analysis”. Em: *HP Laboratories, Technical Report HPL-2011 89* (2011), pp. 1–8.