Associação Propagadora Esdeva
Centro Universitário Academia – UniAcademia
Curso de Engenharia de Software

# Análise de sentimentos em Tweets: Proposta de uma ferramenta de análise de opiniões

*Igor Costa de Oliveira Obeica*
*Centro de Ensino Superior de Juiz de Fora, Juiz de Fora, MG*
*Daves Marcio Silva Martins*
*Centro de Ensino Superior de Juiz de Fora, Juiz de Fora, MG*

*Linha de pesquisa: Engenharia de Software*

**Resumo.** Análise de sentimentos é uma área que vem ganhando bastante atenção nos últimos tempos. O objetivo dessa técnica é classificar sentenças de forma automática como positivas, negativas e neutras. Por meio dessa análise é possível avaliar a opinião da população sobre diversos assuntos, viabilizando mensurar em números a qualidade de estratégias e monitorar a reputação de campanhas e negócios. O objetivo desse trabalho é propor uma arquitetura para extrair, refinar e analisar os sentimentos de comentários no Twitter. Será realizado um estudo de caso relacionado ao tema vacina como forma de validação da ferramenta.

**Abstract.** Sentiment analysis is an area that has been gaining a lot of attention recently. The objective of this technique is to automatically classify sentences as positive, negative, and neutral. Through this analysis it is possible to evaluate the population's opinion about several subjects, making feasible to measure in numbers the quality of strategies and monitor the reputation of campaigns and businesses. This work aims to propose an architecture to extract, refine and analyze the sentiments of Twitter comments. A case study related to the vaccine theme will be carried out as a way of validating the tool.

**Palavras-chave:** Análise de sentimentos, COVID-19, Coronavírus, Vacina, Redes Sociais, Twitter, Google Cloud API, Java, JavaScript.

# 1. Introdução

O crescente uso da Internet e dos serviços prestados para seus usuários têm gerado informações de forma massiva. As pessoas contribuem ativamente no crescimento desse volume de dados utilizando aplicativos, redes sociais, ou até mesmo através de suas pesquisas na internet, colaborando para formar uma espécie de Inteligência Coletiva (O'REILLY, 2007). Este cenário fornece uma proliferação de dados não estruturados, trazendo novos desafios e oportunidades para pesquisadores de todas as áreas (EIRINAKI et al. 2012).

Stephens-Davidowitz (2017) afirma que quando as pessoas compartilham opiniões em redes sociais ou realizam pesquisas na internet, seja sobre fatos, citações, piadas, lugares ou pessoas, podem nos dizer muito mais sobre o que elas realmente pensam, desejam, temem, e fazem do que qualquer um poderia ter imaginado. O ato diário de digitar palavras em dispositivos tecnológicos conectados à internet deixa um pequeno rastro de informação, que quando multiplicado por milhões, eventualmente revelam realidades profundas. Esse conjunto extenso de dados é chamado de *Big Data*.

O que faz do *Big Data* tão importante e poderoso é justamente a quantidade de dados inseridos na *web* todos os dias, somando cerca de 2.5 quintilhões bytes de dados diários (EARTHWEB, 2022). Grande parte desses dados possuem informações que não poderiam ser adquiridas de outra forma. Na internet os usuários têm a sensação de estarem blindados, livres de julgamento ou repressão, por se comunicarem e expressarem através de perfis e avatares dizendo coisas que provavelmente não diriam se perguntadas de forma direta e pessoal (STEPHENS-DAVIDOWITZ, 2017).

Diante desse cenário e da importância de compreender as opiniões sobre diversos assuntos, a proposta desse trabalho se resume na criação de uma arquitetura capaz de coletar dados na rede social Twitter e aplicar ferramentas de Análise de Sentimentos. O tema escolhido para validar a arquitetura é a vacina, por ser um assunto que gera polarização de opiniões em tempos de uma crise sanitária mundial ocasionada pela Covid-19.

Na segunda seção abordou-se o referencial teórico, necessário para extrair os dados do Twitter e realizar a análise.

Na terceira seção foi exposta a proposta do trabalho, explicando a motivação da pesquisa e a metodologia aplicada para analisar e classificar os comentários coletados.

A quarta seção explica a arquitetura do sistema, explorando as características e funcionalidades de cada serviço criado e as tecnologias utilizadas durante o processo de desenvolvimento.

Na quinta seção as análises das informações são explicadas através de dashboards e tabelas.

Por fim, na sexta e útima seção, estão descritas as considerações finais do trabalho.

# 2. Referencial Teórico

## 2.1. Processamento de linguagem natural (PLN) e Análise de sentimentos

Processamento de linguagem natural é uma área de pesquisa da Ciência da Computação e da inteligência. O referido processamento geralmente envolve a tradução de linguagens naturais em dados (números), pelos quais os computadores podem usar para aprenderem sobre o mundo (LANE et al. 2019).

As informações extraídas podem ser utilizadas para resolver problemas e tarefas específicas. Bons exemplos são sistemas de e-mail, que buscam por palavras-chave específicas nas mensagens para classificá-los como Principal, Promoção, Social e *Spam*, como também em *sistemas de mensagens* que interpretam e respondem usuários reais.

Análise de sentimentos, também conhecida como Mineração de Opinião, é um campo de estudo do PLN, que visa analisar a opinião das pessoas, sentimentos, avaliações, atitudes e emoções sobre entidades ou atributos expressados nos textos. Em suma, o objetivo da análise de sentimentos é identificar opiniões positivas, negativas e neutras ou sentimentos expressos ou implícitos nos textos (LIU. 2015).

Em uma plataforma de vendas na internet, por exemplo, seria possível extrair todos os comentários e suas classificações (estrelas) de diversos produtos

com a ajuda de um *web scraper*[1]. Com os dados coletados seria possível treinar um algoritmo para classificar as palavras, n-gramas[2] e frases de acordo com suas ocorrências em comentários positivos e negativos. Desta forma o algoritmo se tornaria capaz de detectar o sentido e o sentimento de qualquer comentário nos produtos do *e-commerce*, auxiliando em tomadas de decisões na empresa.

Neste trabalho foi utilizada a *API* Natural Language do Google Cloud Services[3] e todas as definições abaixo foram retiradas de sua documentação. A citada *API* fornece métodos como: análise de sentimentos, análise de entidades, análise de sintaxe, análise de sentimentos de entidades e classificação de conteúdo.

O serviço de análise de sentimentos do Google Cloud recebe do usuário um documento ou texto e realiza o processamento com seus algoritmos já treinados pela própria plataforma. O retorno do *serviço* é um objeto no formato JSON com as seguintes informações: *score*, *magnitude e language.* A interpretação dos dados deve ser feita seguindo a tabela da imagem 1.

| Sentimento | Amostras de valores |
|---|---|
| Claramente positivo* | "score": 0.8, "magnitude": 3.0 |
| Claramente negativo* | "score": -0.6, "magnitude": 4.0 |
| Neutro | "score": 0.1, "magnitude": 0.0 |
| Misto | "score": 0.0, "magnitude": 4.0 |

Imagem 1 - Fonte: Documentação Google Cloud NLP

- O campo *score* se refere à inclinação emocional geral do texto, variando de -1.0 e 1.0.
- O campo *magnitude* é um valor maior que *zero* que representa a força geral da emoção. Quanto mais palavras existirem no texto, maior será o valor da magnitude. Textos curtos normalmente representam um baixo nível emocional na sentença analisada, ao contrário de textos mais longos.
- O campo *language* representa o idioma do texto ou documento.

1 Sistema responsável por procurar e coletar dados específicos de páginas na internet
2 Sequência contínua de palavras em uma amostra de texto
3 https://cloud.google.com/natural-language/docs/

As imagens abaixo mostram uma análise exemplar com as seguintes frases retiradas de um produto de uma plataforma de vendas: “Bom produto e serviço” e “Não desperdice seu dinheiro com isso”. A primeira frase, claramente positiva, obteve o resultado da imagem 2. A segunda frase com conteúdo negativo apresentou o resultado da imagem 3.


Image 2 - Fonte: Elaborado pelo autor


Image 3 - Fonte: Elaborado pelo autor

## 2.2. Twitter API

O Twitter é uma rede social que foi criada para que as pessoas compartilhassem o que estão pensando. É possível acessar a plataforma pela web ou pelos aplicativos de aparelhos *mobile*. Para um compartilhamento de informações mais amplo, o Twitter também fornece uma API para acesso programático para empresas, desenvolvedores e usuários.

A API do Twitter[4] compreende uma grande variedade de serviços que podem ser divididos em 5 grupos principais:

- Contas e usuários: permite que desenvolvedores gerenciem perfis e configurações, bloqueiem usuários, gerencie seguidores ou solicite informações sobre atividades não autorizadas na conta.

4 https://developer.twitter.com/en/docs

- *Tweets* e respostas: Deixam todos os *tweets* públicos disponíveis para os desenvolvedores, além de permitir que estes criem novos *tweets* pela API. Os *tweets* podem ser pesquisados através de *keywords* ou requisitando amostras de usuários específicos.
- Mensagens diretas: são endpoints que fornecem acesso às conversas através de mensagens diretas de usuários que fornecem permissão de forma explícita para uma determinada aplicação.
- ADS: um conjunto de endpoints voltados para facilitar a criação de gerenciamento de campanhas e propaganda no Twitter.
- Ferramentas de publicação e SDKs: conjunto de ferramentas para que desenvolvedores de software possam incluir tweets, botões de compartilhamento e outros conteúdos do Twitter, em seus *websites*.

Neste trabalho foi utilizado o serviço de tweets e respostas, onde é possível coletar os tweets em tempo real através de um termo de consulta. Cada *tweet* possui uma grande quantidade de dados como data de criação, nome do usuário, conteúdo do texto, número de respostas, número de repostagens e geolocalização (quando autorizada pelo usuário), como mostra a imagem 4. O único dado utilizado na análise foi o texto do *tweet*.


```
{
  "created_at": "Tue Feb 27 21:11:40 +0000 2018",
  "id": "000000000000000000",
  "id_str": "000000000000000000",
  "text": "RT @john_doe: Random comment on something https://t.co/Hash001",
  "truncated": false,
  "user": {
    "id": "000000000000000000",
    "id_str": "000000000000000000",
    "name": "doe",
    "screen_name": "john_doe",
    "location": "bbh iu jjh pcy kjd dks",
    "url": "https://curiouscat.me/baekhyun-l",
    "description": "hi hello I love exo",
    "followers_count": 1142,
    "friends_count": 125,
    "listed_count": 20,
    "favourites_count": 5712,
    "statuses_count": 4011,
    "created_at": "Fri Dec 04 03:44:59 +0000 2015",
    "utc_offset": -28800,
    "time_zone": "Pacific Time (US & Canada)",
    "lang": "en"
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "quote_count": 0,
  "reply_count": 0,
  "retweet_count": 0,
  "favorite_count": 0
}
```


Image 4 - Imagem resumida do retorno da API do twitter. Fonte: Documentação Twitter API

## 2.3. Microsserviços e Programação orientada a eventos

A arquitetura de Microsserviços é uma abordagem para desenvolver uma única aplicação como um conjuntos de pequenos serviços, cada um rodando em seu próprio processo e se comunicando através de mecanismos simples e leves, na maioria das vezes como *APIs* que utilizam o protocolo HTTP (protocolo que especifica como será a comunicação entre um navegador e um servidor web) (FOWLER, 2014).

Segundo Fowler (2014), as principais características desse tipo de arquitetura são:

- Componentização de serviços: são unidades de software que podem ser construídas, evoluídas e substituídas de forma independente.
- Organização por responsabilidades de negócio: os serviços normalmente são separados de acordo com áreas de negócios específicas. Dessa forma, fica claro qual equipe deve atuar nas demandas, assim como quais sistemas devem ser ajustados.
- Governança descentralizada: Um dos problemas da governança centralizada é que em grande parte das vezes existe uma tendência em padronizar as ferramentas e tecnologias. Dividir os sistemas em serviços permite que sejam escolhidas tecnologias diferentes para cada sistema desenvolvido, como linguagem, frameworks e banco de dados que sejam mais pertinentes em cada solução.
- Automação de infraestrutura: Cada serviço desenvolvido permite a criação de uma infraestrutura automatizada diferente. Em uma aplicação monolítica (sistema único, não dividido, não modularizado e que roda em um único processo), uma pequena alteração no sistema implica em realizar o build, test e deployment de todo o sistema. Na arquitetura de microsserviços, cada módulo pode ter seu próprio pipeline de automação[5], simplificando o trabalho das equipes nos ajustes das aplicações.

Nadareishvili et al. (2016) cita algumas das vantagens desse tipo de arquitetura:

5 Conjunto de etapas a serem executadas para disponibilizar uma nova versão de um software

- Reduz as dependências entre os times, resultando em um desenvolvimento mais rápido.
- Permite que várias iniciativas sejam elaboradas de forma paralela.
- Permite a utilização de múltiplas tecnologias/linguagens/*frameworks*.
- Permite degradação graciosa dos serviços, o que acontece quando o sistema mantém funcionalidade quando partes dos serviços passam por algum tipo de problema.

Microsserviços e arquiteturas baseadas em serviços existem há muitos anos, em muitas formas e nomes diferentes. Nestes modelos, os serviços normalmente se comunicam uns com os outros de forma síncrona e direta. Existem também arquiteturas onde seus serviços se comunicam através de mensagens, chamados Microsserviços orientados a eventos (O'REILLY, 2007).

Os microsserviços orientados à eventos seguem as mesmas diretrizes citadas anteriormente, mantendo a simplicidade e coesão dos serviços (divididos por responsabilidades lógicas ou de negócio), mas esses se comunicam por mensagens de forma assíncrona, produzindo e consumindo eventos em listas ou filas (imagem 5). Esses eventos podem ser destruídos ao serem consumidos por uma aplicação, ou podem ser mantidos para que outras aplicações tenham acesso a ele. Essa combinação de eventos e microsserviços permite uma ótima interconexão e flexibilidade entre as aplicações, além de reduzir ainda mais as dependências entre os serviços (BELLEMARE, 2020).

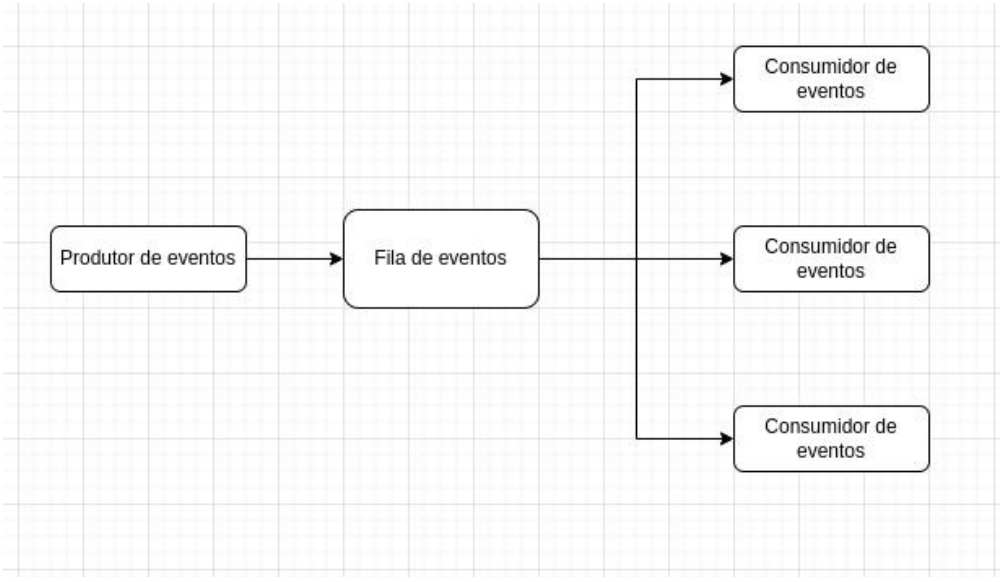


Image 5 - Fonte: Elaborado pelo autor

Para o trabalho foram criados 2 serviços: o primeiro sendo responsável por coletar os dados na API do twitter e o segundo por realizar a análise de sentimentos dos comentários encontrados. Os serviços se comunicam através de uma fila implementada no Redis (imagem 6).

# 3. Proposta de trabalho

Atualmente as redes sociais são o principal meio pelo qual as pessoas expressam suas opiniões. Nas redes estão concentrados volumes massivos de dados: mensagens, fotos, dados de geolocalização, interesses e até mesmo ofertas de produtos e empregos. Se trabalhados de forma correta, esses dados podem gerar ativos de valor (STEPHENS-DAVIDOWITZ, 2017).

O mundo viveu tempos difíceis ao enfrentar uma crise sanitária, milhões de pessoas adoeceram ou morreram por conta do COVID-19. Cientistas do mundo inteiro se dedicaram trabalhando por uma cura e em tempo recorde criaram modelos eficazes de vacinas contra o vírus SARS-CoV-2 (VACCINES EUROPE, 2021).

A vacina é uma importante forma de imunização ativa que garante que o agente causador da doença infecte o corpo do indivíduo e que ele já esteja preparado para responder de forma rápida antes do surgimento de sintomas. A gripe, tétano e a poliomielite são exemplos de doenças que podem ser prevenidas através da vacinação (SANTOS, 2022).

No entanto, mesmo com os resultados positivos apresentados pelo imunizante, a população sempre questionou sua obrigatoriedade e eficácia. Assim como ocorreu na Revolta da vacina, quando pessoas protestaram contra a vacinação obrigatória no Brasil, o movimento antivacina vem crescendo fomentado pela pandemia do COVID-19.

Portanto, a proposta deste trabalho consiste em utilizar uma metodologia e criar uma arquitetura para armazenar dados de tweets, aplicando técnicas de análise de sentimentos para obter conhecimento da opinião das pessoas (positivas, negativas ou neutras) sobre o tema “vacina”, a fim de validar a arquitetura desenvolvida.

Duas análises serão feitas como parte da validação. A primeira com o objetivo de classificar os sentimentos dos comentários que continham a palavra

vacina, enquanto a segunda análise consiste em  agrupar as posições sobre cada um dos fabricantes dos imunizantes para o COVID-19.

# 4. Arquitetura do sistema

## 4.1. Visão geral

A arquitetura utilizada foi a de microsserviços orientados a eventos. O projeto é constituído de dois serviços responsáveis pela análise de sentimentos dos dados coletados e dois bancos de dados para visualização das informações e criação de dashboards.

O primeiro sistema é um API que foi desenvolvido com o framework Node JS[6] e segue a arquitetura MVC. O segundo é um serviço desenvolvido na linguagem Java, utilizando o ecossistema Spring Boot e baseado nos padrões da Arquitetura Limpa.

Todos os componentes do projeto estão hospedados em ambientes em nuvem (Heroku[7], MongoDB Atlas[8] e Redis Cloud[9]), utilizando o CircleCI[10] como pipeline CI/CD (integração contínua e desenvolvimento contínuo). Ademais, os serviços de hospedagem utilizados possuem versão gratuita.

## 4.2. Fluxo do sistema

O fluxo começa com um sistema que coleta os comentários na API do Twitter, contendo a palavra “vacina”. Outra aplicação recebe os *tweets* e os envia para a API Natural Language do Google Cloud para que possam ser processados. Os dados analisados então são salvos em um banco de dados para que seja feita a análise através da criação de dashboards e tabelas, como demonstrado na imagem 6.

6 https://nodejs.org/en/docs/
7 https://devcenter.heroku.com/
8 https://www.mongodb.com/atlas/database
9 https://redis.com/redis-enterprise-cloud/overview/
10 https://circleci.com/docs

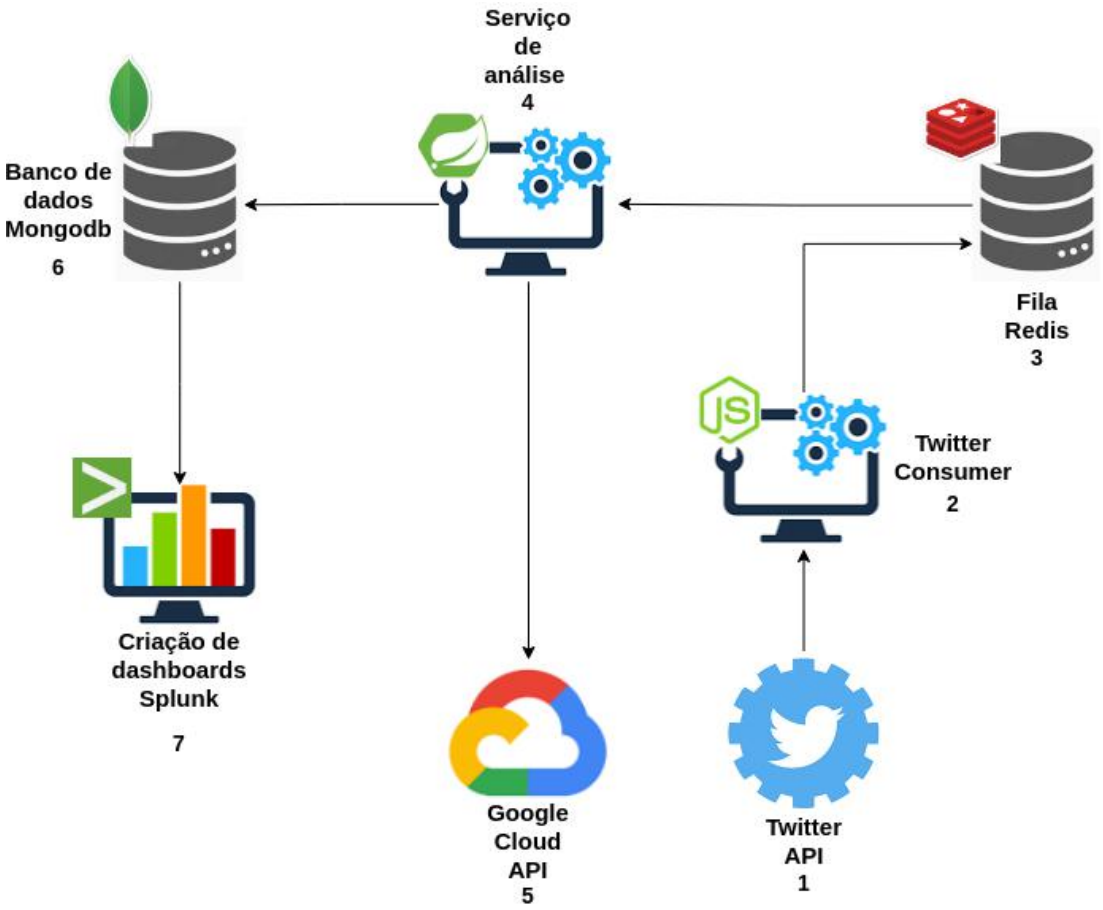


Image 6 - Fonte: Elaborado pelo autor

A aplicação Twitter Consumer[11] e o Serviço de Análise[12] se encontram disponíveis no Github para acesso. Todas as configurações podem ser feitas via variáveis de ambiente, inclusive o próprio termo de busca, o que possibilita pesquisas em diversos temas. Qualquer ferramenta pode ser utilizada para a visualização e criação dos dashboards, basta se consumir os dados armazenados na base final (MongoDB).

Todas as informações necessárias para a execução dos serviços está documentada nos arquivos README.md dos repositórios, como os pré-requisitos do sistema, os endpoints disponíveis e todas as credenciais necessárias. Os dois bancos de dados (MongoDB e Redis) e credenciais (Google Cloud e Twitter API) devem ser criados pelo próprio usuário.

11 https://github.com/IgorCooli/twitter-consumer
12 https://github.com/IgorCooli/twitter-analysis-service

### 4.2.1. Twitter consumer

O serviço inicial do fluxo de sistemas, desenvolvido em Node JS, é uma aplicação consumidora da Twitter API, que possui um *endpoint* para sua ativação. Os termos de busca, tempos de processamento e configurações de conexões externas foram parametrizados através de variáveis de ambiente. Desse modo, ao receber uma requisição, a aplicação começa a busca no Twitter API pelo termo desejado.

Ao término da busca no tempo parametrizado, os dados passam por um tratamento inicial via REGEX[13], os termos indesejados (caracteres especiais e emojis) são removidos dos dados e objetos são criados selecionando somente as informações desejadas. Os dados então são enviados para uma fila implementada no Redis.

### 4.2.2. Serviço de análise

O serviço de análise é o responsável, principalmente, pelo envio dos dados à API Natural Language do Google Cloud. Essa API desenvolvida em Java Spring Boot, é conectada à fila do Redis[14], que contém os dados recolhidos pelo primeiro sistema.

A app recolhe os dados do Redis em ordem de chegada. Os dados são tratados novamente, agora através da remoção de *stopwords* e numerais indesejados.

Com os objetos criados e os textos normalizados, o serviço envia os comentários para serem processados no Google Cloud, retornando os valores de *score* e *magnitude* mencionados anteriormente. Os dados são salvos em uma base de dados MongoDB para serem consumidos e utilizados para análise através de *dashboards* e tabelas.

### 4.2.3. Criação de dashboards

Splunk[15] é um *software* para pesquisar, indexar, analisar e visualizar dados de fontes diversas, como *websites*, aplicações, sensores e dispositivos. É

13 Sequência de caracteres que especifica um padrão a ser pesquisado em determinado texto
14 https://redis.io/docs/
15 https://docs.splunk.com/Documentation

amplamente utilizado em grandes empresas para monitoramento de logs[16] de aplicações, pois possibilita a criação de alarmes, relatórios e visualizações.

Essa ferramenta de *Big Data* tem a capacidade de identificar padrões de dados, provendo métricas e diagnósticos em tempo real. A ferramenta também aceita diversos tipos de dados como: json, *plain text, log format e csv*.

Splunk também desenvolveu uma linguagem de consulta chamada de *Search Processing Language* (SPL). A SPL engloba todos os comandos de pesquisa e suas funções, argumentos e cláusulas (*top*, percentil, média, mediana, etc), o que possibilita a criação de tabelas e dashboards através de consultas sofisticadas.

A versão gratuita do Splunk foi utilizada nesse trabalho para a criação dos dashboards, tabelas e grafos que auxiliaram nas análises e informações obtidas no projeto. Ademais, alguns *plugins* foram utilizados por fornecerem ferramentas específicas para o processamento de linguagem natural (NLP Text Analytics). Todos os *plugins* foram baixados gratuitamente na plataforma splunkbase[17].

É possível conectar o Splunk à uma base de dados permitindo a criação de consultas através da SPL, como demonstrado na imagem 7. Selecionando a aba de "Visualização", a ferramenta fornece uma grande variedade de gráficos possíveis para exibição dos dados (imagem 8). Esse processo auxilia na criação de uma visão sistemática das informações, facilitando o processo de análise e tomada de decisões.

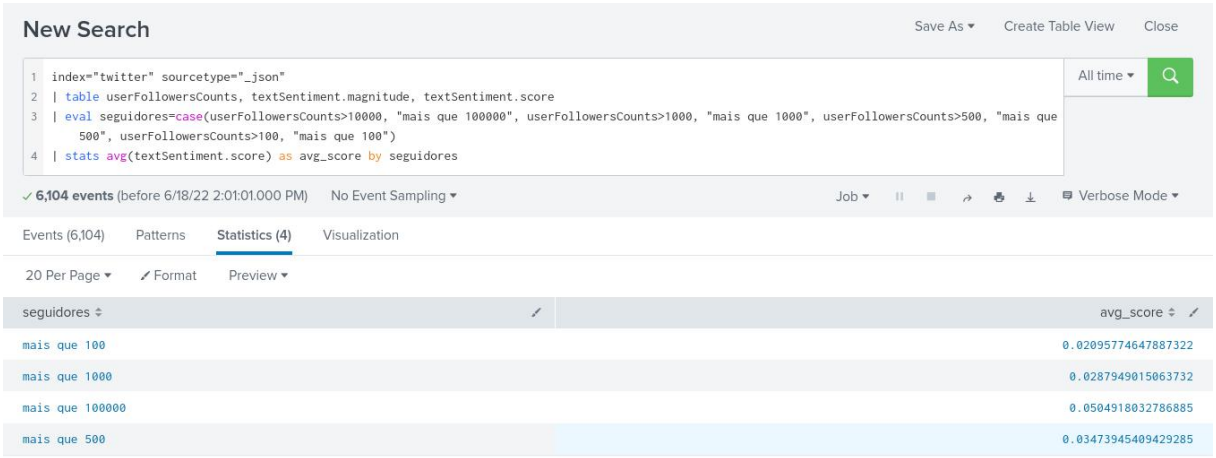


Image 7 - Fonte: Elaborado pelo autor

16 Dados que registram de forma cronológica os eventos de uma aplicação
17 https://splunkbase.splunk.com/

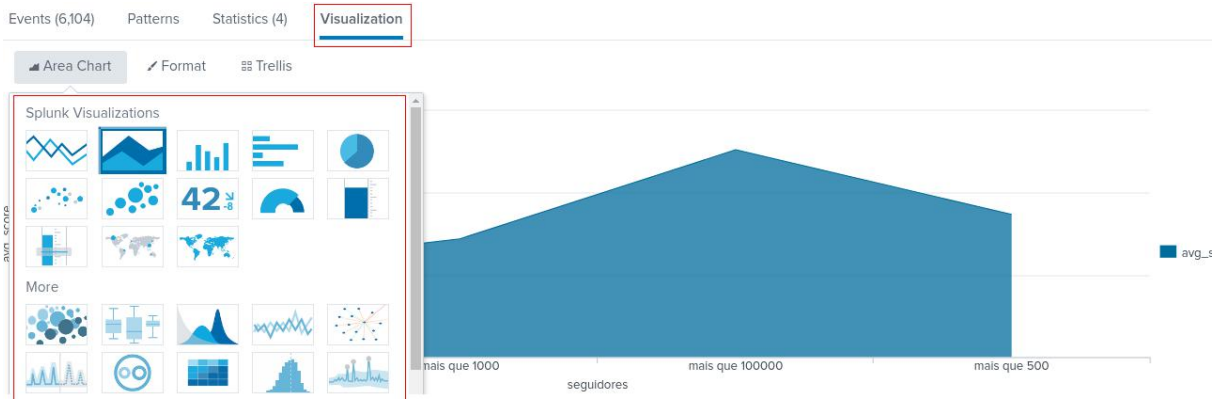


Image 8 - Fonte: Elaborado pelo autor

# 5. Análise do resultado obtido

## 5.1. Sentimento geral do tweets

Para a pesquisa, foram coletados 6085 tweets com comentários contendo a palavra vacina, totalizando 96.164 termos (imagem 9). A busca foi realizada entre os dias 10 de fevereiro de 2022 até o dia 07 de maio de 2022. Todos os dashboards apresentados nessa seção são recortes feitos da ferramenta Splunk. Não há garantia de que o assunto principal dos comentários é de fato a vacina, pois a palavra pode ter sido apenas mencionada em posts com assuntos diversos.


Image 9 - Estatísticas gerais dos dados. Fonte: Elaborado pelo autor

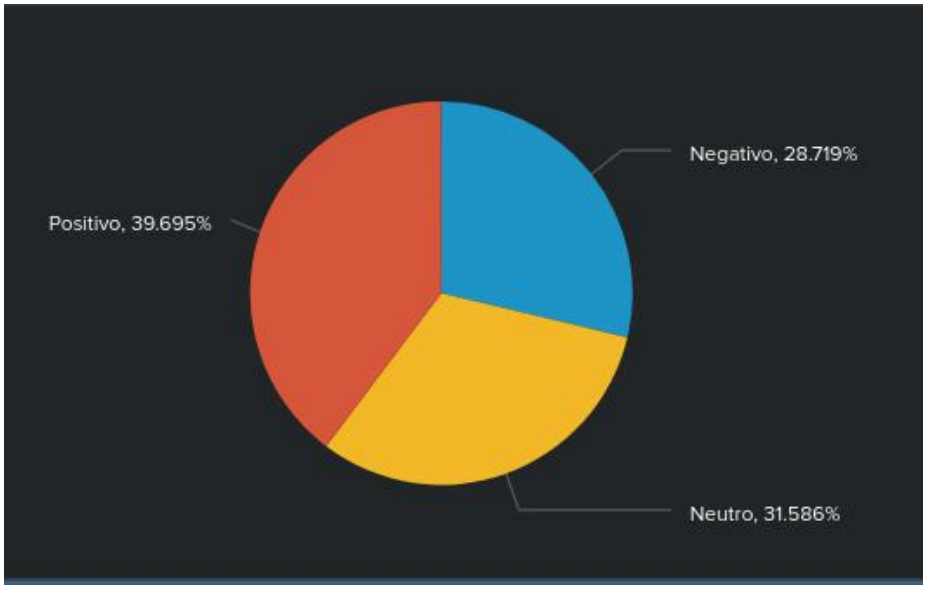


Image 10 - Fonte: Elaborado pelo autor

Como pode ser observado na imagem 10, a opinião das pessoas é significativamente positiva, compreendendo aproximadamente 39% dos comentários. Mesmo com esse resultado, a quantidade de opiniões negativas (29% dos tweets) ainda é grande.

Nas imagens 11 e 12, é possível verificar o trabalho de classificação realizado pelo pipeline de sistemas para gerar o gráfico de opiniões das pessoas nos comentários no Twitter. A imagem 11 é uma amostra de comentários com sentimentos considerados positivos e a imagem 12 de sentimentos considerados negativos.

| text | magnitude | score |
|---|---|---|
| Bom dia, pessoal! É com muita alegria que informamos que estamos há 6 dias sem óbitos por COVID no RN, e também com a menor ocupação de leitos desde o começo da pandemia. A causa tem nome: vacina! Seguimos acreditando na ciência pela vida dos potiguares! Viva o SUS #EquipeFB | 3.4 | 0.5 |
| Excelente notícia. A vacinação no  foi um sucesso. Obrigada SUS, Instituto Butantan e @governosp @jdoriajr . Em São Paulo, tivemos eficiência e organização e fomos os primeiros a sair da pandemia. Mas os cuidados continuam! | 3.4 | 0.3 |
| Vacinas salvam vidas e têm sua eficiência comprovada! Não deixem passar a oportunidade de viver a vida com saúde. A vacina protege você, suas crianças e sua comunidade!#tmjUNICEF#LongLifeForAll@unicefbrasilhttps://t.co/VLinP4QLGr | 3.0 | 0.7 |
| graças à nossa tão amada vacina. Lição aprendida: divirtam-se mas continuem se cuidando. Fui pegar só agora, 2 anos depois estando ivicto, e só tive um resfriado. Obrigado, ciência, e obrigado, Skank! hahaha | 3.0 | 0.6 |

Image 11 - Fonte: Elaborado pelo autor

| text | magnitude | score |
|---|---|---|
| Reação da vacina: me deixou mal pra caramba q isso | 0.8 | -0.8 |
| Meu Deus eu tô muito mal, primeira vez que a vacina do covid me deixa tão ruim desse jeito | 0.8 | -0.8 |
| Tomei a 3 dose tô me sentido doente ' vacina horrível | 0.7 | -0.7 |
| rapa essa vacina me deixou arriado. ta ruim pro meu lado | 0.7 | -0.7 |
| Que ódio que eu tenho de ficar enjoada e não conseguir comer as coisas, reação da vacina | 0.7 | -0.7 |

Image 12 - Fonte: Elaborado pelo autor

## 5.2. Sentimentos divididos por fabricantes das vacinas contra o Coronavírus

Outra análise interessante, evidenciada pelo cenário da pandemia, se baseia no agrupamento das opiniões das pessoas para cada um dos fabricantes de vacinas do COVID-19 reconhecidos no Brasil. O resultado da análise revelou uma preferência da população em relação à vacina Coronavac (score médio 0.2) e uma baixa popularidade da vacina da Janssen (score médio -0.67) (Imagem 13).

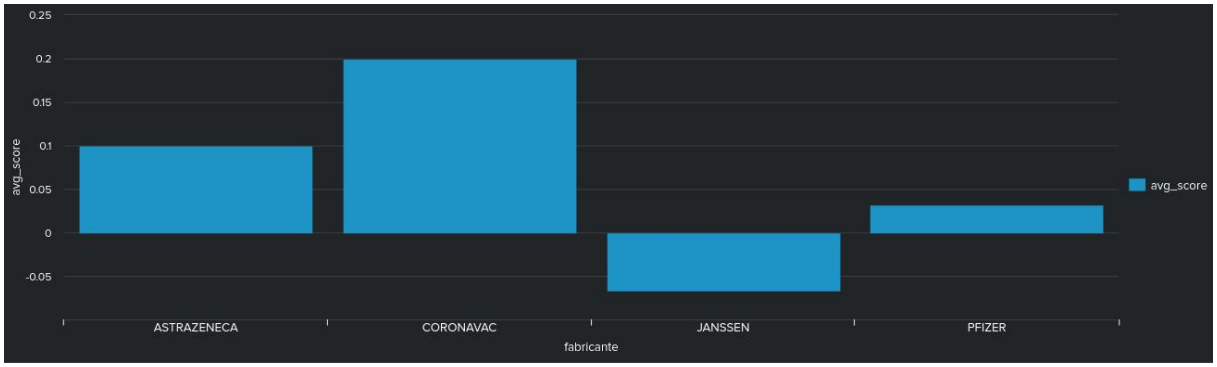


Image 13 - Agrupamento de sentimentos por fabricante. Fonte: Elaborado pelo autor

Analisando os extremos dessa classificação (Coronavac e Janssen) através das nuvens de palavras e de n-gramas, é possível perceber que *o*s comentários sobre a vacina da Janssen se vinculam bastante aos seus sintomas, através de termos como: dolorida, infarto, morte e dor de cabeça (Imagem 14), o que pode explicar sua pontuação mais baixa. Em contrapartida, os comentários relacionados à vacina Coronavac chamam atenção pelas menções ao Instituto Butantã (responsável pela produção da vacina no Brasil) e à ciência (imagem 15), seguida de termos como “eficaz” e “eficiente” que provavelmente elevaram os pontos de sentimentos sobre o fabricante.


Image 14 - Wordcloud e Ngramcloud vacina Janssen. Fonte: Elaborado pelo autor


Image 15 - Wordcloud e Ngramcloud vacina Coronavac. Fonte: Elaborado pelo autor

# 6. Conclusão

As redes sociais são ótimas plataformas para estudos sociológicos. A troca de mensagens, seguidores e sentimentos variados fornecem um ótimo ambiente para entender o comportamento e as opiniões dos usuários no que diz respeito a diversos assuntos.

O estudo apresentado teve como objetivo aplicar ferramentas de análise de sentimentos em comentários sobre a vacina durante a pandemia do Covid-19. Os dados foram coletados em um momento onde a pandemia se encontra mais controlada, entre os dias 11 de fevereiro de 2022 até 28 de abril de 2022. O fim da emergência em saúde pública de importância nacional foi decretado no final do mês de abril do mesmo ano (GOVERNO FEDERAL, 2022).

Dois sistemas foram desenvolvidos para consumir os tweets na plataforma em tempo real (NodeJs) e para processar os dados utilizando os serviços do Google

Cloud Services (Java). Para criação dos dashboards e tabelas foi utilizada a ferramenta Splunk.

Os resultados apresentaram indícios de que a polarização de opiniões em relação à vacina existe, mas cerca de 39% das pessoas demonstram sentimentos positivos sobre os imunizantes, contra 29% dos comentários com carga sentimental negativa. Também foi possível perceber uma preferência da população pela vacina Coronavac, que apresentou a melhor média de pontuação dentre as outras vacinas contra o coronavírus.

Vale ressaltar que, o período de coleta dos dados é um fator que pode influenciar no resultado das análises, tendo em vista que durante a pesquisa boa parte da população já está vacinada e os estudos sobre os imunizantes estão mais aprofundados em relação ao início da pandemia. Desta forma, caso o tempo de extração de dados fosse maior, mais precisos seriam os resultados.

Diversas análises poderiam ser feitas baseadas no banco de dados gerado durante a pesquisa. Uma atividade interessante para expandir a pesquisa seria classificar palavras de maior ocorrência nos comentários positivos e negativos, e coletar mais dados contendo opiniões sobre os diferentes temas encontrados. Outra seria ampliar a coleta dos dados para outras mídias como Instagram e Facebook. Por fim, o sistema gerado como artefato dessa pesquisa poderia ser adotado como modelo para descobrir os sentimentos e emoções da população sobre outros assuntos de importância.

# 7. Referências

LANE, H; HOWARD, C; HAPKE, H. M. (2019). Natural Language Processing in Action - Understanding, analyzing, and generating text with Python, 4-5.

LIU, Bing (2015). Sentiment Analysis - Mining Opinions, Sentiments, and Emotions, 2-8.

FOWLER, Martin, LEWIS, James. Microservices (2014). Disponível em: https://martinfowler.com/articles/microservices.html. Acesso em: 25 abr. 2022.

NADAREISHVILI, I.; MITRA, R; MCLARTY, M.; AMUNDSEN, M..(2016). Microservice Architecture: Aligning principles, practices, and culture, 14.

BELLEMARE, Adam (2020). Building Event-Driven Microservices - Leveraging Organizational Data at Scale, 21-22.

Documentação *API* Natural Language do Google Cloud Services disponível em: https://cloud.google.com/natural-language/docs/.

Documentação Twitter API disponível em: https://developer.twitter.com/en/docs.

O'REILLY, Tim (2007). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Communications & Strategies, 17-37.

EIRINAKI, Magdalini, PISAL, Shamit, SINGH, Japinder (2012). Feature-based opinion mining and ranking. Journal of Computer and System Sciences, 1175–1184.

STEPHENS-DAVIDOWITZ, Seth (2017). Everybody lies - Big Data, New Data and what the internet can tell us about who we really are, 7-24.

EARTHWEB (2022). HOW MUCH DATA IS CREATED EVERY DAY IN 2022? Disponível em: https://earthweb.com/how-much-data-is-created-every-day/#:~:text=15.1)%20Related%20Reading-,Key%20Data%20Creation%20Statistics%202022,photos%20are%20shared%20per%20day.. Acesso em: 26 jun.2022

GOVERNO FEDERAL (2022). Ministério da Saúde declara fim da Emergência em Saúde Pública de Importância Nacional pela Covid-19. Disponível em: https://www.gov.br/saude/pt-br/assuntos/noticias/2022/abril/ministerio-da-saude-declara-fim-da-emergencia-em-saude-publica-de-importancia-nacional-pela-covid-19. Acesso em: 10 jun. 2022.

VACCINES EUROPE (2021). From zero to billions: The story of COVID-19 vaccines. Disponível em: https://www.vaccineseurope.eu/from-zero-to-billions-the-story-of-covid-19-vaccines. Acesso em: 26 jun. 2022.

SANTOS, Vanessa Sardinha dos (2022). "História da vacina"; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/biologia/a-historia-vacina.htm. Acesso em: 28 de jun. 2022.

# Anexo 1

## Repositórios dos sistemas

https://github.com/IgorCooli/twitter-consumer
https://github.com/IgorCooli/twitter-analysis-service

## Endpoins e configurações dos sistemas:

**Endpoints:**

`URL_SISTEMA/run` -> Começa o processo de coleta dos dados

---

**Variáveis de ambiente**

| Chave | Valor |
|---|---|
| ACCESS_TOKEN | {Token Twitter API} |
| ACCESS_TOKEN_SECRET | {Token Secret Twitter API} |
| CONSUMER_KEY | {Consumer Key Twitter API} |
| CONSUMER_SECRET | {Consumer Secret Twitter API} |
| REDIS_BASE_URL | {Redis URL} |
| REDIS_PASS | {Redis Password} |
| REDIS_PORT | {Redis Port} |
| SEARCH_STRING | {Search String} |
| TIMER | {Search Time} |

**Endpoints:**

`URL_SISTEMA/tweet/execute` -> Retorna todos os tweets já analisados

`URL_SISTEMA/text?text={texto para análise}` -> Realiza a análise de sentime

---

**Variáveis de ambiente**

| Chave | Valor |
|---|---|
| GOOGLE_APPLICATION_CREDENTIALS | {Path do GCloud Credentials} |
| MONGO_DB_NAME | {MongoDB Name} |
| MONGO_PASS | {MongoDB Password} |
| MONGO_PORT | {MongoDB Port} |
| MONGO_URL | {MongoDB URL} |
| MONGO_USER | {MongoDB User} |
| REDIS_PASS | {Redis Password} |
| REDIS_PORT | {Redis Port} |
| REDIS_URL | {Redis URL} |