Tweet Utils UI: Uma ferramenta para coleta e visualização de tweets

Pedro P. Wagner
Escola Politécnica
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul
pedro.wagner00@edu.pucrs.br

Isabel H. Manssour
Escola Politécnica
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul
isabel.manssour@pucrs.br

***Resumo*—A cada dia que passa mais empresas e pesquisadores estão interessados na coleta e análise de dados da rede social Twitter, com o intuito de analisar e entender o comportamento e a opinião das pessoas na internet. Sendo assim, neste trabalho propomos o projeto e o desenvolvimento de uma aplicação que possibilite coletar dados do Twitter a partir de um conjunto de palavras e *hashtags* ou por usuários específicos, e disponibilize um conjunto de visualizações para auxiliar na análise destes dados. Os tweets são pré-processados para geração de algumas estatísticas e as representações visuais são geradas de forma automática. O objetivo é que todo o processo seja feito dentro de uma interface simples, com um visual agradável e fácil de usar, sem que seja necessário ter conhecimentos de programação para coletar e explorar dados do Twitter.**

***Palavras-chave*—Coleta e visualização de dados, Twitter**

I. Introdução

As redes sociais, nos dias atuais, estão cada vez mais em ascensão com milhares de postagens por dia [1]. Uma das mais utilizadas é a rede social denominada Twitter[1], através da qual é possível realizar publicações (convencionalmente chamadas de tweet) com 280 caracteres. Chegando a publicar cerca de 500 milhões de tweets por dia, essa rede social começou a atrair a atenção de muitos pesquisadores de diversas áreas a fim de estudar o comportamento das pessoas, que publicam desde um *feedback* de um *reality show*, até uma notícia. Exemplos de estudos feitos nesta rede social incluem análise de sentimento [7] e predição de resultado de eleições [11].

A área de visualização consiste em criar uma representação visual dos dados. Ela é muito utilizada por pesquisadores de diversas áreas pelo fato de facilitar a exploração e análise dos dados, principalmente se houver uma grande quantidade de dados, pois o uso de gráficos facilita a compreensão dos dados de uma maneira rápida e prática. Pessoas com cargos de decisões, por exemplo, podem se beneficiar das visualizações para tomada de decisão, escolhendo caminhos mais lógicos com ideias a partir de fatos. A cada dia que passa, mais empresas adotam essa cultura de realizar decisões a partir de dados [2] (cultura *Data-Driven*) e pesquisadores que não utilizavam tantos dados agora estão começando a usufruir de milhares deles. Entre estes dados estão dados estatísticos sobre biologia, dados abertos do governo, dados de redes sociais e dados de *marketing* e de administração, todos eles sendo usados com o objetivo de se obter conclusões mais precisas e racionais. A fim de auxiliar ainda mais, os gráficos gerados em uma versão interativa se tornam ainda mais atraentes pelo fato de amplificar a qualidade da visualização já que isso permite uma certa independência na análise dos dados para o usuário final. Entre essas interações estão os filtros de dados e o *zoom* por exemplo.

A ascensão da notoriedade do Twitter somada com à implementação da cultura *Data-Driven* fez com que muitos pesquisadores que não estão ligados à áreas da tecnologia da informação se interessassem pelo assunto. E por isso, consequentemente, muitos deles não sabem programar, levando à necessidade de ter uma ferramenta para ajudar na coleta e na análise visual de dados, o que agregaria muito valor para suas pesquisas.

Considerando este contexto, o objetivo deste trabalho foi criar uma ferramenta, chamada Tweet Utils UI[2], que tenha uma interface com o intuito de facilitar e deixar mais amigável a extração e a visualização de tweets. Por isso, o projeto também foca na criação de uma interface mais elegante e intuitiva uma vez que a aplicação poderá ser utilizada por usuários sem conhecimento de programação, além de incluir técnicas de interação para deixá-la mais prática.

Na próxima seção são apresentados alguns trabalhos relacionados à pesquisa proposta. O problema de pesquisa, a metodologia de desenvolvimento e a arquitetura inicialmente desenvolvida, incluindo os elementos de interface e visualizações que foram utilizados na aplicação são apresentados na Seção III. Por fim, são apresentadas as conclusões e são descritos alguns trabalhos futuros para aprimorar a ferramenta.

II. Trabalhos Relacionados

Existem diversos trabalhos com a finalidade de auxiliar na coleta de tweets [6] [12] [13], enquanto outros têm como objetivo a visualização deste tipo de dado [3]–[5]. Por exemplo, o trabalho de Kishore et al. [9] trata-se de uma ferramenta de coleta de tweets. Nele se discute também o custo e a qualidade na coleta de tweets para as pesquisas, e que o uso de API (*Application Programming Interface*) em vez de *webscrapping*

[1]https://twitter.com

[2]Disponível em: https://github.com/DAVINTLAB/Tweet-Utils-UI

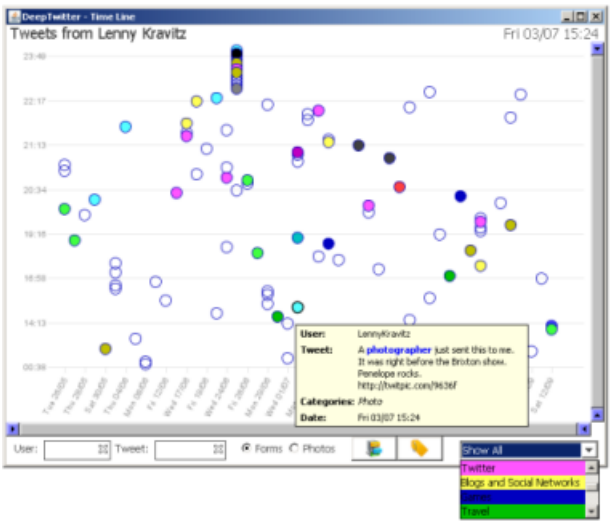

Figura 1. Visualização utilizada no trabalho de Rotta et al. [8]

garante uma alta qualidade nos dados na maioria dos casos. Além disso, a coleta de tweets pode resultar em uma grande quantidade de dados, e a API em questão é fornecida pelo próprio Twitter.

O trabalho de Rotta et al. [8] trata da importância da visualização de dados para a análise e o entendimento de dados do Twitter. Como o volume de tweets nessa rede social cresce de forma muito rápida, principalmente pelo fato de um usuário A receber automaticamente e retransmitir mensagens de um outro usuário B apenas seguindo-o, o uso de técnicas de visualização pode auxiliar na análise do comportamento dos usuários. A figura 1 mostra uma visualização utilizada no trabalho de Rotta et al. que consiste em um gráfico de dispersão que mostra os tweets enviados em cada dia e horário. As cores são usadas para ilustrar a classificação do tweet, por exemplo, se é sobre viagens ou jogos, e ao passar o mouse sobre o círculo que representa o tweet é possível ver o seu conteúdo.

Outros trabalhos de visualização focam em um assunto mais específico, tais como análise de tweets relacionados a programas de televisão ou análise de sentimento. Sanvido et al. [10] apresentam uma ferramenta para visualização de tweets sincronizados com vídeos de programas de televisão, tais como *reality shows*, novelas e jogos de futebol chamada PeakVis. O objetivo é, a partir do fornecimento de um arquivo com tweets coletados e de um vídeo, investigar o que os telespectadores estão comentando sobre o programa em questão para tirar conclusões sobre a transmissão. A interface desta ferramenta, chamada PeakVis, está exemplificada na figura 2, que mostra um gráfico do total de tweets sincronizado com o vídeo e uma lista de tweets na qual é acrescentado um tweet por minuto.

Florence et al. [7], por sua vez, apresentam uma ferramenta de visualização voltada apenas para a analise de sentimentos. Os autores também abordam a importância do estudo de dados de rede sociais em eventos globais ou em eleições políticas, e como o Twitter, por exemplo, influencia nesses assuntos.

Sharma e Rana [14] também propuseram a criação de uma aplicação voltada para analise de sentimentos, mas que realiza a extração de tweets junto com duas principais visualizações. Nesse trabalho é possível, além de exibir o conjunto de dados coletados, visualizar um gráfico voltado à analise de sentimentos e um gráfico no estilo *wordcloud*. A interface proposta é bem simples e os gráficos não têm suporte para interação com o usuário. Esta interface está exemplificada na figura 3. A maior restrição desse projeto é que há um limite na coleta de tweets. É possível extrair no máximo 1000 tweets e esses tweets podem ser coletados apenas através de *hashtags* ou de um perfil específico.

Tabela I
COMPARATIVO DOS TRABALHOS RELACIONADOS

Artigo	Ano	Tipo de Coleta	Visualização	Assunto
[5]	2011	-	Sim	Analise de sentimento
[6]	2012	Crawler	Não	Geral
[4]	2013	-	Sim	Geral
[8]	2013	-	Sim	Geral
[3]	2014	-	Sim	Geral
[7]	2015	-	Sim	Analise de sentimento
[12]	2016	*rest API*	Sim	Geral
[9]	2019	*rest API*	Não	Geral
[14]	2020	*rest API*	Não	Analise de sentimento
[10]	2021	-	Sim	Geral
[13]	2021	Crawler	Não	Geral

A tabela I sintetiza os trabalhos relacionados descritos nesta seção, identificando-os por ano; se caso há coleta qual tipo usada; se há visualização; e se há um escopo proposto no trabalho ou não (geral). Analisando estes trabalhos, podemos perceber que nenhum deles disponibiliza uma interface que permita tanto a coleta como a visualização de tweets sem a necessidade de programar e sem grandes limitações. Nos trabalhos com foco na coleta de dados, diversas formas de coleta são apresentadas, não havendo um padrão e em alguns casos não garantindo a qualidade dos dados (fator importante discutido em alguns dos trabalhos [9] [12]).

III. DESCRIÇÃO DO TWEET UTILS

Considerando os trabalhos apresentados na Seção II, o objetivo deste trabalho é criar uma ferramenta que permita a coleta e a visualização de tweets. Para a coleta, foi utilizada a API do Twitter (Tweepy[3]), que é um método legal, judicialmente falando, para extração de tweets, garantindo qualidade e veracidade nos dados obtidos. Para a visualização foram criados gráficos que auxiliam na análise qualitativa e quantitativa, permitindo a obtenção de *insights* sobre os dados coletados.

A solução proposta neste trabalho seria de grande ajuda para pesquisas como a de Sanvido et al. [10], uma vez que suprimiria o problema da coleta de dados e junto a isso seria possível analisar rapidamente se a repercussão de um determinado

[3]https://www.tweepy.org

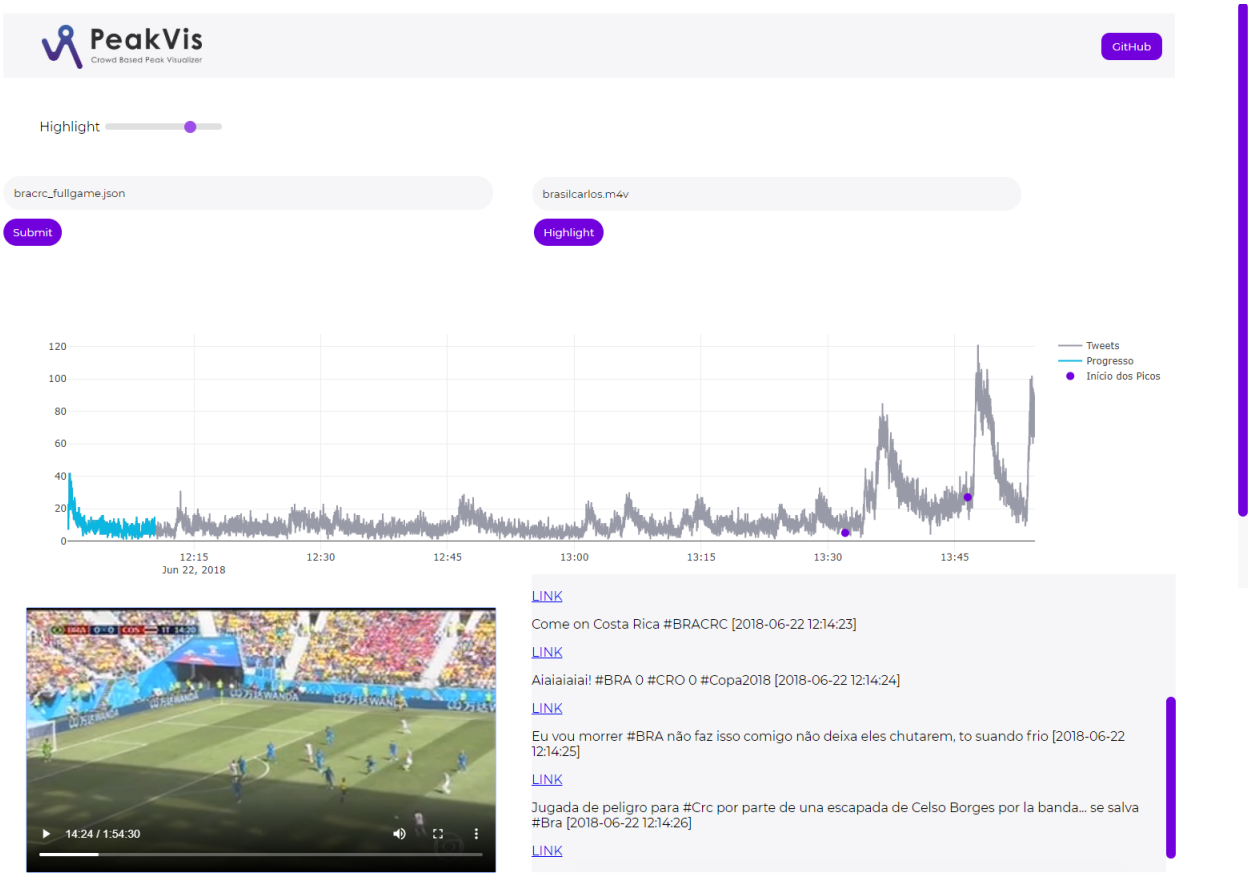

Figura 2. Interface do Peakvis [10].

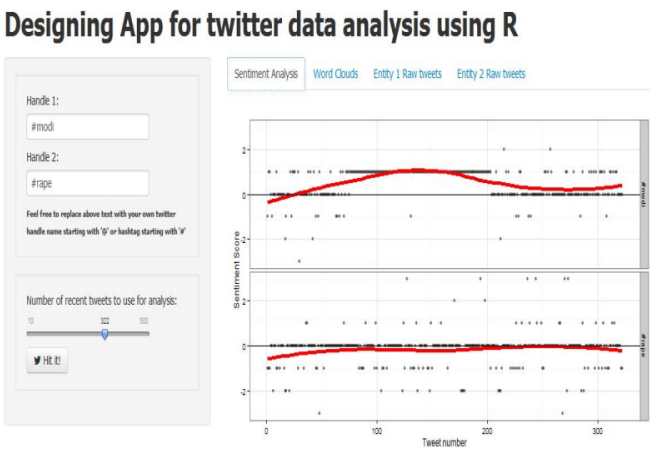

Figura 3. Interface proposta no trabalho de Sharma e Rana [14]

assunto foi da grandeza estimada (tarefa que provavelmente não é trivial dependendo da transmissão). Além disso, poderia auxiliar na avaliação dos dados com visualizações que não estão disponíveis na plataforma já utilizada, corroborando fortemente nas conclusões das pesquisas e, possivelmente, promovendo novos *insights* desses ou de novos dados a serem coletados. Nas próximas seções são detalhadamente descritas todas as etapas seguidas para o desenvolvimento deste projeto.

A. *Motivação e Metodologia de Trabalho*

Sabendo a priori que há muitos pesquisadores em busca de estudar e analisar o que os usuários do Twitter escrevem diariamente, e que desses pesquisadores muitos não são da área da computação e tecnologia, o problema da pesquisa consiste em criar uma aplicação que possua uma interface que permita fazer a coleta e uma análise visual interativa de um conjunto de tweets. Essa interface tem como objetivo facilitar e desmistificar a extração de dados do Twitter e a geração automática de visualizações destes dados. Assim, é possível tirar algumas conclusões mais rápidas sobre o que foi extraído, como, por exemplo, o fluxo de tweets envidados em um período de tempo, as palavras mais utilizadas, os autores dos tweets com mais repercussão, entre outros. A motivação para o desenvolvimento deste trabalho surgiu a partir do entendimento das dificuldades de pesquisadores da área da comunicação e da administração, normalmente sem conhecimentos de programação, para a utilização de *scripts* em *prompt* de comando.

A metodologia utilizada para o desenvolvimento deste trabalho foi sintetizada na figura 4. Inicialmente, foi feito um estudo dos trabalhos relacionados, apresentado na Seção II, para analisar quais aspectos são importantes e fundamentais para

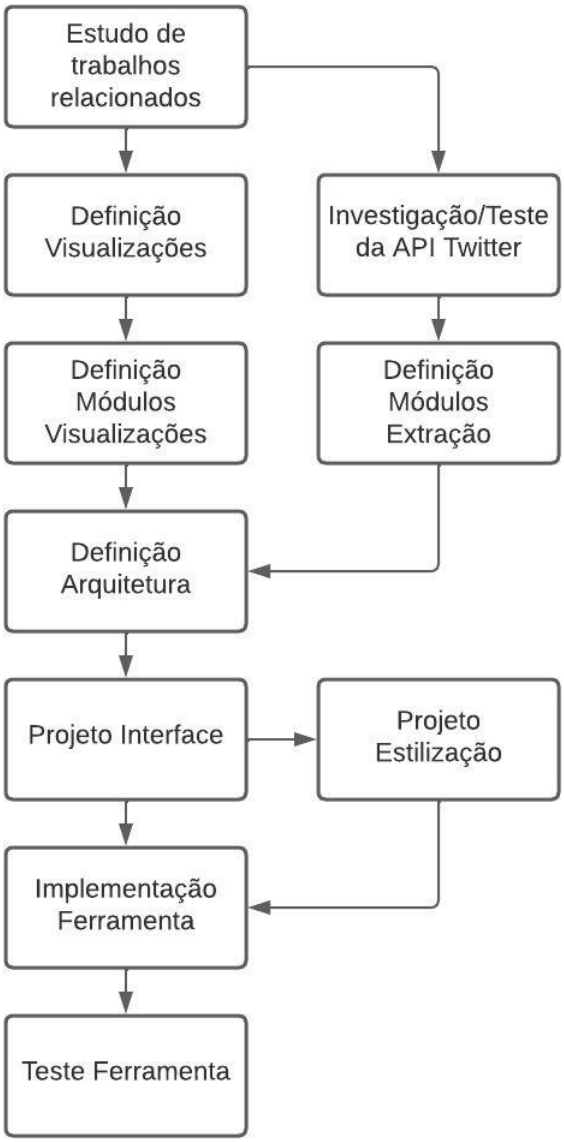

Figura 4. Diagrama da metodologia proposta.

uma extração e visualização de dados. Esse estudo possibilitou a escolha das visualizações que serão disponibilizadas na interface, do porque elas são importantes e o que elas mostram. Além disso, foi possível entender como funciona a extração de dados a partir da API Tweepy e como é possível utilizá-la.

Sendo capaz de extrair dados do Twitter e sabendo com quais técnicas de visualização utilizá-los, foi possível criar os módulos da interface. Os módulos de extração têm o objetivo de gerar *datasets* de conjuntos de tweets a partir de alguma "consulta" como parâmetro. Os módulos de visualização têm por objetivo gerar automaticamente os gráficos solicitados pelo usuário a partir do *dataset* gerado no módulo anterior.

Com os módulos bem estruturados, foi possível definir a arquitetura do projeto. Essa parte é muito importante, pois trata de como funciona o fluxo de utilização da aplicação e como ele deveria ser implementado. Assim, com a arquitetura bem formulada, foi possível pensar na interface do usuário, como ela seria exibida e como seria seu estilo.

Por fim, com a interface final e a aplicação inteira implementadas, foi possível realizar testes a fim de validar todos os passos da metodologia proposta.

B. Ambiente de Desenvolvimento

Para o desenvolvimento da ferramenta foi utilizada a linguagem de programação Python juntamente com algumas API's que visam facilitar a sua implementação. A API Tweepy foi usada para a coleta de dados do Twitter de uma maneira lícita. A interface foi criada usando o *framework* Pyside6[4] e os gráficos foram gerados usando a Plotly[5] e o WorldCould2[6].

Optamos pela utilização da biblioteca Plotly pelo fato dela já ter implementada várias funcionalidades básicas de interação com as visualizações, tais como zoom e movimentação do gráfico. Porém, como essa API, mesmo na sua versão em Python, gera o *output* em um arquivo html, decidimos gerar todas as visualizações em uma página html. Já a biblioteca wordcloud2.js está sendo utilizada porque a Ploty não dá suporte para geração de *worcloud* (ou nuvem de palavras).

Mesmo existindo uma versão da biblioteca de extração de dados do Twitter em Javascript, optamos pela criação da plataforma e a extração de tweets em Python porque a arquitetura do projeto como um todo é mais fácil de estruturar. Além disso, a disponibilidade de uma aplicação facilita o uso para quem não tem conhecimento de programação ou quando não se tem um servidor disponível, além da familiaridade do autor com Python comparado ao Javascript. A biblioteca Pyside6 foi selecionada a fim possibilitar a implementação de uma interface elegante e simples. Sua estilização é feita de maneira fácil, além de ter elementos de animação.

C. Arquitetura

Pensando inicialmente na arquitetura do projeto como um todo, foram projetados módulos de coleta e tratamento desses dados, uma vez que esses são considerados fundamentais. Entre esses módulos estão:

- *Gather Profile*: Coleta *tweets* específicos de um perfil do Twitter, devolvendo esses dados em um JSON.
- *Rest Gathering*: Coleta *tweets* a partir de uma consulta na qual podem ser utilizadas palavras e/ou *hashtags* juntamente com operadores lógicos. O intervalo de tempo da pesquisa pode mudar conforme as chaves fornecidas pelo usuário.
- *Quick Report*: Utiliza uma coleta já realizada como *input* a fim de criar pequenas estatísticas sobre o *dataset*, como número de *tweets* coletados, intervalo de tempo coletado, palavras mais utilizadas, entre outros.
- *Sanitize Tweets*: Utiliza uma coleta já realizada como *input* a fim de limpar os tweets removendo "*stop words*", isto é, artigos, preposições e afins. Além disso, possibilita

[4]https://doc.qt.io/qtforpython/contents.html
[5]https://plotly.com
[6]https://wordcloud2-js.timdream.org

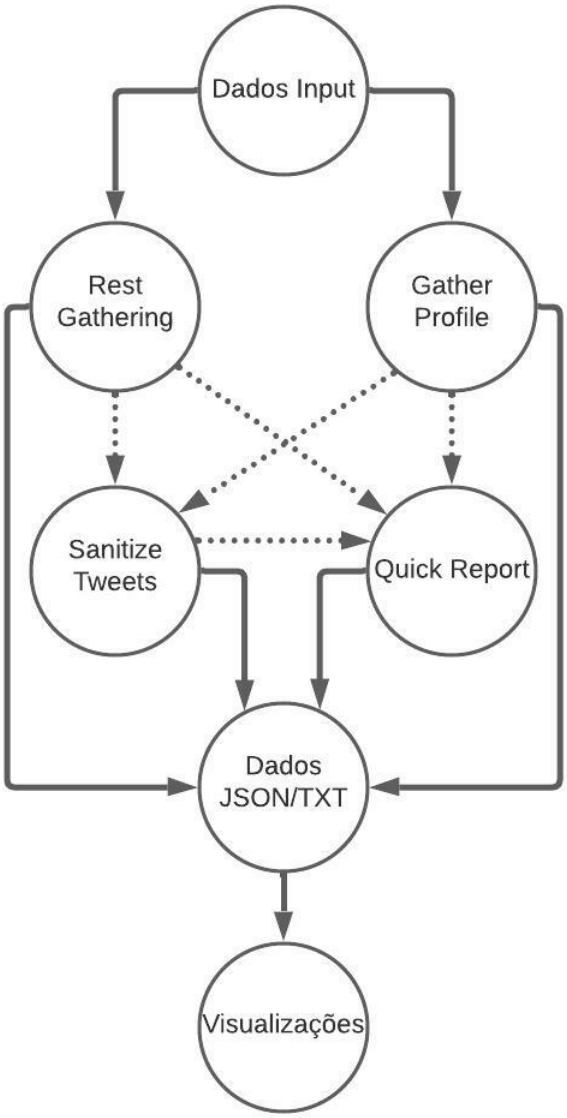

Figura 5. Diagrama da arquitetura do projeto.

a remoção de *retweets* e emojis, retornando um JSON mais enxuto.

Esses módulos estão disponibilizados na plataforma a partir de uma barra de menu na qual é possível selecionar o módulo a ser utilizado. Os *inputs* para esses módulos são os dados necessários para o disparo dos mesmos, sendo que em cada módulo há diferentes tipos de entrada a serem solicitadas, com exceção da chave do usuário que será utilizada em todos os módulos de *gathering* uma vez que a API para a coleta oficial do Twitter a utiliza. Em cada módulo na interface são exibidos todos os campos obrigatórios e opcionais, para que não haja erro na execução dos *scripts*.

A arquitetura projetada está resumida na figura 5, na qual todos os módulos de coleta têm seu *input* e todos eles devolvem um JSON, sendo possível passá-los pelos módulos de tratamento, caso necessário, e após sendo possível gerar as visualizações.

D. Interface

A interface foi elaborada com o objetivo de ser simples, amigável e com uma aparência agradável, possibilitando uma fácil navegação entre os módulos propostos e buscando evitar erros ou dúvidas de uso da aplicação. Para isso foi disponibilizado um menu lateral, através do qual é possível selecionar o módulo desejado para que apareçam as funcionalidades relacionadas a ele. Os módulos disponíveis são: a janela para a inserção da chave para a coleta de tweets; os módulos de coleta *Rest Gathering* e *Gather Profile* citados na seção III-E; os módulos de tratamento de tweets denominados *Quick Report* e *Sanitize Tweets* citados na seção III-C; um módulo para seleção e geração das visualizações; e, por fim, o módulo de exibição dos gráficos gerados. A figura 6 ilustra a interface que é apresentada quando a aplicação é carregada.

Cada módulo tem os campos necessários para a execução da sua tarefa. Esses campos foram projetados visando evitar erros por parte do usuário, e tentando deixar claro quais são opcionais e a função de cada módulo. Para isso, foram disponibilizados botões de ajuda e definidos rótulos intuitivos. Nos módulos de coleta, tratamento e criação de gráficos a partir de dados há uma espécie de *prompt* de comando, com o objetivo de manter o usuário ciente do estado atual do *script*.

Para o tratamento dos dados são disponibilizados dois módulos. O primeiro é o *Sanitize Tweets* que serve para a remoção de *stopwords*. Esse módulo utiliza um arquivo JSON feito a partir de um dos *scripts* de coleta e devolve como saída o mesmo JSON sem artigos, preposições, entre outros. Como parâmetros adicionais é possível escolher o nome do arquivo de saída, além de ter a opção de remover emojis e retweets.

O segundo módulo de tratamento é o *Quick Report*, que utiliza com arquivo JSON também gerado a partir de um dos módulos de coleta a fim de gerar algumas metadados sobre o *dataset*. A partir dele é possível descobrir: o número de tweets coletados; a data do tweet mais antigo e mais recente do *dataset* em questão; e quais foram as palavras, usuários e *hashtags* que mais apareceram na coleta, o número de dados que aparecem no arquivo de saída pode ser modificado pelo usuário. Por fim há um parâmetro opcional para modificar o nome do *output*.

O módulo de visualizações foi dividido na interface em duas partes, uma com o objetivo de criar os gráficos (menu *Visualizations*) e a outra de exibi-los (menu *Dashboard*). No menu *Visualizations* é possível selecionar o conjunto de dados junto com o tipo de visualização desejado. É permitido gerar apenas um tipo de visualização por vez, porém as visualizações selecionadas não precisam ser do mesmo *dataset* podendo gerar uma figura com um *dataset* A e outra com um *dataset* B. A figura 7 ilustra o módulo de criação de visualizações como um todo.

E. Coleta de Tweets

A coleta de tweets, como já mencionado, foi feita a partir da biblioteca Tweepy. Por ela ser uma API original do Twitter,

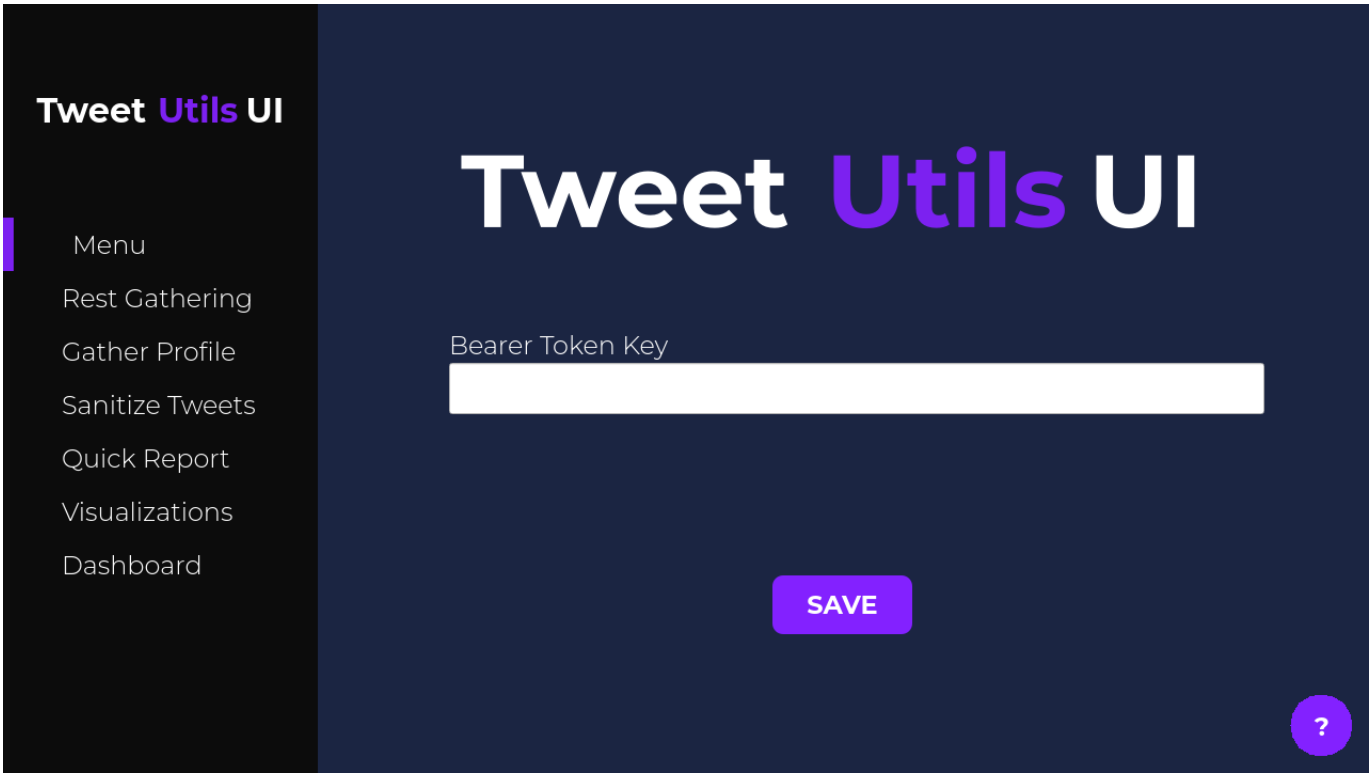

Figura 6. Interface do Tweet Utils UI.

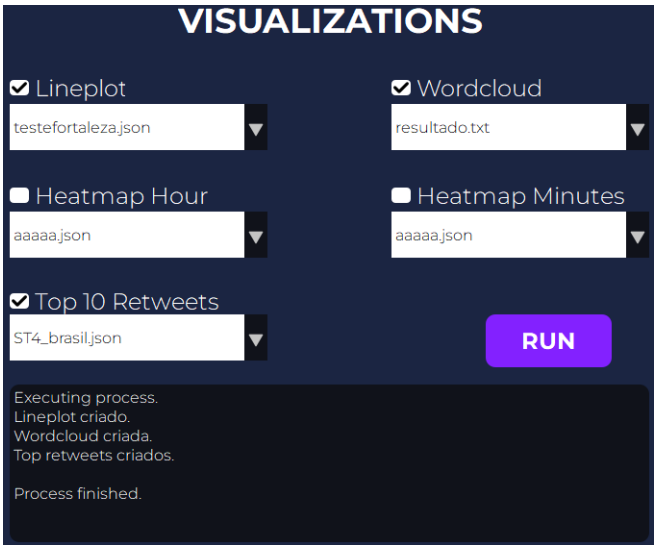

Figura 7. *Inputs* do módulo de criação de visualizações da ferramenta.

é necessário solicitar acesso para sua utilização. Normalmente o Twitter permite o uso para estudantes e pesquisadores, porém é mandatório responder um conjunto de perguntas para isso, como, por exemplo, o objetivo do acesso, expectativa da quantidade de tweets desejados nas coletas, entre outros.

Dependendo das respostas do questionário, o Twitter fornecerá as chaves para a extração de dados a partir de uma *rest API*, e cada chave tem um certo limite de tweets que podem ser coletados. Por isso, cada usuário deverá obter suas próprias chaves, além de ser responsável pelo conhecimento das suas limitações de extração de tweets.

A API disponibiliza diversas requisições *rest* para a manipulação e extração de dados. Porém, como citado na seção III-C, foram utilizadas apenas as requisições de perfil e consultas específicas. Normalmente essas pesquisas podem coletar tweets com uma distância de nove dias após sua publicação, mas esse tempo também difere para cada chave. Os *inputs* necessários para essas requisições são as entradas mandatórias dos módulos de coleta implementados na interface.

Para o módulo de *Rest Gathering*, além de palavras chave e/ou *hashtags* preenchidas pelo usuário no campo "query"da janela, é necessário definir o intervalo de datas para realizar a pesquisa. Junto a isso é possível definir o nome do arquivo de saída e o limite de tweets a serem coletados, mas ambos são *inputs* facultativos. Os *inputs* do módulo *Gather Profile* são mais enxutos, já que é necessário fornecer apenas o ID ou nome do usuário (no twitter, o nome do usuário sempre vem depois de uma arroba) e, opcionalmente, o nome do *output* do módulo. Os campos de entrada dos módulos *Rest Gathering* e *Gather Profile* estão representados na figura 8.

F. Visualizações

Após analisar as visualizações disponíveis nos trabalhos relacionados, foram escolhidas três visualizações para serem

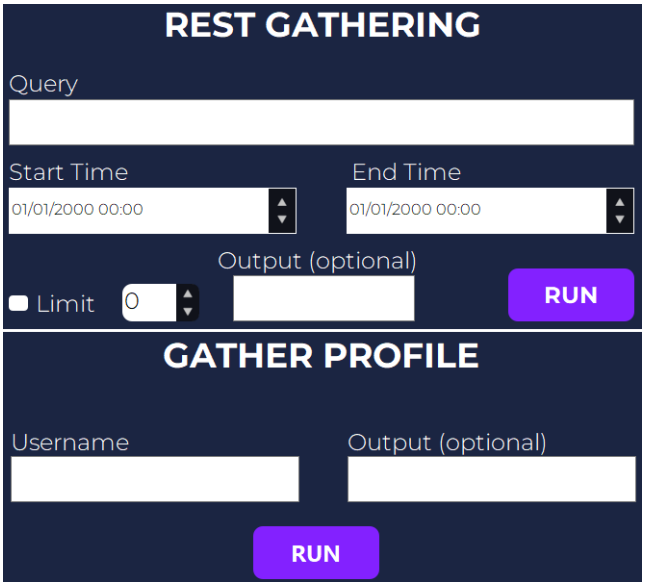

Figura 8. *Inputs* dos módulos *Rest Gathering* e *Gather Profile* dentro da interface respectivamente.

incluídas, além de um relatório: *lineplot*, *heatmap*, *wordcloud* e relação de tweets com mais *retweets*.

O *lineplot* serve para avaliar o fluxo de tweets por segundo da coleta realizada. Neste caso, o eixo X é o horário e o eixo Y é o número de tweets que foram coletados naquele horário. Normalmente ele é utilizado para analisar e validar a repercussão do assunto que está sendo estudado e, assim, verificar se ele é relevante ou se seria necessário trocar as palavras-chaves na parte de *gathering*, por exemplo.

O *heatmap* serve para complementar o *lineplot* no sentido de trazer uma análise mais abrangente em relação ao fluxo de tweets. Ele mostra quantos tweets foram coletados por hora durante o intervalo de busca realizado. Cada linha do *heatmap* corresponde aos dias pertencentes ao intervalo da coleta e as colunas, num total de vinte e quatro, estão relacionadas às respectivas horas do dia. O *heatmap* pode servir para um possível "corte"no *dataset*, eliminando horas nas quais não houve a repercussão esperada, por exemplo. Junto ao *heatmap* por hora há também a opção de apresentar um por minuto, com a única diferença de ser uma coluna por minuto (totalizando sessenta em vez de vinte e quatro).

A *wordcloud* auxilia na análise de quais palavras são mais utilizadas pelos usuários do Twitter considerando a coleta realizada. Este gráfico é literalmente uma "nuvem de palavras", na qual o tamanho das palavras é diretamente relacionado ao número de vezes que ela apareceu no *dataset*. Esse gráfico serve para ter uma confirmação mais rápida se os usuários estão comentando sobre um determinado fato que já passou e que chamou a atenção da maioria, ou se os comentários dos usuários acompanham o decorrer de um programa, por exemplo.

Por último há a relação dos tweets em destaque, que correspondem aos dez tweets mais retweetados da coleta. Quando um tweet é retweetado todos os seguidores da pessoa que retweetou passam a recebê-lo, sendo suscetível a ser cada vez mais retweetado por outros usuários. Caso no conjunto de tweets coletado não haja dez tweets retweetados, ele exibirá o número máximo de tweets retweetados.

IV. Estudo de Caso

Um estudo de caso foi realizado a fim de validar e demonstrar o uso da ferramenta como um todo. O tema escolhido foi o lançamento da quarta temporada de *Stranger Things*[7] pela Netflix. A opção por este tema foi pelo fato dessa série ser bastante aclamada por diversos usuários, e, portanto, havia a probabilidade de haver centenas de milhares de tweets entre os dias de seu lançamento.

A série em questão foi lançada às sete da manhã horário UTC (*Universal Time Coordinated*) do dia vinte e sete de maio de 2022 com sete episódios, totalizando por volta de nove horas de duração. Por isso, o intervalo da coleta realizada começou um dia antes, para que pudesse ser analisado o aumento do tráfego de tweets antes e depois do lançamento junto com o potencial interesse na série antes do lançamento. O final do intervalo da coleta foi de três dias após o lançamento, para que houvesse tempo dos espectadores assistirem a série e fazerem seus comentários no Twitter. Assim, o intervalo da busca foi do dia vinte e seis até o dia trinta de maio, totalizando mais ou menos quatro dias. As palavras-chave utilizadas na coleta foram as *hashtags* #StrangerThings e #StrangerThings4.

Com esse intervalo de busca, selecionando apenas por tweets da língua portuguesa (esse filtro foi feito por um *script* auxiliar externo), foram coletados 65.665 tweets únicos, totalizando 290.890 contando com retweets. Gerando o *lineplot*, *heatmap* de hora, *wordcloud* e *top* 10 retweets foram obtidas as visualizações da figura 9. A partir do *lineplot* é possível perceber um fluxo relativamente grande de tweets, principalmente depois do lançamento da série. Esse fluxo diminui drasticamente apenas nas madrugadas dos próximos dias.

A utilização do *heatmap* de horas foi de grande ajuda nesse estudo, uma vez que foi possível analisar facilmente o fluxo de tweets de forma mais detalhada, principalmente nos momentos de maior fluxo de tweets. Assim foi possível inferir que os maiores horários de repercussão foi na parte da noite, em torno das vinte e uma horas até meia noite. Vale ressaltar que os horários estão em formato UTC, portanto os valores estão transladados três horas na frente em relação ao horário de Brasília.

A *wordcloud* nesse estudo de caso, mostrou que a maioria dos usuários comentam de forma geral sobre a série, não abordando comentários ou discussões específicas de alguma cena ou episódio, por exemplo.

Com os tweets mais retweetados foi possível entender de forma mais específica o que mais repercutiu nas redes sociais. A maioria dos usuários simpatizaram com memes feitos a partir de personagens específicos e de uma cena provavelmente

[7]https://www.netflix.com/br/title/80057281

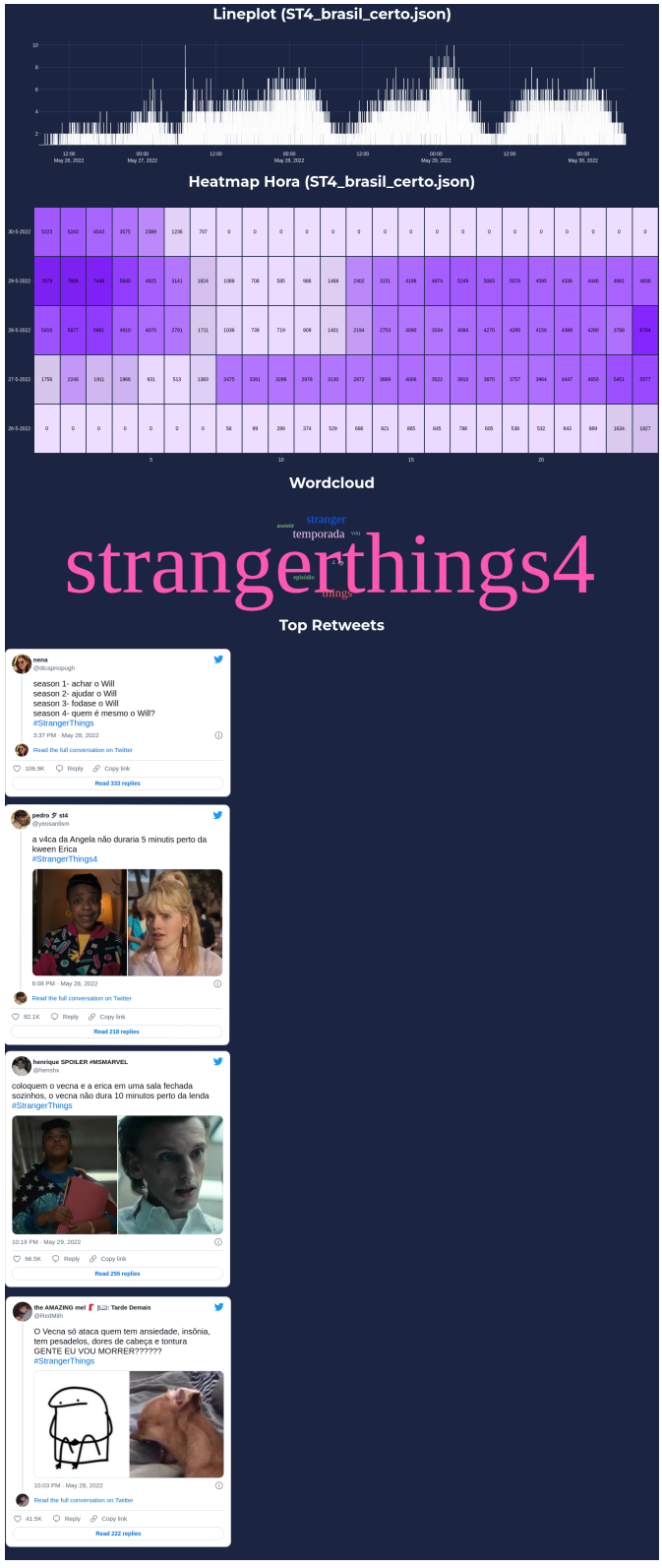

Figura 9. Visualizações obtidas dos tweets da quarta temporada da série Stranger Things.

de grande impacto na temporada. A partir de um desses tweets foi possível assistir o trecho dessa cena específica, apenas assistindo-o pode-se perceber a sua relevância. Porém, ela não é auto explicativa, assim sendo possível de entendê-la apenas junto ao seu contexto.

Sumarizando o estudo, podemos perceber a grande capacidade de coleta de tweets da ferramenta, além de conseguir entender um pouco sobre a opinião dos espectadores de *Stranger Things*. Mesmo não sabendo nada a priori da série, pode-se entender que foi uma temporada bem aguardada pela boa repercussão que causou nos primeiros dias após seu lançamento. Junto a isso foi possível perceber uma grande parcela de comentários positivos da temporada junto com algumas sátiras sobre o rumo incoerente da série em alguns momentos. A maioria dos seus espectadores brasileiros provavelmente preferiram assistir os novos episódios no turno da noite, o que pode ser apenas um horário de preferência ou o horário que a maioria deles têm disponível para assistir, deixando aberto mais um ramo de estudo antropológico.

V. Conclusões e Trabalhos Futuros

Após a implementação do projeto pode-se afirmar que o objetivo de criar uma aplicação para coletar dados da rede social Twitter a fim de desmistificar e facilitar esse processo, auxiliando qualquer tipo de usuário que necessite foi atingido. A ferramenta ficou com uma aparência simples e intuitiva. A coleta de tweets da rede social ficou mais rápida e prática pelo fato de todos os *inputs* estarem claramente indicados na interface, além da janela "simulando"um *prompt* de comando funcionar, sem congelar a aplicação durante a execução dos *scripts*. Assim a aplicação pode ser considerada interessante e certamente ajudará principalmente pesquisadores não programadores na extração de tweets.

Em relação à trabalhos futuros cogita-se na criação de outros *scripts* de tratamento de dados, principalmente com as funções de corte de *datasets* e filtro por língua a fim de de remover parte de tweets ou tweets inteiros indesejados ou de dividir a coleta em vários *datasets*. Junto aos novos *scripts* planeja-se de ser possível realizar o *import* dos dados para a extensão CSV (*Comma-separated values*) com o objetivo de ser possível manipular os dados em outras ferramentas já utilizadas por muitos pesquisadores que não sabem programar, como, por exemplo, o Excel.

Junto a novos *scripts* pensa-se a inserção de uma visualização semelhante à um mosaico com o objetivo parecido com a visualização dos tweets mais retweetados sendo possível resumir a opinião e as discussões dos usuários, porém, focando em fotos e vídeos a fim de complementar a análise. Com o intuito de implementar essa visualização, através do módulo *rest gathering* a coleta já inclui as URLs das fotos e vídeos dos tweets coletados. Por fim, pretendemos fazer testes com usuários para validar e aperfeiçoar a interface criada.

Referências

[1] InternetLiveStats, "Twitter Usage Statistics," Acessado em 28 de março de 2022 https://www.internetlivestats.com/twitter-statistics.

[2] Chatterjee, Sheshadri; Chaudhuri, Ranjan; Vrontis, Demetris, “Does data-driven culture impact innovation and performance of a firm? An empirical examination,” in Annals of Operations Research. Springer, 2021, pp. 1-26

[3] Stojanovski, Dario; Dimitrovski, Ivica; Madjarov, Gjorgji, “Tweetviz: Twitter data visualization,” in Proceedings of the data mining and data warehouses, 2014.

[4] S. Malik et al., “TopicFlow: Visualizing topic alignment of Twitter data over time,” 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 2013, pp. 720-726, doi: 10.1145/2492517.2492639.

[5] Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. J. “Sentiment analysis of twitter data,” in Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38). 2011, June.

[6] Boanjak, Matko; Oliveira, Eduardo; Martins, José; Mendes Rodrigues, Eduarda;Sarmento, Luis. “TwitterEcho - A distributed focused crawler to support open research with twitter data”, 2012, doi: 10.1145/2187980.2188266.

[7] Florence Ying Wang; A. Sallaberry; K. Klein; M. Takatsuka; M. Roche, “SentiCompass: Interactive visualization for exploring and comparing the sentiments of time-varying twitter data,” in IEEE Pacific Visualization Symposium (PacificVis), 2015, pp. 129-133, doi: 10.1109/PACIFICVIS.2015.7156368.

[8] Rotta, G.C.; de Lemos, V.S.; da Cunha, A.L.M.; Manssour, I.H.; Silveira, M.S.; Pase, A.F. “Exploring Twitter Interactions through Visualization Techniques: Users Impressions and New Possibilities,” in Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds) Human-Computer Interaction – INTERACT 2013.

[9] Kishore, Shohil; Peko, Gabrielle M; Sundaram, David, “Looking Through the Twitter Glass: Bridging the Data – Researcher Gap,” in The University of Auckland Business School Research Paper Series, 2019, Twenty-fifth Americas Conference on Information Systems (AMCIS), Cancun, Mexico, Aug 15-17, AIS Electronic Library (AISeL), Available at SSRN: https://ssrn.com/abstract=3966006

[10] Sanvido, Pedro Henrique M; Kurtz, Gabriela B; Teixeira, Carlos RG; Wagner, Pedro P; Leuck, Lorenzo P; Silveira, Milene S; Tietzmann, Roberto; Manssour, Isabel H, “PeakVis: a Visual Analysis Tool for Social Network Data and Video Broadcasts,” in IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), pp 418-427, 2021.

[11] Machado, Caio, et al. “Consumo de notícias e informações políticas no Brasil: Mapeamento do primeiro turno das eleições presidenciais brasileiras de 2018 no Twitter.” in Memorando de dados Comprop 4, 2018.

[12] I. Basaille; S. Kirgizov; É. Leclercq; M. Savonnet; N. Cullot; “Towards a Twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets,” in IEEE Tenth International Conference on Research Challenges, Information Science (RCIS), 2016, pp 1-10, doi: 10.1109/RCIS.2016.7549324.

[13] J. You; J. Lee; H. -Y. Kwon; “A Complete and Fast Scraping Method for Collecting Tweets,” in IEEE International Conference on Big Data and Smart Computing (BigComp), 2021, pp. 24-27, doi: 10.1109/BigComp51126.2021.00014.

[14] A. Sharma; R. Rana; “Analysis and Visualization of Twitter Data using R,”2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2020, pp. 455-459, doi: 10.1109/PDGC50313.2020.9315740.