UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JONATHAN CARLETTI SILVA

Analysis of Twitter Users and Posts Based on Hashtags Regarding COVID-19 Vaccines

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Dante Augusto Couto Barone
Coadvisor: Ms. Eduardo Gabriel Cortes

Porto Alegre
April 2023

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos André Bulhões
Vice-Reitora: Prof^{a}. Patricia Pranke
Pró-Reitora de Graduação: Prof^{a}. Cíntia Inês Boll
Diretora do Instituto de Informática: Prof^{a}. Carla Maria Dal Sasso Freitas
:
Coordenador do Curso de Ciência de Computação: Prof. Marcelo Walter
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro
:

ACKNOWLEDGMENT

I thank my beloved wife, Thaísa Tamusiunas, who always encouraged and supported me throughout this journey. This work is dedicated to her. I also thank my Advisor and Coadvisor for the opportunity and guidance to make this work happen.

ABSTRACT

With advance of COVID-19, researchers around the world have developed vaccines in record time to try to mitigate the pandemic situation. When the vaccines were about to arrive in Brazil, different opinions (pro-vaccine and anti-vaccine) have raised on internet, probably motivated by the fastness of immunizing was made, by the political polarization Brazil was facing and perhaps by the fake news spreading. The main objective of this work is to make an analysis of posts collected from Twitter containing hashtags pro-vaccine and anti-vaccine, during COVID-19 pandemic period, studying the possible factors that driven the usage of those hashtags. A pre-processing was made to filter only relevant fields from a Twitter dataset and a support software was built to help on analysis, plotting graphs and calculating numbers of tweets with hashtags in favor or not in favor to vaccines. In this work we analysed 43935 users and 89851 tweets and we could see that 93% of users have used pro-vaccine hashtags, 6% used anti-vaccine and only 1% have changed their hashtags overtime.

Keywords: Vaccination. data processing. data mining. twitter. hashtags. covid.

RESUMO

Com o avanço da COVID-19, pesquisadores ao redor do mundo desenvolveram vacinas em tempo recorde para tentar mitigar a situação de pandemia. Quando as vacinas estavam para chegar ao Brasil, diferentes opiniões (pró-vacina e antivacina) surgiram na internet, provavelmente motivadas pela rapidez com que os imunizantes foram feitos, pela polarização política que o Brasil enfrentava e possivelmente também pela disseminação de fake news. O principal objetivo deste trabalho é fazer uma análise de postagens coletadas do Twitter contendo hashtags pró-vacina e antivacina, durante o período de pandemia da COVID-19, estudando os possíveis fatores que motivaram o uso dessas hashtags. Um pré-processamento foi feito para filtrar apenas os campos relevantes de um conjunto de dados do Twitter e um software de apoio foi construído para ajudar na análise, plotando gráficos e calculando números de tweets com hashtags a favor ou contra a vacinação. Neste trabalho analisamos 43.935 usuários e 89.851 tweets e pudemos constatar que 93% dos usuários usaram hashtags pró-vacina, 6% usaram hahstags antivacina e apenas 1% mudou suas hashtags ao longo do tempo.

Palavras-chave: Vaccination. data processing. data mining. twitter. hashtags. covid. mineração de dados. vacinação. processamento de dados.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

With advance of COVID-19, researchers around the world have developed vaccines in record time to try to mitigate the pandemic situation. When the vaccines were about to arrive in Brazil, different opinions (pro-vaccine and anti-vaccine) have raised on internet, probably motivated by the fastness of immunizing was made, by the political polarization Brazil was facing and perhaps by the fake news spreading.

Facing this scenario, it is natural that some questions are raised about what instigates people to express themselves in favor or not in favor of COVID-19 vaccination on social medias, like Twitter. This questioning is also the main motivation for this work. It is that possible to relate the computational data analysis with posts from Twitter (tweets) to try to understand what are the factors contributing to people express their sentiment related to vaccination?

Motivated by this, the main objective of this work is to make an analysis of tweets containing hashtags pro-vaccine and anti-vaccine, during COVID-19 pandemic period - from January 2019 to September 2021 - studying the possible factors that driven the usage of those hashtags. The specifics objectives of this work are: simplifying the tweets dataset extracted; creating graphs based on pro-vaccine and anti-vaccine hashtags; analysing graph spikes over the proposed period; and associating the results with external events.

The methodology consists on pre-processing files with dataset extracted from Twitter and build a support software to generate graphs. The result shows the usage of pro-vaccine and anti-vaccine hashtags of all users tweets from dataset, including tweets from verified users (Influencers), which may be shown in a separated graph. The software also build graphs showing the hashtags changes (anti-vaccine to pro-vaccine and vice-versa) over time.

This work may increase the users comprehension facing the spread of fake news, not only on political and Twitter context, but to understand in more general contexts and to another social media platforms. Furthermore, the pre-processing script could be used to simplify future dataset extraction, and the software support can automatically executes new analysis based on future uploaded data.

To achieve those objectives, this work is organized with the following structure: chapter 2 presents the background related to this work, contextualizing the pandemic context in Brazil; an explanation of dataset origin, the tools used to achieve this work and also the related work. In chapter 3, we have the methodology showing how is was the implementation and steps to achieve the objectives. Chapter 4 presents the results and analysis obtained with the implementations. Finally, we have the conclusion in chapter 5, showing the results and possible future works.

2 BACKGROUND

2.1 Vaccination on Pandemic Situation in Brazil

During the pandemic, Brazil was among the countries most affected by COVID-19, affecting the sectors of economy, education [1] and also causing the growth of unemployment, that have passed 15 millions on the first half of 2021 [2]. Also, the political Brazilian situation was not good in pandemic time, the public opinion was divided between government supporters and the opposition. The government and its supporters used to minimize the effects of COVID-19 and, also, the president at that time used to discouraging the use of vaccines against the virus [3]. Beyond that, a lot of fake news regarding the vaccines and COVID-19 was spread [4] that contributed to have more Brazilians with anti-vaccine opinion. On the other hand, a lot of Brazilians disagree with the government position regarding vaccines and COVID-19, causing a great struggle beyond the pandemic situation.

In urgency of having a solution to pandemic, researchers around the world have developed COVID-19 vaccines in record time [5]. With the polarization of Brazilians and the spreading fake news as never seen before, is natural that social medias has reflected this different opinions and Twitter is one of the most used platforms that we could use as a tool to analyse those different sentments regarding vaccination.

2.2 Dataset

This work have used files with dataset with tweets extracted from Twitter in the period from January 2019 to September 2021. (HALLBERG; CORTES; BARONE, 2021) and (MARTINS, 2022) have extracted those data into 33 files of 45GB, containing several information about each tweet from that period. The tweets were extracted in two steps: first obtaining ID from Tweets that express ideas pro-vaccine and anti-vaccine, with an

[1] https://www.worldbank.org/pt/country/brazil/brief/impactos-da-covid19-no-brasil-evidencias-sobre-pessoas-com-deficiencia-durante-a-pandemia

[2] https://veja.abril.com.br/economia/ibge-desemprego-durante-a-pandemia-foi-maior-que-o-estimado/

[3] ttps://www.bbc.com/portuguese/brasil-55939354

[4] https://www.cartacapital.com.br/sociedade/como-o-brasil-foi-arrebatado-por-uma-epidemia-de-fake-news-e-desinformacao-durante-a-pandemia/

[5] https://www.mckinsey.com/featured-insights/sustainable-inclusive-growth/chart-of-the-day/mind-over-matter-how-the-world-developed-covid-19-vaccines-in-record-time

open source module from Python known as Twint [6] and, then, using an open source tool called Twarc [7] to take several information regarding those tweets IDs.

2.3 Implementation Background

All scripts in this work were made using Python language with Pandas [8] and Matplotlib [9] libraries.

2.3.1 Pandas

Pandas is an open source tool for data manipulation and analysis for Python. It facilitates the handle of large dataset converting the data into Dataframe, with elements with easy access and optimized performance. It's efficient reading json and csv large files, using Chunks, that split the reading of data into a bunch of small pieces avoiding computer memory overload. It offers also a series of structures to convert datas into organized table with easy an simple access.

2.3.2 Matplotlib

According to its website [10], *"Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible."* . In this work the Matplotlib was used to generate charts and facilitate the analysis of tweets.

2.4 Related Work

Twitter hashtags are a very good tool to measure the positioning regarding pro-vaccine and anti-vaccine opinions, since users tend to insert specific hashtags in their

[6] https://github.com/twintproject/

[7] https://github.com/docnow/twarc

[8] https://pandas.pydata.org/

[9] https://www.worldbank.org/pt/country/brazil/brief/impactos-da-covid19-no-brasil-evidencias-sobre-pessoas-com-deficiencia-durante-a-pandemia

[10] https://www.worldbank.org/pt/country/brazil/brief/impactos-da-covid19-no-brasil-evidencias-sobre-pessoas-com-deficiencia-durante-a-pandemia

tweets to associate to their thoughts. In this work, we used two predefined sets of Twitter hashtags to make the analysis of pro-vaccine and anti-vaccine tweets. Those hashtags were categorized in (HALLBERG; CORTES; BARONE, 2021) work, generating two sets: one composed by hashtags commonly used by users that expressed themselves in favor to vaccines and another set composed by hashtags used by users that expressed themselves not in favor to vaccines. Two methods were used in his work to categorize those hashtags: one consists in use of search engines to find papers and news that cited hashtags used in association with opinions pro-vaccine and anti-vaccine on Twitter; and another method is the use of Twitter search to find relevant keywords and find appropriated hashtags. The hashtags set result is presented in table 2.1.

Table 2.1: Twitter hashtags categorization.

Pro-vaccine Hashtags	**Anti-vaccine Hashtags**
#VacinaParaTodosJa	#VacinaNao
#VacinaParaTodos	#EuNaoVouTomarVacina
#VacinaSim	#VacinaMata
#VacinaJa	#NaoVouTomarVacina
#VemVacina	#NaoÀsVacinas
#VacinaSalva	#ChegaDeVacina
#TodosPelasVacinas	#VacinasMatam
#VacinaFunciona	#NaoQueroVacina
#VacinaÉAmorAoPróximo	#NaoVouTomar
#VacinaAgora	#ContraVacina
#QueroSerVacinado	#VacinasCausamAutismo
#QueroSerVacinada	#NaoAVacina
#ExijoVacina	#NaoTomoVacina
#VivaAVacina	#ForaVacina
#QueroVacina	
#VacinasFuncionam	
#ProVacina	
#VacinasPelaVida	
#VacinasSalvamVidas	
#Vacinese	
#EuQueroVacina	
#EuQueroVacina	
#VacinasSalvam	
#VacinasJa	
#VacinasFuncionam	

In (HALLBERG; CORTES; BARONE, 2021) work, the focus was on determining and understanding the factors that lead Brazilian users on Twitter to be pro-vaccines or anti-vaccines and how demographic factors can contribute to influence users opinion. As a conclusion, it was determined that political keywords used in Twitter profiles is the factor with the most significant influence on users opinions. Also, age and localization are strong factors of influence.

In (MARTINS, 2022) work it was used Machine Learning techniques on large set of data to try to predict the opinion of users about vaccination. And, then, use sentiment analysis techniques in tweets about vaccines to see if they have a very positive, positive, neutral, negative or very negative sentiment. It was conclude that the most expressed sentiment by pro-vaccine users is the positive, while sentiment by anti-vaccine users is neutral.

Another related works are (RECUERO; ZAGO; BASTOS, 2014), which used computational techniques to understand the main features of speeches from Brazilian protests in June 2013, (DARWISH et al., 2020), where it was presented a highly effective unsupervised framework for detecting the stance of prolific Twitter users with respect to controversial topics, (BECHINI et al., 2020), which discusses a crucial aspect in structuring the data processing pipeline in intelligent systems aimed at monitoring the public opinion through Twitter messages, and (GOMIDE et al., 2011), which analyzes how Dengue epidemic is reflected on Twitter and to what extent that information can be used for the sake of surveillance.

The results of our work shows that fake news are one of the main causes that leads people to use more hashtags anti-vaccine (that will be explained in Chapter 4) and the work (HAYAWI et al., 2022) explores vaccine tweets with misinformation about COVID-19 vaccines. It was collected and annotated COVID-19 vaccine tweets and trained machine learning algorithms to classify vaccine misinformation. More than 15,000 tweets were annotated as misinformation or general vaccine tweets using reliable sources and validated by medical experts and as a result the best classification performance was obtained using BERT, showing that Machine learning–based models are effective in detecting misinformation regarding COVID-19 vaccines on social media platforms.

A significant time was invested on pre-processing the Twitter dataset in our work to make feasible all analysis. Matching with the idea that pre-processing is important, in work (ALAM; YAO, 2019) it was studied the impact of different pre-processing steps on the accuracy of three machine learning algorithms for sentiment analysis, applying different text pre-processing techniques and studying their impact on accuracy for sentiment classification using three well-known machine learning classifiers including Naïve Bayes (NB), maximum entropy (MaxE) and support vector machines (SVM). It was calculated accuracy of the three machine learning algorithms before and after applying the pre-processing steps. Results proved that the accuracy of NB algorithm was significantly improved after applying the pre-processing steps. Slight improvement in accuracy of SVM algorithm was seen after applying the pre-processing steps. This research work proves that text pre-processing impacts the accuracy of machine learning algorithms.

Also, in (NASEEM; RAZZAK; EKLUND, 2021) it was shown that pre-processing plays an essential role in disambiguating the meaning of short-texts, not only in applications that classify short-texts but also for clustering and anomaly detection. This paper analyzes twelve different pre-processing techniques on three pre-classified Twitter datasets on hate speech and observes their impact on the classification tasks they support. It also proposes a systematic approach to text pre-processing to apply different pre-processing techniques in order to retain features without information loss. Two different word-level feature extraction models are used and the performance of the proposed package is compared with state-of-the-art methods. The experimental results suggest that some pre-processing techniques impact negatively on performance, and these are identified, along with the best performing combination of pre-processing techniques.

3 METHODOLOGY

This chapter presents the methodology of dataset analysis used in this work and all information related to implement it.

3.1 Implementation

The implementation planning for this work is to take the dataset already extracted by (HALLBERG; CORTES; BARONE, 2021) and (MARTINS, 2022), as mentioned before, pre-processing the data to let it more feasible to work, with only relevant information and make an analysis based on the hashtags classified on (HALLBERG; CORTES; BARONE, 2021) work, creating a support program that take the pre-processing data and give information about COVID-19 vaccine positioning (pro-vaccine or anti-vaccine) of all tweets, also analysis of verified Twitter accounts; making an analysis of users that have changed their opinions about vaccines over time to and, finally, help on understanding what motivated users to change their minds and what are the facts that driven their positioning.

3.1.1 Pre-processing

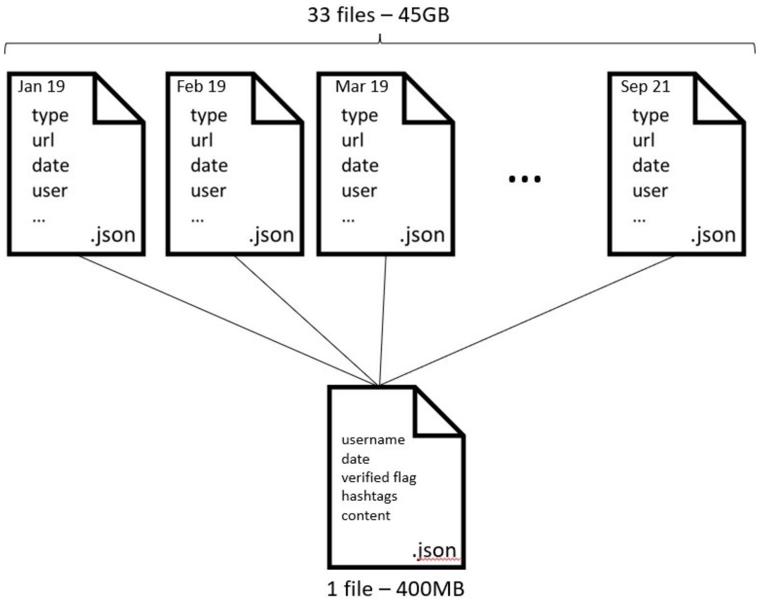
In this work it was used a dataset extracted by (HALLBERG; CORTES; BARONE, 2021) and (MARTINS, 2022), as mentioned before, first obtaining ID from Tweets that express ideas pro-vaccine and anti-vaccine, with an open source module from Python known as Twint [1] and, then, using an open source tool called Twarc [2] to take several information regarding those tweets IDs. The dataset was divided into 33 files extracted from Twitter with tweets from January 2019 to September 2021. In those files it has the information of each tweet date, user, id, date, content, hashtags and several other information about the posts. The files were around 45GB in total which made unfeasible to use the files as it is. First we have difficult to run the functions of the support program (explained later) due to out of memory issues. This problem was solved using a Pandas feature called *chunks*, that is a parameter to read large json or csv files. With *chunks*, a value is passed in *chunksize* parameter when read json file to determining how many lines Pandas will

[1] https://github.com/twintproject/

[2] https://github.com/docnow/twarc

read from a file. With this feature, the problem of memory was solved but another issue arose: the program was taking around 36 hours to run through tweets, identifying users with hashtags pro-vaccine and anti-vaccine and, then, plotting a chart with the result. A lot of optimizations was made and the 36 hours turned into 24 hours, but this time was not good enough to make analysis feasible. The final solution to make the functions faster was making a filter only with necessary fields from Twitter dataset. For this work we only need some few elements from those tweets information, so it was necessary to have a pre-processing of the the dataset before starts to implement itself. The script implemented went through all 33 files, searching in each tweet the information about username, tweet date, a flag about account verified, the hashtags and content; and only tweets containing hashtags (not empty). This procedure of making a filter ran in about 24 hours, then all those information were stored in a Pandas Dataframe and saved as one single json file, but it was just one shot. The previous 33 files (one per month) summing 45GB was transformed into one file with around 400 MB. With this pre-processing, the support software is running in an average of 5 minutes each function, a gain on 99.7% of performance, making all analysis feasible for this work.

Figure 3.1: Scheme on pre-processing implementation.

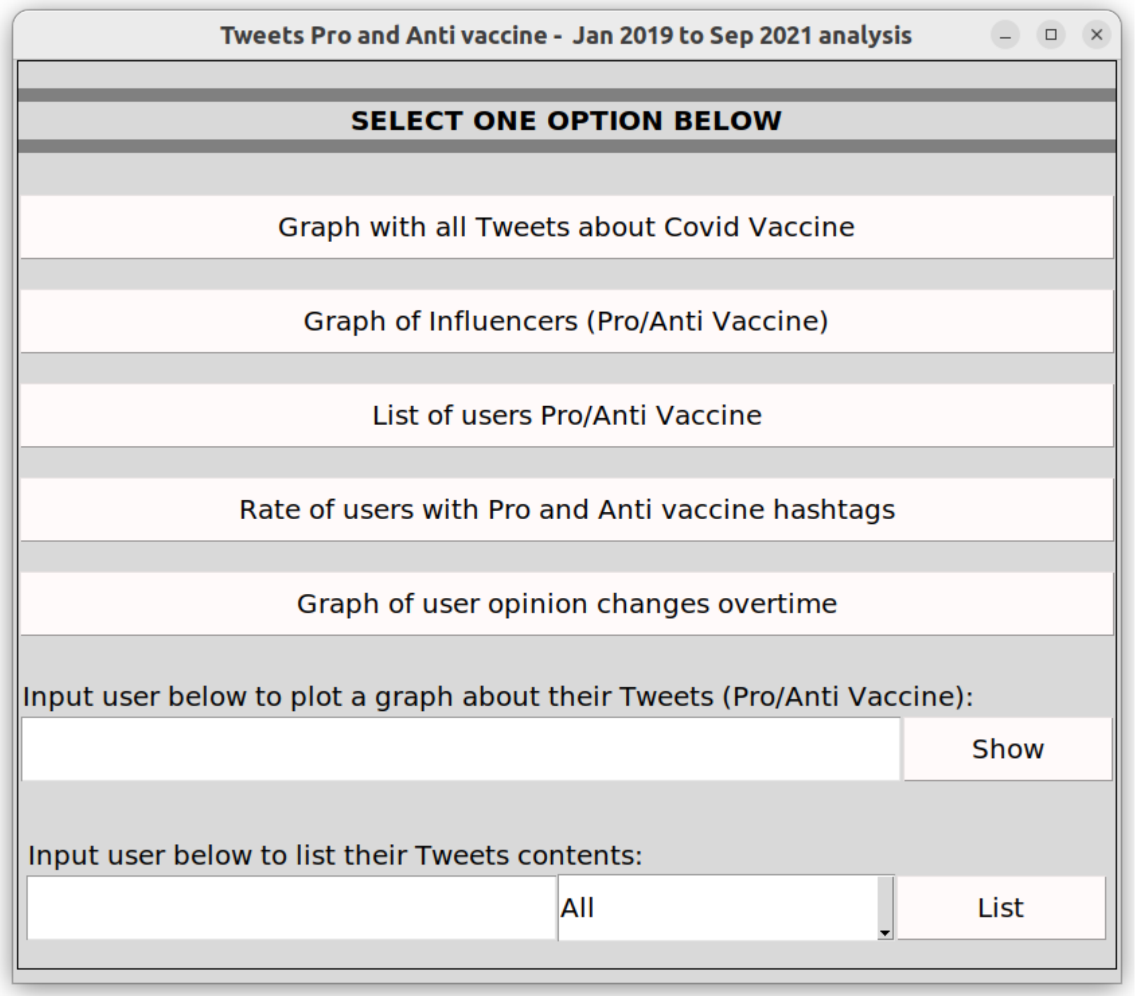


Source: The Authors

3.1.2 Support Software

A software was built to provide support on analysis of tweets, it consists in 7 main procedures that takes the dataset pre-processed, plotting charts and making analysis about tweets, helping on getting automatically overall and also specific results about users, tweets and hashtags changes. The scripts also plot graphs showing the quantity of hashtags pro-vaccine and anti-vaccine and the changes of the user positioning over time.

Figure 3.2: Interface of Support Software developed to help on analysis for this work.


Tweets Pro and Anti vaccine - Jan 2019 to Sep 2021 analysis
SELECT ONE OPTION BELOW
Graph with all Tweets about Covid Vaccine
Graph of Influencers (Pro/Anti Vaccine)
List of users Pro/Anti Vaccine
Rate of users with Pro and Anti vaccine hashtags
Graph of user opinion changes overtime
Input user below to plot a graph about their Tweets (Pro/Anti Vaccine):
Show
Input user below to list their Tweets contents:
All
List

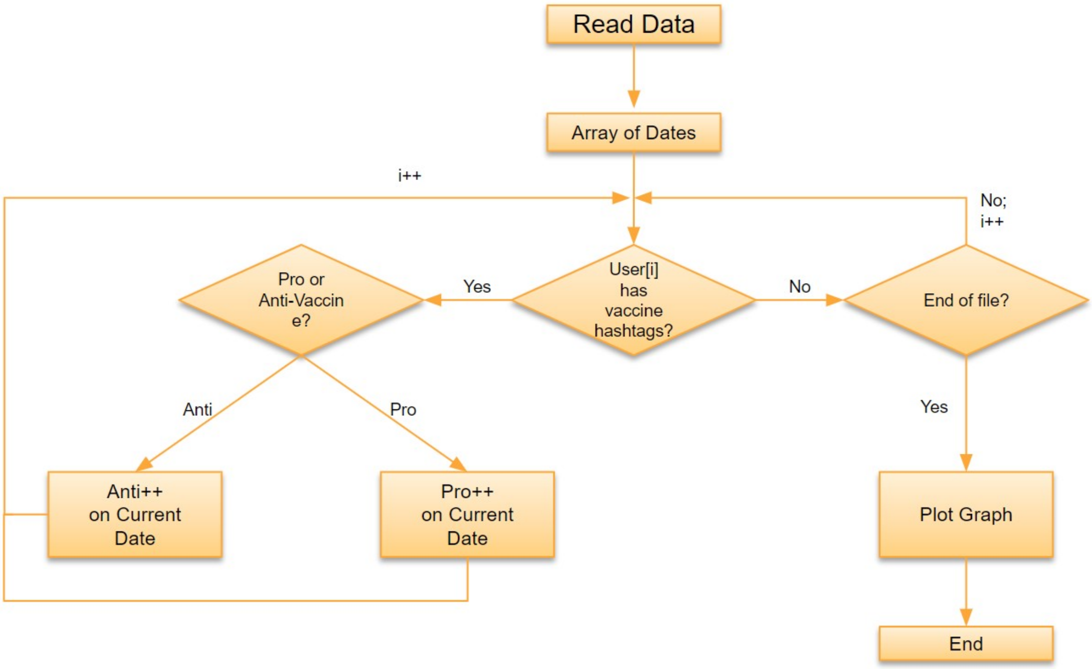
Source: The Authors

Before the start of interface, the program take the pre-processed file and transform it into Pandas Dataframe to facilitate the manipulation.

3.1.3 Line Graphs

There are three procedures that plot Line Graphs: one plots using all Tweets about COVID-19 vaccine, another plots using tweets from verified accounts (Influencers), and the last one plots a line graph from a determined user. The functions run through tweets and check if the tweets are pro-vaccine or anti-vaccine - according hashtags set present in one of the groups of hashtags pro-vaccine or anti-vaccine from (HALLBERG; CORTES; BARONE, 2021) work. Finally, it plots a graph with Axis *x* showing the months and Axis *y* showing the number of tweets pro-vaccine and anti-vaccine. The function of Influencers checks if the profile has a flag *"account verified"* also.

Figure 3.3: Fluxogram of basic scheme from procedures that plot graphs in line.



Source: The Authors

3.1.4 Users with Pro-vaccine and Anti-vaccine hashtags and list of tweet contents

There are two functions listing information about all users: one showing with pro-vaccine, anti-vaccine and both pro/anti-vaccine hashtags (selected by whom is using the Software), and another lists the content of tweets of a determined user - here it can be chosen all tweets or only tweets with vaccine hashtags. The first function saves 3 files, one with unique users only with pro-vaccine hashtags, another with only anti-vaccine hashtags and another file with unique users having both hashtags. The file contains the username and how many tweets with pro-vaccine or anti-vaccine hashtags. It uses also the pro-vaccine and anti-vaccines hashtags mapped before to determine which ones have one or another hashtags (or both).

3.1.5 Rate of Users with Pro-vaccine and Anti-vaccine hashtags

This procedure calculates the rate of users that have changed their hashtags related to total users, showing the number of users that changed from pro-vaccine to anti-vaccine hashtags and vice-versa.

3.2 Expantion to Other Subjects

The support software make analysis between two sets of hashtags. So the hashtags can be easily replaced by another hashtags from a different subjects in the code. The new hashtags should be inserted without the simbol "#", then the program will search for those different sets of hashtags and will plot a new graph using those values. Also, the pre-processing script can be used on new files with updated data from Twitter, as long as dataset is taking with Twint and Twarc methods (json file) as mentioned before. The Python codes are available on Github [3].

[3] https://github.com/jonafui/AnalysisOfTwitterHashtags

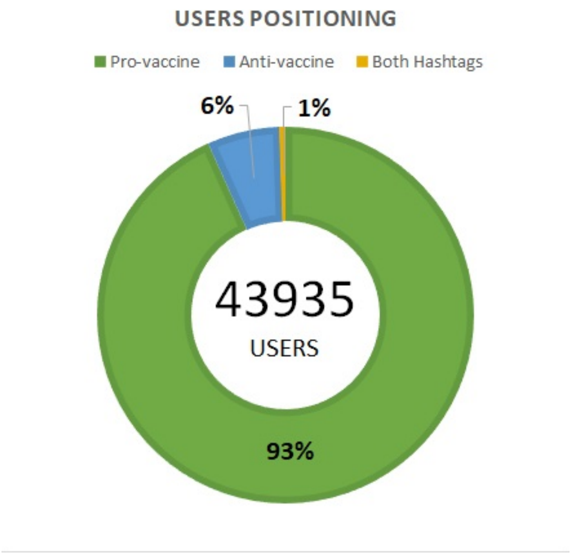
4 RESULTS AND ANALYSIS

This chapter presents the results and analysis of users and number of tweets with pro-vaccine and anti-vaccine hashtags between January 2019 and September 2021.

4.1 Dataset

A total of 43935 users were identified having pro-vaccine and anti-vaccine hashtags on their tweets (according to (HALLBERG; CORTES; BARONE, 2021) hashtags), between January 2019 and September 2021. 40884 users have used pro-vaccine hashtags (around 93% of total); 2716 users have used anti-vaccine hashtags (around 6% of total); and 235 users have used both hashtags - pro-vaccine and anti-vaccine - over time (less than 1%).

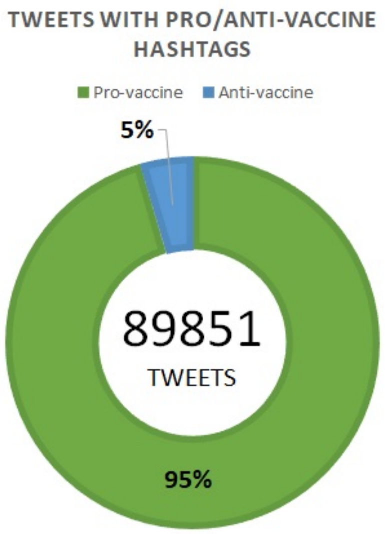
Figure 4.1: Total of users identified with anti-vaccine, pro-vaccine and both hashtags between January 2019 and September 2021.



Source: The Authors

Regarding number of tweets, a total of 89851 posts were identified with pro-vaccine and anti-vaccine hashtags, between January 2019 and September 2021. 85797 tweets with pro-vaccine hashtags (around 95% of total) and 4054 tweets with anti-vaccine hashtags (around 5% of total).

Figure 4.2: Total of tweets identified with pro-vaccine and anti-vaccine hashtags between January 2019 and September 2021.



Source: The Authors

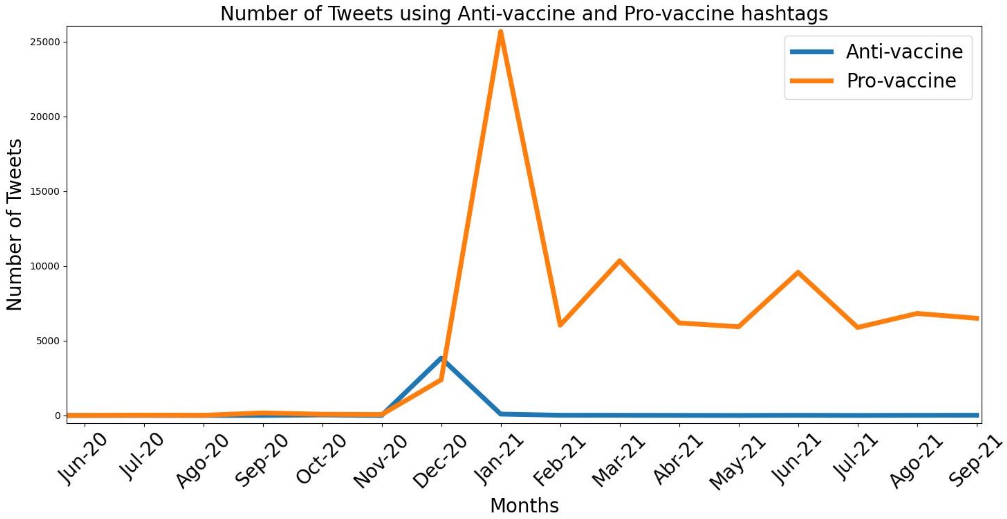
4.2 Dataset Analysis

We could identified the periods when tweets have been posted and put in charts (Figure 4.3 and 4.4) where Axis *x* represents when the tweets has been posted (per month) and the Axis *y* means the number of tweets in each month. The period analysed is January 2019 to September 2021, as mentioned before, but from January 2019 to May 2020 it wasn't enough significant data to show in charts, so it was decided to plot numbers starting in June 2020.

4.2.1 Analysis Based on Total Tweets

In Figure 4.3, we can see a chart with the most part of 89851 tweets (from June 2020 to September 2021, disregarding a few tweets before those dates, as mentioned earlier) and its corresponding positioning hashtags (pro-vaccine and anti-vaccine).

Figure 4.3: Number of Tweets with Pro and Anti-Vaccine over time.



Source: The Authors

We can see an anti-vaccine spike on December 2020. In this month, Brazilian Department of Health was confirmed the first case of COVID-19 re-infection in Brazil. Also, several countries around the world have started vaccination programs. In Brazil, the government announced the National Vaccination plan against COVID-19. We can observe also that in December the Supreme Court of Brazil judged the constitutionality of compulsory vaccination. In the end of December, it was decided that Brazilian States governments would have autonomy to decide whether or not vaccines would be mandatory in their States. With the imminent arrival of COVID-19 vaccines, several fake news were spread in Brazil in December 2020: fake news talking about crematory ovens sent from China to Argentina due to the adoption of CoronaVac in the country [1]; another saying that Pharmaceuticals Pfizer and BioNTech vaccine for Covid-19 causes infertility in women [2]; also images showing wounds caused by the Pfizer and BioNTech vaccine on the feet of a volunteer in the US [3]; another message saying that Peru suspended tests with CoronaVac due to neurological problems in a volunteer [4]; other saying that nurse died in Tennessee after taking Covid-19 vaccine and collapsing in public [5]; saiyng that Pfizer

[1] https://g1.globo.com/fato-ou-fake/coronavirus/noticia/2020/12/03/e-fake-que-imagem-mostre-fornos-crematorios-enviados-da-china-para-a-argentina-por-conta-da-adocao-da-coronavac-no-pais.ghtml

[2] https://g1.globo.com/fato-ou-fake/coronavirus/noticia/2020/12/08/e-fake-que-vacina-das-farmaceuticas-pfizer-e-biontech-para-covid-19-cause-infertilidade-em-mulheres.ghtml

[3] https://g1.globo.com/fato-ou-fake/coronavirus/noticia/2020/12/08/e-fake-que-imagem-mostre-ferimentos-causados-pela-vacina-da-pfizer-e-da-biontech-em-pes-de-voluntaria-nos-eua.ghtml

[4] https://g1.globo.com/fato-ou-fake/coronavirus/noticia/2020/12/16/e-fake-que-peru-suspendeu-testes-com-coronavac-por-problemas-neurologicos-em-um-voluntario.ghtml

[5] https://g1.globo.com/fato-ou-fake/coronavirus/noticia/2020/12/20/e-fake-que-enfermeira-morreu-no-

CEO said he won't take his own vaccine [6], and so on. All those news are fake and were spread on December 2020. We can see that those factors contributed to users post more tweets against vaccination on December 2020, it indicates that probably users were influenced by several information coming from government mainly raising the debate about if it is reasonable the government force a person to become vaccinate even against their will. It's interesting to observe also that this was the only month that we had more anti-vaccine tweets than pro-vaccine ones, indicating that probably the spread of fake news was a factor that contributed to this anti-vaccine spike.

On January 2021, we had the biggest spike of tweets with pro-vaccine hashtags. In this month Brazil has a growth in COVID-19 cases and a new variant was discovered (Gamma). The Brazilian Health Regulatory Agency (Anvisa) approved the use of Coronavac and Oxford/Astrazeneca vaccines on emergency basis, and in the end of January, vaccination has begun in Brazil. Those were possibly the main factors that generated the huge spike on this month.

We can also observe another spikes on March 2021, possibly due to the first mass vaccination in Brazil; and on June 2021 when Anvisa authorized the emergency importation of vaccines from Russia to North and Northest of Brazil.

4.2.2 Analysis Based of Influencers

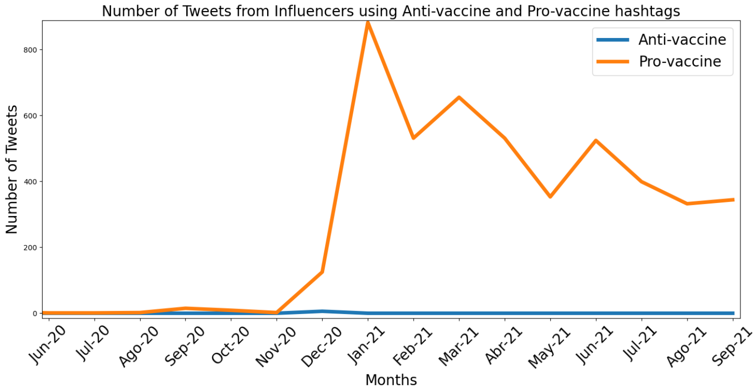
In Figure 4.4, we have the chart of 4797 tweets from users with verified accounts, that we considers as Twitter Influencers. According to Twitter, a verified account is any account of public interest that's been authenticated by the company itself [7].

tennessee-apos-tomar-vacina-contra-covid-19-e-desmaiar-em-publico.ghtml

[6] https://g1.globo.com/fato-ou-fake/coronavirus/noticia/2020/12/29/e-fake-que-ceo-da-pfizer-disse-que-nao-vai-tomar-a-propria-vacina.ghtml

[7] https://www.theverge.com/23199155/verified-twitter-account-how-to

Figure 4.4: Number of Tweets from Influencers with Pro and Anti-Vaccine over time.



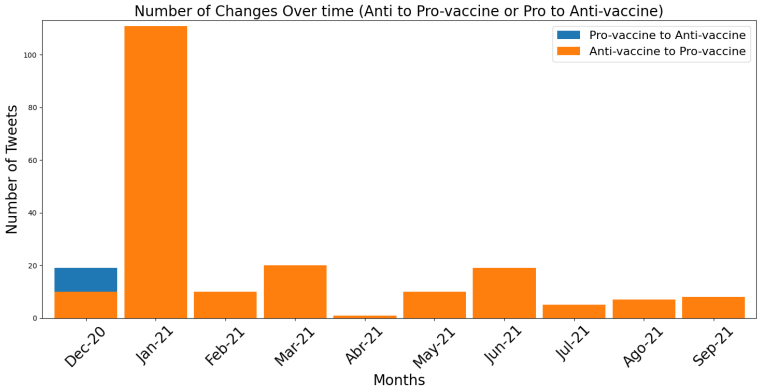
Source: The Authors

We can see that the vast majority of tweets have pro-vaccine hashtags. It was identified only 6 tweets with anti-vaccine hashtags. The spikes on chart of Figure 4.4 occur in the same months as showing on chart of Figure 4.3, indicating that verified accounts could have influenced the tweets from general users.

4.3 Analysis of Hashtags Changes

From those 235 users that used both hashtags, as mentioned earlier, almost 8% changed their hashtags from pro-vaccine to anti-vaccine; around 85% changed their hashtags from anti-vaccine to pro-vaccine; and 7% of users have changed their hashtags more than once or have used both hashtags in same Tweet.

Figure 4.5: Number of Tweets with Pro and Anti-Vaccine over time.



Source: The Authors

Figure 4.5 shows the users that changed their hashtags positioning over time, from anti-vaccine to pro-vaccine or the opposite. We can observe that the months with most changes coincide with the periods with spikes from Figures 4.3 and 4.4. December 2020 is the only month that has changes from pro-vaccine to anti-vaccine, probably driven by the several fake news spread, as mentioned before. From January 2021 to September 2021, we have only occurrences from anti-vaccine to pro-vaccine. The spike on January 2021 is linked to earlier explanation about chart on Figure 4.3: the growth of COVID-19 cases and the approve vaccination from Anvisa (see section 4.2.1).

4.4 Limitations

The results and analysis of this work was based on a fixed database extracted from Twitter using the tool Twint [8] and API Twarc [9], as mentioned before, and the dataset files was previously generate by (HALLBERG; CORTES; BARONE, 2021) and (MARTINS, 2022) with all tweets from January 2019 to September 2021. The algorithm will not take dynamically the Tweets from Twitter database. For updates it will be necessary to use Twint and Twarc again to take necessary new Tweets. All analysis were based on hashtags previously classified on (HALLBERG; CORTES; BARONE, 2021) work, so it doesn't take into account that some users can use hashtags as an irony, for example.

4.5 Results

As a result of the analysis of this work, we can see that the majority of tweets about COVID-19 vaccination have pro-vaccine hashtags which is an evidence that most of users are in favor of vaccination and, also, with the low rate of changing hashtags (anti-vaccine to pro-vaccine and vice-versa) it indicates that most of the users tend to not change their opinion. On the other hand, the minority of users that seems to be not in favor of vaccination are more inclined to change their minds over time. Also, the vast majority of pro-vaccine tweets from verified users could be an indication of influence in general users about COVID-19 vaccination. We can also observe that fake news seems to be one of the main factors to users use more anti-vaccine hashtags and also to change their opinion from pro-vaccine to anti-vaccine (based on hashtags).

[8] https://github.com/twintproject/

[9] https://github.com/docnow/twarc

5 CONCLUSION

In this work we have taken the fixed dataset generated by (HALLBERG; CORTES; BARONE, 2021) and (MARTINS, 2022), making a pre-processing taking only necessary attributes and converting the 33 files with 45GB into one simple file of 400MB with only relevant information for this work, that is, username, tweet date, flag of verified account, hashtags and content of each tweet. Then, a support software was built that takes the pre-processed file and identify tweets and users that are pro-vaccine and anti-vaccine according to hashtags classified in (HALLBERG; CORTES; BARONE, 2021) work. The scripts also plotting charts to help on analysis of the tweets.

With those procedures we could realize that political situation and spread of fake news were main factors that influence users to use more hashtags anti-vaccine and also tend to change their hashtags from pro-vaccine to anti-vaccine, as we could see on December 2020 numbers, where we have a bunch of fake news spread and we saw more hashtags anti-vaccine than hashtags pro-vaccine, and also is the only period when we have changes of pro-vaccine to anti-vaccine hashtags on tweets.

We can see also that the majority of tweets about COVID-19 vaccination have pro-vaccine hashtags which is an evidence that most of users are in favor of vaccination and, also, the low rate of changing hashtags it is an indicative that most of the users tend to not change their opinion regarding COVID-19 vaccination.

Another point to observe is that the vast majority of pro-vaccine tweets from verified users could be an indication of influence in general users about COVID-19 vaccination.

As future work, it would be interesting of having those same analysis extend to other social media platforms like Instagram and Facebook. Another point that could be a future improve is the tentative of analysing irony on tweets, since some users use hashtags opposite of what they really mean on tweet. Another point is trying to expand the analysis to a political vision, trying to understand the political position of users and try to relate it to their opinion regarding vaccination.

REFERENCES

ALAM, S.; YAO, N. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. **Computational and Mathematical Organization Theory**, Springer, v. 25, p. 319–335, 2019.

BECHINI, A. et al. Stance analysis of twitter users: the case of the vaccination topic in italy. **IEEE Intelligent Systems**, IEEE, v. 36, n. 5, p. 131–139, 2020.

DARWISH, K. et al. Unsupervised user stance detection on twitter. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2020. v. 14, p. 141–152.

GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In: **Proceedings of the 3rd international web science conference**. [S.l.: s.n.], 2011. p. 1–8.

HALLBERG, A. G.; CORTES, E. G.; BARONE, D. A. C. An analysis of twitter users opinions on vaccines using machine learning techniques. In: **2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)**. [S.l.: s.n.], 2021. p. 1311–1315.

HAYAWI, K. et al. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. **Public health**, Elsevier, v. 203, p. 23–30, 2022.

MARTINS, G. F. Um estudo utilizando-se de análise de sentimentos e aprendizado de máquina para a classificação de tweets sobre a vacinação no brasil. In: **Trabalho de Conclusão de Curso - Universidade Fedeal do Rio Grande do Sul**. [S.l.: s.n.], 2022.

NASEEM, U.; RAZZAK, I.; EKLUND, P. W. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. **Multimedia Tools and Applications**, Springer, v. 80, p. 35239–35266, 2021.

RECUERO, R.; ZAGO, G.; BASTOS, M. T. O discurso dos# protestosbr: análise de conteúdo do twitter. **Galáxia (São Paulo)**, SciELO Brasil, v. 14, p. 199–216, 2014.