A Look into COVID-19 Vaccination Debate on Twitter

# A Look into COVID-19 Vaccination Debate on Twitter

Larissa Malagoli[1], Júlia Stancioli[1], Carlos H. G. Ferreira[1,2], Marisa Vasconcelos[3]
Ana Paula Couto da Silva[1], Jussara Almeida[1]
[1]Universidade Federal de Minas Gerais, Brazil
[2]Universidade Federal de Ouro Preto, Brazil
[3]IBM Research, Brazil
{larissagomes,juliastancioli,chgferreira}@dcc.ufmg.br,marisaav@br.ibm.com,{ana.coutosilva,jussara}@dcc.ufmg.br

## ABSTRACT

Twitter is one of the most popular social media applications used by the general public to debate a wide range of topics. It is not surprising that the platform has become an effervescent channel where people are talking about the COVID-19 pandemic. After one year of a severe pandemic, we are now giving the first steps towards its ending: the production and distribution of vaccines as well as the start of vaccination campaigns in several countries worldwide. However, the relatively quick emergence of alternative vaccines raised several concerns and doubts among the general people, leading to lively online and offline debates.

In this paper, we investigate the public perception of this topic as it unrolls in the real world, analyzing over 12 million tweets during two months corresponding to the early stages of vaccination in the world. Our investigation includes the analyses of user engagement as well as content properties, including sentiment and psycholinguistic characteristics. In broad terms, our findings offer a first look into the dynamics of the online debate around a topic – COVID-19 vaccination – at its early stages of development, evidencing how people use the online world, notably Twitter, to share their impressions and concerns about it. As a means to allow reproducibility and foster follow-up studies, we release our collected dataset for public use.

## KEYWORDS

COVID-19 vaccines, Twitter, Textual analysis, Online discussions

**ACM Reference Format:**
Larissa Malagoli[1], Júlia Stancioli[1], Carlos H. G. Ferreira[1,2], Marisa Vasconcelos[3] and Ana Paula Couto da Silva[1], Jussara Almeida[1]. 2021. A Look into COVID-19 Vaccination Debate on Twitter. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447535.3462498

## 1 INTRODUCTION

Since December 2019, the new COVID-19 pandemic has devastated the world. Health care professionals and researchers have been working around the clock to contain the spread of the virus and develop effective solutions to fight against it. However, the lack of strong knowledge about the new virus, the challenges of a new life under (sometimes severe) restrictions and uncertainties, and the need for social distancing in the physical world fostered a sharp increase in the search for information online, notably in social media applications [22]. Indeed, many platforms have launched updates to help users find reliable information, connect with others and follow real-time events more easily. Twitter is one such example[1]. With an original prediction of growth in the user base below 3%, by October 2020, the platform was expected to grow more than 8% in the year[2], reaching a record number of users during pandemic[3].

After one year of severe pandemic and, as mentioned, a lot of online activity driven by it, the focus has recently narrowed down to a much-expected topic – the COVID-19 vaccination. The world is finally seeing the emergence of a number of vaccines to stop the virus spread, and the vaccination campaign is starting in several countries. It is common sense that the success of such a campaign mostly relies on massive people's engagement, once vaccines are, so far, the unique solution to fight the spread of the virus. However, the relatively quick emergence of alternative vaccines, much faster than priorly developed vaccines, and the lack of fundamental knowledge about them raised several concerns and doubts among the general people. Such concerns have led to lively online and offline debates as well as a high level of anxiety, mistrust, and hesitancy[4], aggravated by the already established discourse of the anti-vax community, for whom no vaccine is safe or acceptable.

In this paper, we aim at taking a look into the *dynamics of the online debate around the COVID-19 vaccination through the lens of Twitter.* Twitter has been the focus of a number of studies on the modeling and analysis of online discussions [1, 13], offering a broad view of people's perceptions of various themes, from politics [1, 9] to the new coronavirus itself [16, 17]. These studies shed light on how these topics are explored in the online world, evidencing potential effects from and on the real (offline) world. We here complement these prior studies by focusing on a different, timely topic, with great impact on the global society – the COVID-19 vaccination. We aim at analyzing how the debate around this topic evolves, in particular in light of important events that took place in the real world. Our work greatly contributes to a few very recent and preliminary studies on the topic [6, 14, 23], which provided only a very coarse-grained, quantitative analysis. Instead, we here offer a broader study, covering a richer set of analyses, a larger dataset, and a longer time period.

To that end, we analyze over 12 million English-language tweets covering the months of December 2020 and January 2021, which correspond to the early stages of vaccination in several countries and during which debates on Twitter related to COVID-19 vaccination have emerged and spread. We focus our investigation on two major dimensions for analyzing information spread, namely user

[1]https://blog.twitter.com/en_us/topics/company/2020/covid-19.html

[2]https://blog.hootsuite.com/twitter-statistics/

[3]https://www.washingtonpost.com/business/economy/twitter-sees-record-number-of-users-during-pandemic-but-advertising-sales-slow/2020/04/30/747ef0fe-8ad8-11ea-9dfd-990f9dcc71fc_story.html,
https://news.sky.com/story/coronavirus-lockdowns-drive-record-growth-in-twitter-usage-12034770

[4]https://time.com/5925467/covid-19-vaccine-hesitancy/

engagement, and content properties, analyzing how their characteristics evolve over time as real-world events mark the discussions. Our main findings reveal that the volume of tweets is highly correlated with external events such as vaccine rollout campaigns and reports of vaccine efficacy studies. Examining the polarity of the discussions about vaccination we notice the negative sentiment towards topics related to the anti-vaccination movement. Finally, we observe that terms associated with health, work, and religion occur more often in tweets with vaccine-related words. In contrast, tweets mentioning anti-vax terms often contain more terms related to death, anger, and negative emotions. As a means to allow reproducibility and foster follow-up studies, we release our collected dataset for public use[5].

The rest of this paper is organized as follows. Section 2 summarizes prior related work. Section 3 describes our dataset while Section 4 presents our main results. Finally, Section 5 concludes the paper and offers possible directions for future work.

## 2 RELATED WORK

A number of prior studies have analyzed online discussions on vaccination in general and on the covid-19 pandemic [2, 3, 5–8, 10, 11, 13–18, 20, 23, 24], as we briefly review next.

**Online Discussions on Vaccination in General:** Mitra *et al.* [13] used four years of longitudinal data capturing vaccine discussions on Twitter to identify users who persistently hold pro and anti attitudes as well as those who newly adopt anti attitudes towards vaccination. Shah *et al.* [15], in turn, estimated the proportion of vaccine-related tweets linked to Web pages of low credibility and measured the potential reach of those posts. Aiming at assisting the public health communication of vaccines, the authors of [10] examined the sentiment towards the vaccination topic in social media by constructing and analyzing semantic networks of vaccine-related information from highly shared websites of Twitter users. Bonnevie *et al.* [2], in turn, reported on vaccine opposition and misinformation promoted on Twitter, highlighting Twitter accounts that drive the conversation. They identified that users who received the highest engagement on their tweets (e.g., top users) were responsible for almost 60% of vaccine-opposition messages. These top users coordinate with other vaccine opponent users to promote misinformation on the network. The authors of [20] used Twitter to monitor the public opinion propensity towards vaccination in the Italian context. They showed that the proportions of positive and negative tweets are influenced by vaccine-related events and the publication of relevant information on vaccine-preventable diseases.

**Online Discussions on COVID-19 Pandemic.** Various studies have analyzed the spread of information related to COVID-19, notably misinformation, on social media platforms. Focusing on the early periods of the pandemic, Dimitrov *et al.* [7] released TweetsCOV19, a publicly available knowledge base, with currently more than 8 million tweets, and provided a knowledge graph of COVID-19 related online discourse. Silva *et al.* [17] examined COVID-19 related tweets to characterize the spread of misinformation with respect to user engagement and bot-behavior. Similarly, Cheng et al. [24] characterized the prevalence of low-credibility information on Twitter during the coronavirus outbreak and evidence of bot coordinated activity amplifying this type of content. Gallotti *et al.* [8] analyzed a large dataset of tweets to classify the reliability of the news being disseminated. They showed that human response to falsehood exhibits early-warning signals that might be mitigated with adequate communication strategies. Shahi *et at.* [16], in turn, conducted an exploratory study on the propagation, authors, and content of misinformation present in tweets about COVID-19. Their analyses showed the presence of verified authors disseminating misinformation and discrediting other information on social media. More broadly, Cinelli *et al.* [5] addressed the diffusion of information about the COVID-19 in five social platforms, namely Twitter, Instagram, YouTube, Reddit, and Gab. By analyzing the engagement and interest in the COVID-19 topic, the authors found that information spread is driven by the interaction paradigm of each system and by specific interaction patterns of users engaged with a topic.

**Online Discussions on COVID-19 Vaccination.** We are aware of only a few studies on the dissemination of information related to COVID-19 vaccination on social media. These studies are still very preliminary and offer only a quite coarse view of the online discussions on the topic. In particular, Wu *et al.* analyzed Reddit posts related to COVID-19 vaccines, observing that the proportion of comments containing conspiracy theories outweighed that of any other topic [23]. In contrast, Pierri *et al.* [14] monitored online conversations of Italian Twitter users since the official start of the Italian vaccination campaign, finding that, despite the sharing of low-credibility information, there was a higher prevalence of high-credibility information. Finally, a dataset and a dashboard of English-language tweets about COVID-19 vaccines showing basic statistics (e.g., tweets over time, hashtags used, and websites shared) are presented in [6]. The investigation we offer here complements the aforementioned studies and build on their findings by offering a much broader set of analyses on a much larger dataset. Our results enrich the understanding of how online discussions, notably those about COVID-19 vaccines, evolve on social media during a period when the topic itself is still being developed and attracting great notoriety in the real world.

## 3 DATASET OVERVIEW

Our data collection focuses on gathering a corpus of English-language tweets that would be informative of the online debate on COVID-19 vaccines worldwide. To that end, we used the Twitter API Search[6] to collect tweets based on specific keywords related to COVID-19 vaccination. We built a list of such keywords that include terms related to both pro and anti-vaccine discourse as well as words related to the most well-known COVID-19 vaccines available so far. Specifically, we consider the following list of keywords: *vaccine, vaccination, anti-vaccination, antivax, anti-vaccine, anti-vax, anti-vaxxers, NoForcedVaccination, getvaccinated, pfizer, moderna, astrazeneca, covaxin, biontech, novavax, coronavac, sputnikv, bnt162b2.*

In total, we gathered over 12 million tweets, covering 9 weeks, from December 1st, 2020 to January 31st, 2021. This is an important

[5]https://zenodo.org/record/4721643#.YI6URbVKhPY

[6]https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

**Table 1: Overview of our dataset on a weekly basis.**

| Week | Start | #Tweets | #Retweets | #Unique Users |
|---|---|---|---|---|
| 1 | 12-01-2020 | 368,509 | 874,211 | 595,386 |
| 2 | 12 -08-2020 | 380,190 | 900,100 | 587,929 |
| 3 | 12-15-2020 | 366,076 | 1,091,348 | 706,678 |
| 4 | 12-22-2020 | 284,514 | 729,433 | 520,239 |
| 5 | 12-29-2020 | 467,213 | 1,193,621 | 701,812 |
| 6 | 01-05-2021 | 421,030 | 998,966 | 670,713 |
| 7 | 01-12-2021 | 409,322 | 1,059,747 | 705,176 |
| 8 | 01-19-2021 | 401,670 | 1,029,714 | 666,323 |
| 9 | 01-26-2021 | 460,000 | 1,171,250 | 622,772 |

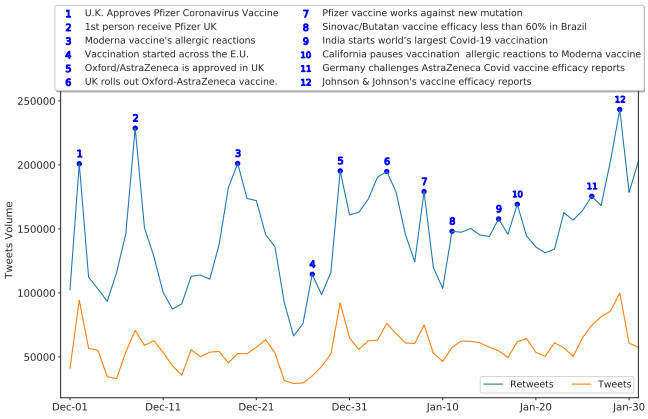



**Figure 1: Daily numbers of tweets and retweets with annotation of important external events.**

period that includes the launch of the first worldwide COVID-19 vaccination campaign (launched on December 8th in the United Kingdom[7]), as well as several other important real-world events that influenced and dictated people's discussions.

Table 1 provides an overview of our dataset, showing the numbers of tweets, retweets, and unique users per week. Here, we label as tweets both original tweets and quoted tweets, which are replies to someone else's original tweet. Despite some fluctuations, there is a general trend towards an increase in user engagement in discussions on vaccination topics, especially in the early weeks of 2021. This general trend reflects the momentum that the topic gains as several important related events occur (e.g., several vaccination campaigns start in different countries of the world).

We delve deeper into the temporal evolution in the vaccination-related discussions by showing the daily time series of the numbers of tweets and retweets in Figure 1. The figure also highlights the days of some important real-life events related to COVID-19 vaccination including, the approval of the Pfizer vaccine in the UK (event #1 in the figure), the start of the COVID-19 vaccination campaign worldwide in India (event #9) and the pause of vaccination in California due to an observed adverse reaction (event #10). Note that there are some significant spikes in the volume of tweets and retweets, which coincide with the identified events. As an illustration, on the same day that the first person received the Pfizer vaccine in the UK, the number of tweets and retweets around the

[7]https://www.bbc.com/news/uk-55227325

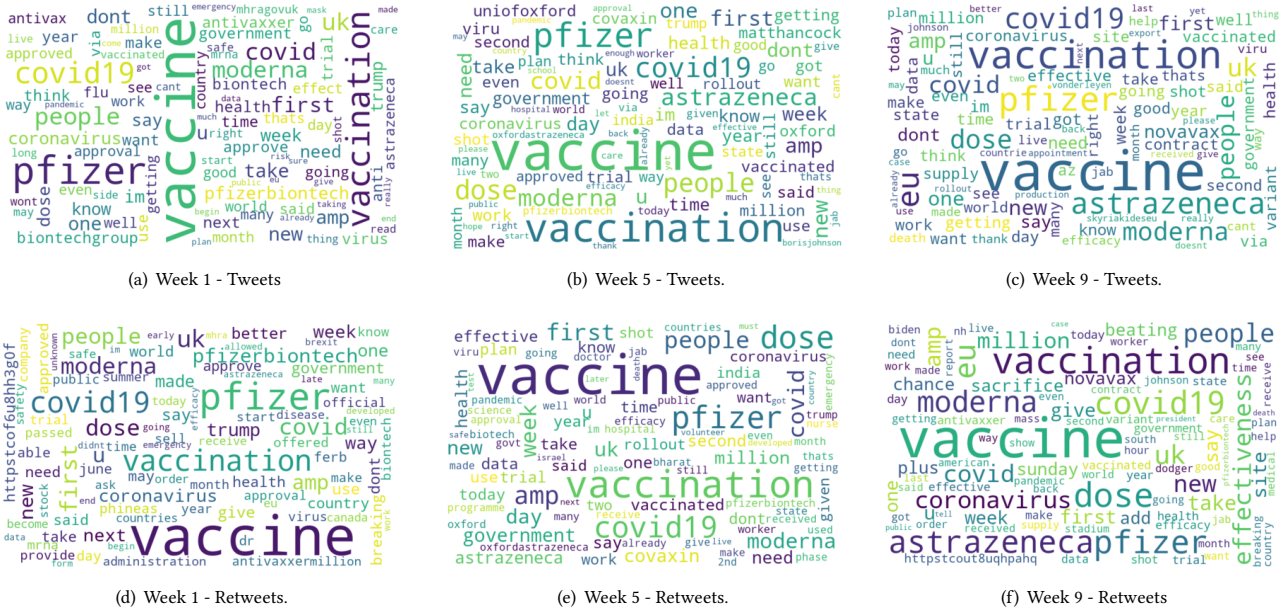
**Table 2: Top-10 most tweeted and retweeted keywords.**

| Keywords | Total | | # Retweet/Tweet |
|---|---|---|---|
| | #Tweets | #Retweets | |
| *vaccine* | 965,813 | 2,378,455 | 2.46 |
| *vaccination* | 904,175 | 2,206,146 | 2.43 |
| *pfizer* | 689,193 | 1,771,997 | 2.57 |
| *moderna* | 319,854 | 874,882 | 1.26 |
| *astrazeneca* | 201,280 | 509,264 | 2.53 |
| *biontech* | 153,439 | 501,570 | 3.26 |
| *anti-vaccine* | 93,846 | 119,898 | 2.19 |
| *anti-vaxxers* | 67,733 | 156,740 | 2.31 |
| *covaxin* | 42,673 | 198,866 | 4.66 |
| *anti-vax* | 33,308 | 67,076 | 2.01 |

vaccination topic increased by 31% and 56%, respectively (event #2). Similarly, an increase of 16% and 20% in the number of tweets and retweets, respectively, was observed on the same day when the efficacy of the Johnson & Johnson's vaccine was officially publicized (event #12). Thus, several external events such as the approval of new vaccines in a country, the start of vaccination campaigns, and news about adverse reactions to the use of some specific vaccine may have propelled the increase in the debate about COVID-19 vaccines on Twitter during the selected period.

To provide a broader view of the online debate about COVID-19 vaccines on Twitter, we look into the popularity of the different keywords (among those selected to guide our data collection) in our dataset. Table 2 lists the top-10 most mentioned keywords, with the total number of tweets and retweets that mentioned each of them as well as the fraction of retweets per tweet for each keyword. The rankings of keywords by numbers of tweets and retweets are very similar, except for the positions occupied by the *covaxin* and *anti-vaxxers* keywords, which have switched positions. The three most mentioned keywords – *vaccine*, *vaccination* and *pfizer* – account for around 75% of all tweets and retweets we gathered, with an average of around 2.5 retweets per tweet. The higher popularity of these keywords is somewhat expected as they reflect the general discussions, not focused on any particular vaccine, as well as the discussions around one of the first vaccines to be approved and used in different countries. Interestingly, despite being one of the least popular keywords in the ranking, *covaxin* generated the largest cascades of retweets per tweet. Also, we found that keywords related to the anti-vaccine discourse (*anti-vaccine*, *anti-vaxxers*, *anti-vax*) are among the least mentioned ones. However, these keywords are also highly diffused over the network, with an average number of retweets per tweet approximately similar (and even larger in some cases) to those for the most popular keywords.

Next, we look into the contents of the tweets and retweets during three different periods covered by our dataset. We do so by showing in Figure 2 the word clouds with the top 100 most frequent words (in numbers of tweets and retweets) during the $1^{st}$, $5^{th}$, and $9^{th}$ weeks. As expected, *vaccine* and *vaccination* are the most frequent keywords in tweets and retweets in all three weeks. We can see that *astrazeneca* becomes relatively more frequent over the weeks, *pfizer* fluctuates a bit, whereas *moderna* seems to peak up by the end of the collection. Moreover, we also observe the emergence of



(a) Week 1 - Tweets

(b) Week 5 - Tweets.

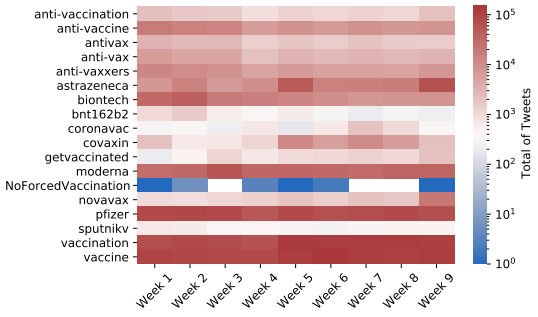
(c) Week 9 - Tweets.

(d) Week 1 - Retweets.

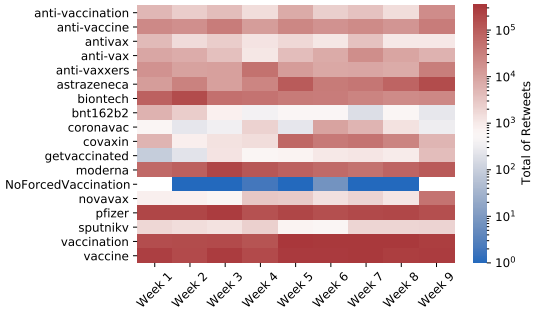
(e) Week 5 - Retweets.

(f) Week 9 - Retweets

**Figure 2: Word clouds with the top-100 most popular words in numbers of tweets and retweets.**




(a) Tweets.




(b) Retweets.

**Figure 3: Keywords popularity over the weeks.**

new words which most likely reflect the reaction of users to what was going on in the real world. For instance, the name of the UK Health secretary *Matt Hancock* appears in the word cloud of the $5^{th}$ week (see Figure 2(b)) most probably as users discussed the stricter lockdown rules installed in the UK at the time[8]. By the end of the period covered by the dataset, other words such as *variant* and *novavax* appear, probably reflecting users reaction to news about vaccine efficacy against the new SARS-COV-2 mutation[9] and the results of the clinical trials for Novavax vaccine[10]. Concerns about vaccine effectiveness seem to increase with time, as words such as *safe*, *efficacy* and *effectiveness* appear more frequently over the last weeks (see Figures 2(d), 2(e) and 2(f)).

As a final overview of our dataset, we look into the popularity of each keyword used to guide our data collection during each week. Keyword popularity is measured in terms of the number of tweets and retweets that mention the keyword. Figure 3 shows these numbers as heatmaps, where the color of each cell reflects the total number of tweets (retweets) on the week (x-axis) on a logarithm scale.

Looking first at the number of tweets, Figure 3(a) shows that *vaccine* and *vaccination* remain as the most popular keywords during all weeks, suggesting that discussions on the vaccination topic in general dominate over the whole period. Regarding keywords related to specific vaccines and manufacturers, *pfizer*, *biontech* and *bnt162b2* are the most popular keywords most probably because they are aliases of the first vaccine used for mass vaccination. Over the weeks, the names of other vaccines (*moderna*, *astrazeneca* and *covaxin*) start gaining more attention. Interestingly, keywords expressing support to vaccination, such as *getvaccinated*, also seem

[8]https://www.bbc.com/news/uk-55489932
[9]https://www.reuters.com/article/us-health-coronavirus-vaccines-variant-idUSKBN29Z0I7
[10]https://edition.cnn.com/2021/01/29/europe/novavax-covid-vaccine-uk-sa-trials-gbr-intl/index.html

to gather more people engagement over time. In contrast, words related to the anti-vaccine discourse, such as *anti-vaccination, anti-vaccine, antivax, anti-vax, anti-vaxxers, Noforcedvaccination* remain with popularity roughly stable throughout all weeks. Similar conclusions hold for retweets, as shown in Figure 3(b).

## 4 ANALYSES AND RESULTS

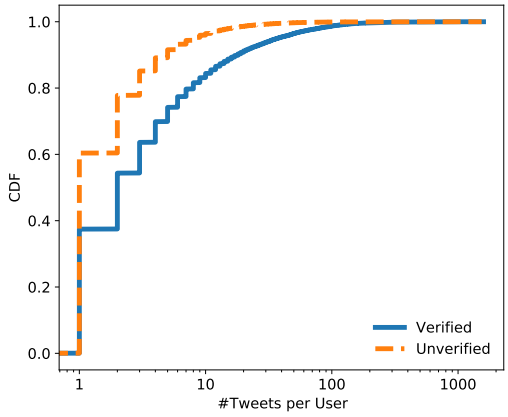
In this section, we characterize how users engage in the vaccination debate using the selected keywords, and we analyze the textual content of the tweets. As mentioned, we use the term *tweet* to refer to original tweets and quoted tweets, which are retweets containing comments.

### 4.1 User Profile

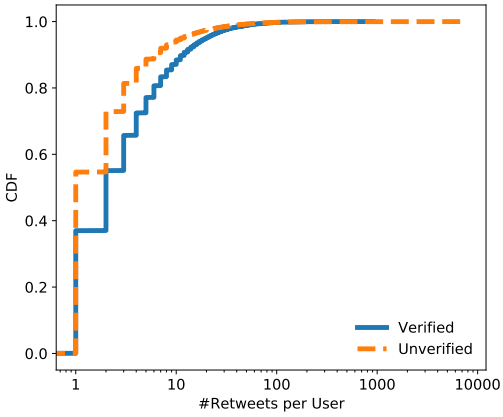
As in other studies [4, 11], we distinguish between verified and unverified user accounts. Verified user accounts are identified by Twitter as of public interest and to be authentic accounts. They are usually the most active accounts when major events occur [4]. Figures 4(a) and 4(b) show the cumulative distribution functions (CDFs) of the numbers of tweets and retweets per verified and unverified user accounts, respectively. As expected, verified user accounts tend to post more tweets and retweets: according to Figure 4, 80% of the verified user accounts posted up to 18 tweets and 12 retweets during the collected period, against to 5 tweets and 6 retweets from the same fraction of non-verified user accounts. Nevertheless, we do observe a few very active users with both verified and unverified accounts, with over 1,500 tweets.

We delve deeper into the five most active verified and unverified user accounts to explore the information they are posting (or relaying). We thus rank each user in terms of their total numbers of tweets and retweets shared. Table 3 shows for each selected user, the number of followers and the total number of times the used mentioned one of the keywords used in our dataset crawling. We omitted personal information due to privacy issues. Note that users mention both more general keywords and specific ones (for instance, *bnt162b2*).

We observe that the verified accounts, the most active users in terms of the number of tweets, include news agencies located in countries as the UK, India, Russia, and the Philippines, while active retweeters include public figures. Influential users and news sources usually weigh in the latest news and particularly, the vaccination debate, using Twitter to amplify their messaging [4].




(a) Tweets.




(b) Retweets.

**Figure 4: Distributions of the number of tweets and retweets per Twitter account.**

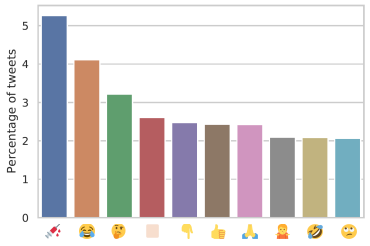
### 4.2 Usage of Emojis

On social media, emojis are used as an alternative means to communicate rather than textual content. They express a visual representation of an emotion, idea, or symbolism. By employing the package *emoji*[11], we were able to find only a smaller fraction of tweets (8.7%) and retweets (9.4%) containing at least one emoji. Figure 5 exhibits the most prevalent emojis in our dataset, while Table 4 presents some examples of tweets and retweets that use some of those emojis. Overall, the most popular emoji in the tweets was the syringe emoji, representing COVID-19 vaccines[12]. Looking at some of the tweets with that emoji (Table 4), we note a diversity of opinions, varying from expressions of happiness and hope with vaccine development to mistrust in the vaccines and conspiracy theories dissemination. Other popular emojis such as "rolling on the floor laughing", “person-shrugging” and “face with rolling eyes” express user emotional states such as fun, indifference, and annoyance, respectively. Other emojis reflecting cognitive states such as “thinking face” (pondering), “thumbs up” (approval or agreement), the need to point some content (“hand point down”) are also among the most popular ones in our dataset. For instance, in Table 4, “thinking face” was used when the tweet author was in doubt or questioning the vaccine’s efficiency or discussing potential conspiracy theories (e.g., Bill Gates vaccine). Finally, the laughing emoji is used ironically and to make fun of some aspect of the debate.

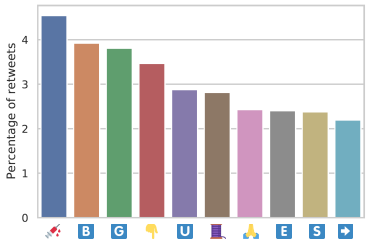
[11] https://pypi.org/project/emoji/
[12] https://blog.emojipedia.org/vaccine-emoji-comes-to-life/

**Table 3: Top-5 most active verified user accounts.**

| Tweets | | |
|---|---|---|
| **User** | **# Followers** | **Keywords (#)** |
| 1 | 247,854 | pfizer(372), astrazeneca(288), biontech(260), vaccination(223), vaccine(216), moderna(187), novavax(21), coronavac(12), covaxin(10), anti-vaccine(3) |
| 2 | 23,026,130 | pfizer(306), astrazeneca(290), vaccine(222), moderna(173), biontech(168), vaccination(157), novavax(19), covaxin(12), coronavac(11), anti-vaccine(10), bnt162b2(1) |
| 3 | 7,889,473 | vaccination(341), pfizer(144), astrazeneca(120), covaxin(116), biontech(96), vaccine(84), moderna(83), novavax(13), anti-vaccine(11), coronavac(6), sputnikv(3), anti-vaccination(2) |
| 4 | 322,931 | sputnikv(237), vaccination(182), pfizer(169), moderna(99), vaccine(84), astrazeneca(71), biontech(56), novavax(8), anti-vaccine(7), anti-vaccination(5), covaxin(4), coronavac(1), anti-vax(1) |
| 5 | 1,179,479 | pfizer(251), astrazeneca(209), vaccination(187), moderna(94), biontech(73), vaccine(52), novavax(29), anti-vaccine(11), coronavac(10), covaxin(3), anti-vaccination(3), bnt162b2(1), anti-vaxxers(1) |
| **Retweets** | | |
| **User** | **# Followers** | **Keywords (#)** |
| 1 | 1,129,136 | vaccination(337), covaxin(258), pfizer(103), astrazeneca(70), moderna(51), biontech(39), vaccine(22), sputnikv(15), anti-vaccine(5), anti-vaccination(4), novavax(2), coronavac(1) |
| 2 | 22,065 | vaccination(84), vaccine(80), pfizer(74), moderna(54), biontech(26), anti-vaccine(7), astrazeneca(7), novavax(5),anti-vaxxers(4), anti-vaccination(3), getvaccinated(3), antivax(2), anti-vax(1), bnt162b2(1), covaxin(1) |
| 3 | 67,242 | vaccination(84), vaccine(77), pfizer(66), moderna(50), biontech(23), novavax(14), astrazeneca(10), anti-vaccine(10), anti-vaccination(7), coronavac(1), anti-vaxxers(1) |
| 4 | 127,639 | vaccine(69), moderna(68), pfizer(65), vaccination(60), biontech(27), anti-vaccine(15), astrazeneca(8), anti-vaccination(5), anti-vaxxers(4), anti-vax(4), antivax(3), novavax(3), getvaccinated(2), covaxin(1) |
| 5 | 38,726 | covaxin(110), vaccination(57), astrazeneca(47), pfizer(41), moderna(27), biontech(21), vaccine(9), novavax(2), anti-vaccine(2), anti-vaxxers(1) |




(a) Tweets




(b) Retweets

**Figure 5: Top-10 most frequent emojis.**

## 4.3 Sentiment Analysis

We now analyze whether the expressed opinion in the tweets is mostly positive, negative, or neutral. To that end, we perform sentiment analysis employing SentiStrength[13] to estimate the strength of positive and negative sentiment in shorter texts [21]. The tool provides an integer score, ranging from -4 (strongly negative) to +4 (strongly positive). Score 0 implies a neutral sentiment. We here consider as *negative*, *neutral* and *positive* tweets and retweets with scores smaller than, equal to and greater than 0, respectively.

[13]http://sentistrength.wlv.ac.uk/index.html

**Table 4: Example tweets with emojis.**

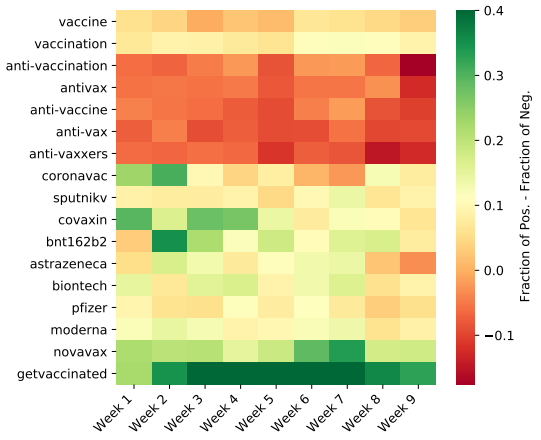
| Emoji | Tweet |
|---|---|
| 📢 | *I'm ready for my vaccine!!!!!", "I want the vaccine and I want it now.* |
| | *The vaccine is not about us. It's about drug companies making trillions for their wealthy investors. Where did Covid come from? Still we do not have the answer but we got a vaccine made from the aids virus and dead babies. Genetic medication for future mutants?.* |
| | *It seems people have short memories !!!!!! THINK BEFORE YOU INJECT YOURSELF WITH A VACCINE THATS ONLY TAKEN 8 MONTHS TO PRODUCE !* |
| 🤔 | *35 years an effective HIV vaccine remains elusive....#COVID19 9 months latter a #vaccine and I'm not a #Antivaxxer* |
| | *Why do Pfizer need legal indemnity for this new vaccine? Why have the government granted it?* |
| | *I don't trust Bill Gates vaccination when he wants to depopulate the world* |
| 😂 | *I love the covid vaccine talk glad people aren't really that stupid* |
| | *@realDonaldTrump But it's a hoax. Yet they were all happy when trump (made) the vaccinef or a virus he said didn't exist. I think they actually think trump made the vaccine!* |

To understand how people's sentiment varies towards each keyword, we contrast tweets (and retweets) mentioning each keyword and classified them as positive and negative over the weeks. Specifically, we build a heatmap, shown in Figure 6, summarizing the differences using a *contrastive score*, calculated as the difference between the fractions of positive and negative tweets (and retweets). The keyword *NoForcedVaccination* was filtered out from the heatmap, as it was not mentioned in all weeks.

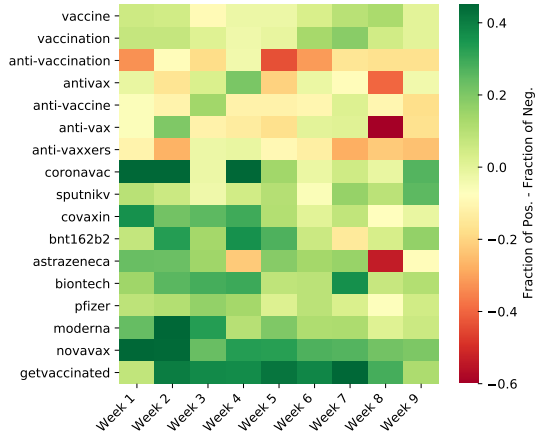
Figure 6(a) shows the results for tweets. Predominantly, users tend to express positive comments towards *vaccine*, *vaccination* and the names of specific vaccines (*bnt162b2, covaxin, biontech, novavax*) keywords over the weeks. However, users make negative posts when mentioning keywords closely related to anti-vaccination movements. We also observe that the positive sentiment towards the possibility of being vaccinated (*getvaccinated*) increases significantly over the weeks as the media reports new vaccine approvals and vaccination campaigns proceed worldwide. Unlike keywords associated with other vaccines, we observe that tweets mentioning *astrazeneca* attracts more negative scores on some weeks (8th and 9th weeks). Looking at media news archives, we found that the *Oxford/AstraZeneca* vaccine was facing some operational problems during that period, which raised doubts about its efficiency for people over 65 years[14].

Since retweets may be seen as a measure of the strength of information diffusion, we also analyze their sentiment scores in Figure 6(b). We observe only slight changes in sentiment polarity as well as in the intensity of the final sentiment scores, compared to tweets (shown in Figure 6(a)). Regarding keywords of vaccine names, people tend to amplify and diffuse more positive tweets, which can be interpreted as a supportive opinion towards the vaccination. Looking at keywords related to anti-vaccine, we can also observe more positive comments. However, this does not necessarily mean that

[14]https://www.theguardian.com/world/2021/jan/22/covid-oxfordastrazeneca-vaccine-delivery-to-eu-to-be-cut-by-60
https://www.theguardian.com/world/2021/jan/26/german-government-challenges-astrazeneca-covid-vaccine-efficacy-reports
https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-01-28-21/index.html




(a) Tweets




(b) Retweets.

**Figure 6: Contrasting sentiment score of users towards the keywords over the weeks.**

people are supporting anti-vax movements, as can be illustrated by the following positive scored retweet: *"Tucker Carlson delivered an anti-vaccine rant tonight, which I will not tweet, encouraging millions of people not to trust the COVID vaccine. It's not complicated: This kind of misinformation will kill people. Any corporation that advertises on Tucker Carlson is complicit."*

## 4.4 Co-occurrence Networks

Now, we analyze how the selected keywords co-occur in the vaccination debate. For that, we here build a co-occurrence network in which vertices represent keywords and edges indicate the co-occurrence of two keywords in the same tweet (retweet). An edge is weighted by the number of such co-occurrences, i.e., the number of tweets (retweets) containing the two keywords. In our visualization, the size of each node is proportional to its degree and the edge thickness is proportional to the number of co-occurrences of the pair of keywords interconnected.

The co-occurrence networks produced for tweets and retweets (again for the $1^{st}$, $5^{th}$, and $9^{th}$ weeks) are presented in Figure 7. Unsurprisingly, some particular pairs of keywords, notably *vaccine* and *vaccination* as well as *vaccine* and *<vaccine-name>*, co-occurred quite frequently in the tweets and retweets. In the first weeks, *vaccination* and *vaccine* keywords are strongly linked to the Pfizer vaccine. However, as time passes, we can see that the *vaccine* term becomes also more strongly associated with other *<vaccine-name>* keywords (e.g., moderna, astrazeneca, coronavac), once these vaccines started being used worldwide. Interestingly, anti-vax-related terms seem to be associated with all vaccine names.

## 4.5 Psycholinguist Analysis

Finally, we study the psycholinguistic properties of tweets and retweets, aiming at finding similarities and differences in the way users communicate. We rely on the Linguistic Inquiry and Word Count (LIWC) lexicon [19] to categorize words in each tweet/retweet in linguistic style, affective and cognitive attributes. For each keyword, we compute the average frequency of the attributes over the tweets and retweets. In our data, we identify 75 attributes, out of the 125 available in LIWC's English dictionary.

We then identify attributes that characterize the discourse around each keyword. To that end, we search for statistical differences across keywords based on the average frequencies of their respective attributes. We first use Kruskal's non-parametric test [12] with a statistical confidence level of 95% to select only attributes for which there is a significant difference across keywords. We identified 69 attributes with statistically significant differences.

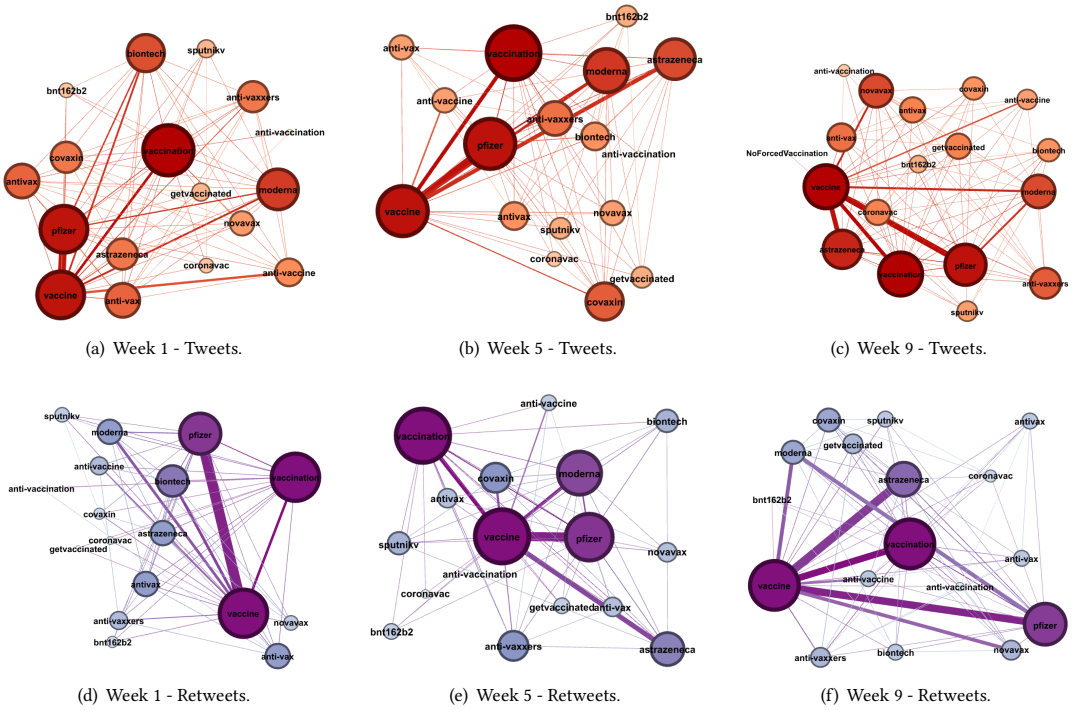
Having identified those attributes, we then rank them according to their capacity to discriminate across different keywords, estimated by the Gini Coefficient [25]. We use the top-20 to create visualizations that can better highlight topics associated with keywords.

Figure 8 shows a heatmap for the top-20 ranked attributes. The heatmap cells in a column indicate the relative deviation of the given attribute for the given keyword from the other keywords. In other words, each column (attribute) is normalized following the z-score – i.e., $z = (x - mean)/std$. Thus, each value gets subtracted from the average of the column, then divided by the standard deviation of the column. Locations are color-coded red (resp. blue) when the attribute is more (resp. less) present than the average.

The results in Figure 8 show that tweets and retweets containing the target keywords are quite different with respect to the selected attributes. For instance, tweets and retweets with anti-vaccination keywords frequently use words regarding *death*, *anger* and *negative emotion*, but seldom words related to *health*. Tweets and retweets mentioning vaccine-related keywords focus mostly on *health* (safety, efficacy and necessity), *work* (economic motives) and *religion* (moral and religious concerns), corroborating previous studies performed by the First Draft Organization.[15] Interestingly, posts including *getvaccinated* keyword strongly express *anxiety* feelings.

In summary, LIWC is a useful tool to analyze the content of the ongoing debate on Twitter, providing a nice picture of the

[15]https://firstdraftnews.org/long-form-article/under-the-surface-covid-19-vaccine-narratives-misinformation-and-data-deficits-on-social-media/




(a) Week 1 - Tweets. (b) Week 5 - Tweets. (c) Week 9 - Tweets.

(d) Week 1 - Retweets. (e) Week 5 - Retweets. (f) Week 9 - Retweets.

**Figure 7: Co-occurrence networks for tweets and retweets.**

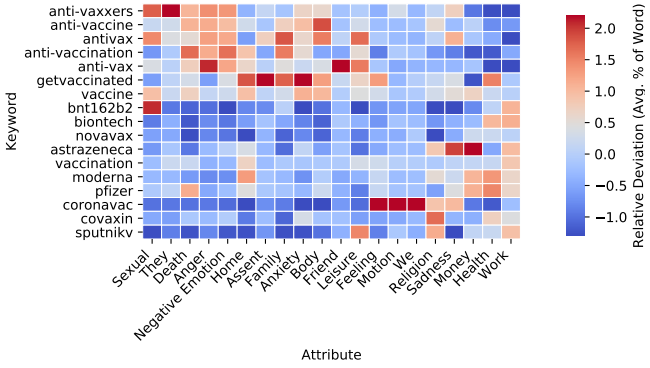



**Figure 8: Top-20 LIWC attributes extracted from the collected tweets and retweets.**

dominant COVID-19 vaccines narratives in our dataset. At last, it is important to emphasize that this work is a preliminary study about the vaccination debate on Twitter. Deeper analyses on text content should be performed to identify users who are either criticizing the anti-vax movements or supporting them.

## 5 CONCLUDING REMARKS

In this paper, we investigate the dynamics of the online debate around the COVID-19 vaccination through the lens of Twitter. For conducting our analyses, we used over 12 million tweets crawled during two months corresponding to the early stages of vaccination in the world.

Our results show that the volume of tweets is highly correlated with external events such as vaccine rollout campaigns and reports of vaccine efficacy studies. Examining the polarity of the discussions about vaccination we notice the negative sentiment towards topics related to the anti-vaccination movement. Finally, we observed that health, work, and religious attributes were more common in tweets with vaccine-related words. On the other hand, tweets mentioning anti-vaxx terms were more focused on terms related to death, anger, and negative emotions. This work is a preliminary study and because of the volume of data and the recency of the topic, we chose to use simpler and more general analysis. As future work, we plan to focus on the different narratives, discriminating the actors in the vaccine debate such as the real anti-vaccination supporters.

**Acknowledgments**: The research leading to these results has been partially supported by CNPq, FAPEMIG, CAPES, and the Covid Data Analytics project, funded by PrPq/UFMG/SESU/MEC (23072.211119/2020-10).

## REFERENCES

[1] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11-7 (2016).

[2] Erika Bonnevie, Jaclyn Goldbarg, Allison Gallegos-Jeffrey, Sarah Rosenberg, Ellen Wartella, and Joe Smyser. 2020. Content Themes and Influential Voices Within Vaccine Opposition on Twitter. *American Journal of Public Health* 110, S3 (2020), S326–S330.

[3] David Broniatowski, Amelia Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra Quinn, and Mark Dredze. 2018. Weaponized Health

Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. 108, 10 (10 2018), 1378–1384.
[4] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill* 6, 2 (May 2020).
[5] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *Scientific Reports* 10, 1 (2020), 16598.
[6] Matthew DeVerna, Francesco Pierri, Bao Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Cristopher Torres-Lugo, Kai-Cheng Yang, Fil Menczer, and John Bryden. 2021. CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines. arXiv:2101.07694 [cs.SI]
[7] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proc. of CIKM*.
[8] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics'in response to COVID-19 epidemics. *Nature Human Behaviour* 4, 12 (2020), 1285–1293.
[9] Kiran Garimella and Ingmar Weber. 2017. A Long-Term Analysis of Polarization on Twitter. *Proc. of ICWSM*.
[10] Gloria Kang, Sinclair Ewing-Nelson, Lauren Mackey, James Schlitt, Achla Marathe, Kaja Abbas, and Samarth Swarup. 2017. Semantic network analysis of vaccine sentiment in online social media. *Vaccine* 35, 29 (2017), 3621 – 3638.
[11] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie Akl, and Khalil Baddour. 2020. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus* 12, 3 (3 2020), e7255.
[12] William Kruskal and Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.
[13] Tanushree Mitra, Scott Counts, and James Pennebaker. 2016. Understanding Anti-Vaccination Attitudes in Social Media. In *Proc. of ICWSM*.
[14] Francesco Pierri, Silvio Pavanetto, Marco Brambilla, and Stefano Ceri. 2021. VaccinItaly: monitoring Italian conversations around vaccines on Twitter. arXiv:2101.03757 [cs.SI]
[15] Zubair Shah, Didi Surian, Amalie Dyda, Enrico Coiera, Kenneth Mandl, and Adam Dunn. 2019. Automatically Appraising the Credibility of Vaccine-Related Web Pages Shared on Social Media: A Twitter Surveillance Study. *J Med Internet Res* 21, 11 (2019).
[16] Gautam Shahi, Anne Dirkson, and Tim Majchrzak. 2020. An Exploratory Study of COVID-19 Misinformation on Twitter. arXiv:2005.05710 [cs.SI]
[17] Mirela Silva, Fabrício Ceschin, Prakash Shrestha, Christopher Brant, Juliana Fernandes, Catia Silva, André Grégio, Daniela de Oliveira, and Luiz Giovanini. 2020. Predicting Misinformation and Engagement in COVID-19 Twitter Discourse in the First Months of the Outbreak. arXiv:2012.02164 [cs.SI]
[18] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. arXiv:2003.13907 [cs.SI]
[19] Yla Tausczik and James Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[20] Lara Tavoschi, Filippo Quattrone, Eleonora Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni, and Pier Luigi Lopalco. 2020. Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines & Immunotherapeutics* 16, 5 (2020), 1062–1069.
[21] Mike Thelwall. 2017. The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. In *Cyberemotions*. Springer, 119–134.
[22] Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid Butt. 2021. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health* 3, 3 (2021), e175–e194.
[23] Wei Wu, Hanjia Lyu, and Jiebo Luo. 2021. Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story. arXiv:2101.06321 [cs.SI]
[24] Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. 2020. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. *CoRR* abs/2004.14484 (2020). arXiv:2004.14484 https://arxiv.org/abs/2004.14484
[25] Shlomo Yitzhaki. 1979. Relative deprivation and the Gini coefficient. *The quarterly journal of economics* (1979), 321–324.