Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

VaccinEU: COVID-19 Vaccine Conversations on Twitter in French, German and Italian

Marco Di Giovanni[1, 3,*], **Francesco Pierri**[1,2,*] **Christopher Torres-Lugo**[2] **and Marco Brambilla**[1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
[2]Observatory on Social Media, Indiana University, Bloomington, USA
[3]Università di Bologna, Italy
francesco.pierri@polimi.it

Abstract

Despite the increasing limitations for unvaccinated people, in many European countries there is still a non-negligible fraction of individuals who refuse to get vaccinated against SARS-CoV-2, undermining governmental efforts to eradicate the virus. We study the role of online social media in influencing individuals' opinion towards getting vaccinated by designing a large-scale collection of Twitter messages in three different languages – French, German and Italian – and providing public access to the data collected. Focusing on the European context, our VaccinEU dataset aims to help researchers to better understand the impact of online (mis)information about vaccines and design more accurate communication strategies to maximize vaccination coverage. Data can be fully accessed in a Dataverse repository and a GitHub repository.

Introduction

Less than a year into the COVID-19 pandemic, the first vaccine was approved and made available to the public[1], providing an effective tool to fight the spread of the virus (Orenstein and Ahmed 2017). Vaccination programs started towards the end of 2020 in most European countries, and as of December 2021 over 700 M doses have been administered according to Our World in Data[2]. However, despite the large availability of vaccines, vaccine uptake exhibits a large variability across different countries, ranging from 40% of people vaccinated with at least one dose in Romania to 90% in Portugal[3]. This indicates that a considerable number of people are still hesitant to get vaccinated, and that it will be hard to reach herd immunity .

Research in the past highlighted the role of online social media in promoting and amplifying negative views about vaccines (Burki 2019; Broniatowski et al. 2018; Johnson et al. 2020). Specifically to the COVID-19 pandemic, concern has recently risen around the 'infodemic' (Zarocostas 2020; Yang et al. 2021; Gallotti et al. 2020) of misleading information about the virus spreading online, and it has been shown that online misinformation might negatively influence individuals' opinion towards getting vaccinated (Pierri et al. 2021a; Loomba et al. 2021).

In this paper, we describe a data resource which will allow researchers and academics to study the impact of online conversations about COVID-19 vaccines on Twitter in three different languages: French, German, and Italian.

Specifically to the Italian context, Righetti (2020) and Cossard et al. (2020) analyzed the debate on Twitter around the 2017 mandatory child vaccination law, observing the spread of problematic information and highlighting the presence of echo chamber effects (Cinelli et al. 2021). Gargiulo et al. (2020) obtained similar results when analyzing French data, finding that defenders and critics of vaccines focus on different topics, and that, while there are more defenders, critics are more active and coordinated. To the best of our knowledge, there is no previous work which analyzes vaccine conversations on social media in German langauge.

Our contribution is manifold. We curated a list of vaccine-related keywords as complete as possible with the help of native speakers, using a snowball sampling approach (DeVerna et al. 2021), and collected over 70 million tweets in three different languages, from November 1st 2020 to November 15th 2021, using a combination of streaming and historical search Twitter APIs. To the best of our knowledge there are no such datasets publicly available, with the only exception of VaccinItaly (Pierri et al. 2021b) in Italian language. We provide public access to this data in agreement with Twitter terms of service by releasing *ids* of tweets which can be used to retrieve full objects via APIs. For each language, we further collected a list of hashtags which strongly state a stance in favor or against vaccination, and we manually annotated a random sample of 1,000 tweets with four labels (Pro-vaccines, Anti-Vaccines, Neutral, Out-of-context). We provide full access to this metadata, which can be used to better understand the polarized debate around vaccinations and train machine learning classifiers to automatically detect anti-vaccination messages (Di Giovanni et al. 2021). Finally, we provide some preliminary analyses of the dataset in terms of volumes, hashtags, sources, geolocation and coordinated activity.

The outline of the paper is the following: we first overview

[*]These authors contributed equally.

[1]https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine

[2]https://ourworldindata.org/covid-vaccinations

[3]https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html

existing datasets which relate to our work. Then, we describe in detail the data collection process. Next, we provide some preliminary analyses of the data, leaving more sophisticated analyses for future work. Finally, we discuss limitations and potential uses of this dataset.

Related Datasets

Here we describe some public data resources recently released to study conversations around COVID-19 vaccines on social media.

At the beginning of 2021, DeVerna et al. (2021) released the first Twitter dataset conceived to investigate English language online conversations around COVID-19 vaccines. They used a snowball sampling approach to curate a list as complete as possible of terms related to vaccines, and they provide public access to *ids* of tweets collected since the beginning of January 2020. They also have an associated online dashboard (CoVaxxy[4]), where they provide an interactive visualization of the relationship between online misinformation spreading on Twitter and the evolution of the US vaccination program. Associations between online misinformation and vaccine hesitancy were reported in Pierri et al. (2021a) leveraging their data.

Pierri et al. (2021b) released a public dataset of Italian language tweets related to vaccines and collected since December 2020 to October 2021[5]. They also set-up a collection of public posts about vaccines shared by public Facebook pages and groups and gathered through Crowdtangle. Similar to CoVaxxy, they provide an online dashboard where they show visualizations of the interplay between Twitter conversations and the vaccination program in Italy[6].

Muric, Wu, and Ferrara (2021) focused on antivaccine narratives on Twitter and publicly released two data collections, one streaming keyword–centered with more than 1.8 million tweets, and another historical account–level collection with more than 135 million tweets. Both collections are based on English language keywords. They showed that Twitter users who engaged the most in antivaccination narratives are politically right-wing leaning, and that questionable news sources are very active in promoting negative views about vaccines.

Hayawi et al. (2021) focused on online misinformation around COVID-19 vaccines. After collecting over 15 million tweets, they manually labeled a sample of 15k tweets with the help of medical experts in order to identify unsubstantiated claims and misleading information about vaccines. They eventually trained and test machine learning classifiers on these tweets, reaching up to 98% of F1-score in the task of classifying vaccine misinformation.

In addition to the aforementioned resources, several datasets have been released to study the COVID-19 pandemic on Twitter, providing oftentimes useful metadata (geolocation, sentiment, gender, etc) in addition to raw tweet *ids* (Banda et al. 2021; Chen et al. 2020; Lopez and Gallemore 2021; Imran, Qazi, and Ofli 2021).

[4]https://osome.iu.edu/tools/covaxxy

[5]https://github.com/frapierri/VaccinItaly

[6]http://genomic.elet.polimi.it/vaccinitaly/

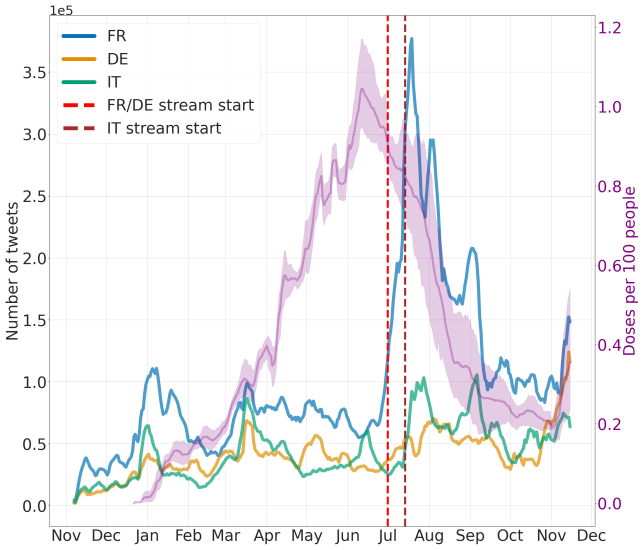

Figure 1: Daily number of vaccine-related tweets collected in different languages (left y-axis), along with the daily number of doses administered per million population (right y-axis) in several European countries (Austria, Belgium, France, Germany and Italy). All time series are smoothed with a 7-day average. Daily vaccinations are obtained from Our World in Data (Mathieu et al. 2021) and correspond to the average over different countries with 95% C.I. Vertical dashed lines indicate the beginning of the streaming collection for France and Germany (red) and Italy (brown).

Data Collection

In this section, we describe our data collection process. We detail every design choice made to obtain a dataset as complete and unbiased as possible.

Twitter APIs

We use both the standard streaming Filter API v1.1[7] and the new historical Search API v2[8] to collect tweets related to vaccines in three different languages: French, German, and Italian.

The Filter API filters tweets that match a defined query in a real-time fashion, up to 1% of the global stream. Approximately 500 million tweets are shared every day on Twitter[9], and as shown in Figure 1 we collected at most 350k tweets in a day, thus we likely never incur in this limitation. We started the streaming collection of German and French tweets on July 1st, 2021 and Italian tweets on July 14th, 2021. We filtered tweets by language specifying the *lang* parameter in the queries.

We experienced network malfunctioning issues in some cases, and to fill them we used the Historical Search API, which was released at the beginning of 2021, that allows

[7]https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview

[8]https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

[9]https://www.internetlivestats.com/twitter-statistics/

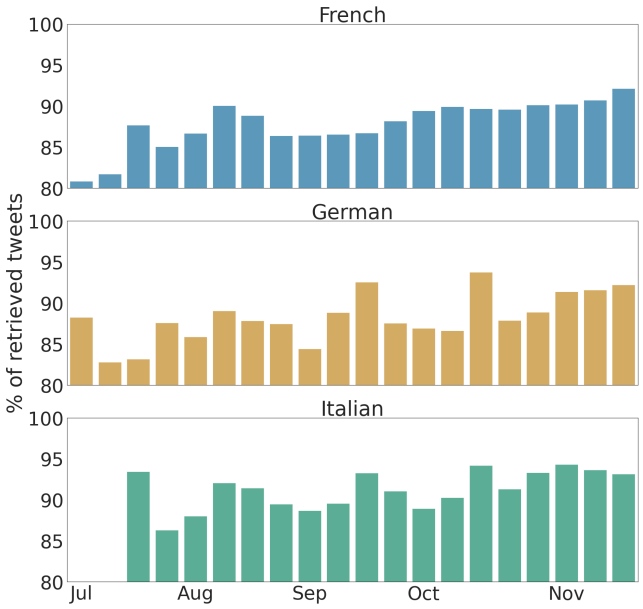

Figure 2: Percentage of tweets successfully retrieved using the *GET statuses/lookup* endpoint. Each point corresponds to a different week, for which we extract a random sample of 10k tweets which we attempt to retrieve. The procedure was done on December 16th, 2021.

academics and researchers to perform a full-archive search with a set of selected keywords. We also employed it to recover all tweets shared since November 1st, 2020 to June 30th, 2021 (N.B. July 13th for the Italian language).

We remark that data collected through the historical Search is not complete, due to Twitter's Terms of Service. Twitter does not allow to retrieve deleted tweets nor those shared by protected or suspended accounts[10]. Nevertheless, we believe that it is still useful to obtain a collection of vaccine-related tweets as complete as possible. To provide a rough estimate of the amount of tweets that we might lose in the process, we hydrate a random selection of 10k tweets per week collected with the streaming API. We show the percentage of tweets recovered running the *GET statuses/lookup* endpoint on December 16th, 2021 in Figure 2. We can see that we lost between 5 and 20% of shared tweets, and that this number likely increases as we search farther in the past.

Query Keywords

Both Filter and Search APIs require one or more keywords to collect relevant tweets. An accurate selection of keyword is crucial to obtain a comprehensive dataset.

We iteratively selected the keywords with the help of three native speakers for each language using a snowball sampling approach (DeVerna et al. 2021). We selected as initial set of keywords the translation of very generic vaccine-related words such as "vaccine" and "vaccination" in French, German, and Italian. We made sure to include every grammatically correct variation of words since Twitter APIs perform case-independent *exact* match of keywords and the tokenized texts of tweets (e.g., the tweet "Vaccines are necessary." will be selected if we include in our query the keyword "vaccines", but it will not be collected when including the keyword "vaccine"). This might be problematic for languages like German, where words can appear with four different cases (nominative, accusative, dative, and genitive).

At each round, we used the historical API to filter tweets in the entire period November 2020 - June 2021, and we inspected the most frequent co-occurring words with those in the query. Then, we augmented our list of keywords with those clearly related to vaccines, including specific hashtags, as indicated by native speakers. For instance, we include "#Igetvaccinated" because tweets containing this hashtag will not be collected by simply using "vaccinated" as keyword.

The final list of keywords for each language is available in our Dataverse[11].

Gold Hashtags

The goal of our project is to understand the influence of positive and negative opinions about vaccines shared on Twitter. To this aim, we collected sets of hashtags that indicate the stance (Pro or Anti vaccines) of tweets with high likelihood. We define them as Gold Hashtags (GH), and similarly to our query keywords, we used a snowball sampling approach to obtain a set of hashtags for each language with the help of annotators. We assume that tweets sharing one or more GH from the same stance express that specific view about vaccines, but this might not always hold true.

We begun with the selection of one GH for each stance, respectively the translation in different languages of "Iwillgetvaccinated" for Pro and "Iwillnotgetvaccinated" for Anti[12]. We iteratively added new GHs inspecting those that co-occurred the most with the initial set of hashtags, based on whether they clearly expressed a stance on vaccines. We discarded hashtags when they generically referred to the topic of vaccines, but whose stance was unidentifiable (such as *#vaccine*). We also discarded hashtags that, although their stance seemed clear to the annotators, highly co-occurred with GH of both stances. We iterated this procedure three times. The final list of hashtags is available in our repository.

Table 1 shows statistics of GHs. Manually inspecting a small set of tweets which included both a Pro and Anti GH, we noticed that most often they do not state a clear stance and usually include questions and pools.

Gold Labels

In addition to hashtags which express a specific stance towards vaccines, we asked our native speakers to manually annotate a sample of random tweets. We randomly

[10] As a matter of fact, the streaming API also does not provide tweets which are shared by protected accounts.

[11] https://doi.org/10.7910/DVN/NZUMZG

[12] We checked that these hashtags were actually shared by Twitter users in each language.

Gold Hashtags	French	German	Italian
Pro	161,871	41,933	53,374
Anti	129,926	115,512	83,097
Both	585	1,224	451

Table 1: Statistics of tweets sharing Gold Hashtags.

Gold Labels	French	German	Italian
Pro-Vaccines	419	547	314
Anti-Vaccines	135	108	151
Neutral	279	169	458
Out-of-Context	167	176	77

Table 2: Statistics of manually annotated tweets.

	French	German	Italian
Tweets	38,198,048	15,573,108	16,581,210
Users	1,586,071	615,317	656,578
URLs	4,749,359	2,808,657	2,686,055

Table 3: Breakdown of the datasets in terms of unique tweets, users and URLs shared, for each language.

picked 1,000 unique tweets for each language, thus discarding retweets, and we asked two annotators to attach one of four "Gold Labels": Pro-vaccines, Anti-vaccines, Neutral, Out-of-Context. We gave them the following guidelines: Pro- and Anti-vaccines tweets should clearly express a stance about vaccines; Neutral tweets should not express any stance, or their stance is unclear; finally Out-of-Context tweets are tweets not related to COVID-19 vaccines (e.g., animal vaccines). A third annotator solved the conflicts by picking one of the two labels for the tweets when they did not agree. We report statistics of the labels in Table 2.

Data Availability

In agreement with Twitter terms of service, we provide public access to the entire list of tweet *ids* in our Dataverse dataset[13] and Github repository[14]. These can be "hydrated", i.e., fully retrieved using the *GET statuses/lookup* endpoint of Twitter API, unless they were deleted or their author suspended in the meantime.

In addition to the raw list of *ids*, organized in daily files, we provide the list of *ids* of tweets which contain Pro and Anti vaccine Gold Hashtags (as defined in previous subsection). We also provide the text of tweets labelled using the four Gold Labels defined in the previous subsection.

Data Characterization

In this section we provide descriptive statistics of the data collected in terms of volumes, hashtags, news sources and geolocation. These should be seen as potential uses of the dataset, whereas we leave more sophisticated analyses for future work. In Table 3 we provide basic statistics of the data in terms of tweets, users and URLs for each language.

[13]https://doi.org/10.7910/DVN/NZUMZG

[14]https://github.com/DataSciencePolimi/VaccinEU

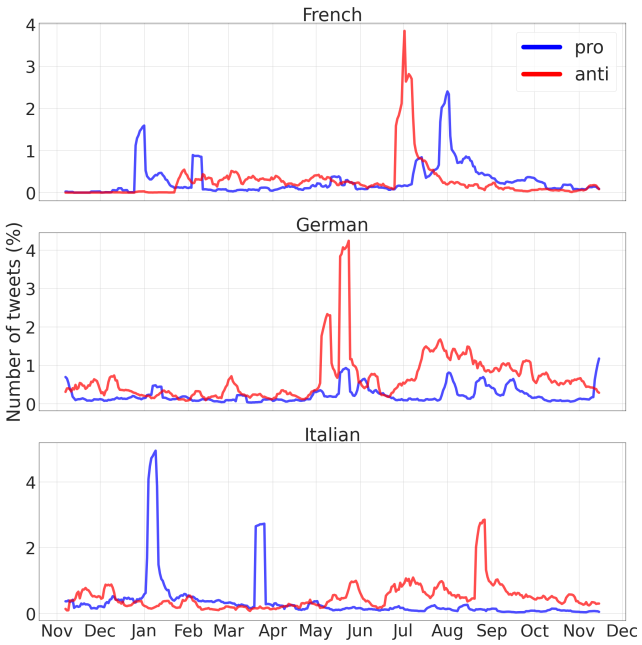

Figure 3: Daily percentage of tweets and retweets sharing pro and anti vaccine hashtags, respectively in blue and red, for each language. We count tweets which contain only hashtags belonging to one of the two classes. Time series are smoothed with a 7-day average.

Volumes

In Figure 1 we show the daily number of tweets collected for each language, highlighting with two vertical lines when the streaming collection starts for French and German (July 1st 2021), and for Italian (July 14th 2021). As a reference for the COVID-19 vaccination programs, we show the daily number of vaccine doses administered (per 100 people) (Mathieu et al. 2021) averaged over different European countries where these languages are spoken, namely Austria, Belgium, France, Germany, and Italy.

We can see that overall the daily volume of French tweets is much higher compared to the other two languages, and this might be due to the fact that it is more widespread, especially in the African continent (cf. also Table 3).

We bserve a peak of activity across all languages in January, corresponding to the beginning of the vaccination program, and another one in March when alleged links between the AstraZeneca vaccine and blood clots became viral in mainstream media. In summer there is an outstanding increase of French and Italian tweets, probably linked to the introduction of the restrictions for unvaccinated people, whereas towards fall we can see that the topic is trending across all languages (especially German) following a slight increase in the number of vaccinations. In fact, there is a significant Pearson correlation between the daily volumes of tweets collected in different languages (in the range 0.56-0.71, $P \sim 0$).

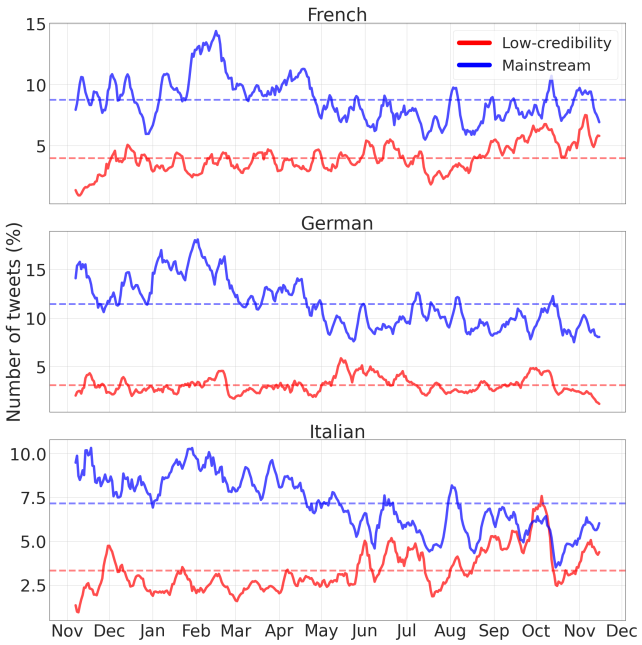

Figure 4: Daily percentage of tweets and retweets sharing links to low-credibility and mainstream news websites, respectively in red and blue, for each language. Time series are smoothed with a 7-day average. Dashed lines represent the mean value over the entire period of observation.

Hashtags

When we look at the top-10 most shared hashtags in the three languages, we observe that they mostly contain generic references to the pandemic (e.g. ”vaccin”, ”covid19”, ”corona”), the debate around the introduction of vaccination documents (e.g. ”passsanitaire” in French or ”greenpass” in Italian) and politicians (e.g. ”macron” and ”draghi”).

In Figure 3 we show instead the daily percentage of tweets sharing Pro and Anti vaccine Gold Hashtags (computed over the total number of tweets shared in that day), using the list of GHs specified in the Data Collection section, for each language. For each day we count tweets and retweets which contain hashtags belonging to only one of the two classes. For what concerns French, we notice a peak of activity for Pro vaccine hashtags at the beginning of the campaign (January 2021) and another in late summer, which follows a strong peak of Anti vaccination hashtags. For what concerns German, we notice little sharing activity for Pro vaccine hashtags, whereas Anti vaccination ones exhibit a peak at the beginning of summer, and then show an increasing trend towards the beginning of fall. Finally, for what concerns Italian, we notice a large number of Italian Pro vaccine hashtags at the beginning of the campaign in January, and likewise in correspondence of the AstraZeneca blood clots ’event’. Towards summer, similarly to other langauges, we notice an increase in the sharing of Anti vaccination hashtags. Overall, daily volumes stay in a similar range across different languages (0-4%).

News Sources

We now investigate the prevalence of low-credibility by using a source-based approach to label news articles, i.e., we label sources based on lists compiled by journalists, researchers and fact-checkers and we propagate the label to all URLs linking to these websites. This approach is limited, since not all stories published on a disinformation website are fake, but it is widely adopted in the literature to study low-credibility content at scale (Yang et al. 2021; Bovet and Makse 2019; Shao et al. 2018; Caldarelli et al. 2021; Brena et al. 2019). As a reference, we consider publishers of mainstream news as a proxy for reliable information similar to (Yang et al. 2021).

Specifically, we aggregate three different sources of labels:

- a list of 60+ Italian low-credibility websites which were flagged by Italian fact-checkers and journalists for sharing disinformation, misinformation, fake news, etc introduced in (Pierri, Artoni, and Ceri 2020) and employed in (Pierri 2020; Pierri, Piccardi, and Ceri 2020; Guarino et al. 2021; Pierri et al. 2021b). It is available in our repository.
- a list of over 600 low-credibility domains based on information provided by the Media Bias/Fact Check website (MBFC, mediabiasfactcheck.com) (Yang et al. 2021). It is available in our repository.
- a list of credibility scores in the range $[0, 100]$ provided by NewsGuard (https://www.newsguardtech.com/it/), a journalistic organization that rates websites on their tendency to spread true or false information. In particular, we consider publishers with a score less than 60 as low-credibility (as suggested by NewsGuard), and those with a score higher than 60 as mainstream. We cannot disclose this list because the data is proprietary.

In Figure 4 we show the daily percentage of tweets and retweets containing a link to low-credibility and mainstream news websites. We can see that the amount of low-credibility is smaller yet non negligible compared to mainstream news. It is also stationary around the mean value of the entire period (in the range $[2.5\%, 4.8\%]$) in all languages, whereas mainstream coverage of vaccines exhibits a decreasing trend towards summer for German and Italian. Interestingly, we can notice that around October-November 2021 the amount of Italian misinformation circulating on Twitter was higher than mainstream news. However, we remark that our lists are not exhaustive, and that these estimations should be consider as a lower bound for both low-credibility and mainstream information.

We further investigate which are the most shared low-credibility news websites in different languages. In Figure 5 we provide the Top-15 ranking of such websites. We can see a similar prevalence on Twitter of most popular misinformation websites, with the uppermost websites being shared over 100k times. In French: ”francesoir.fr” is a popular tabloid which has been criticised for publishing false claims about the COVID-19 pandemic. In German: ”reitschuster.de” is the blog of a political commentator (Boris Reitschuster) which has a borderline score according

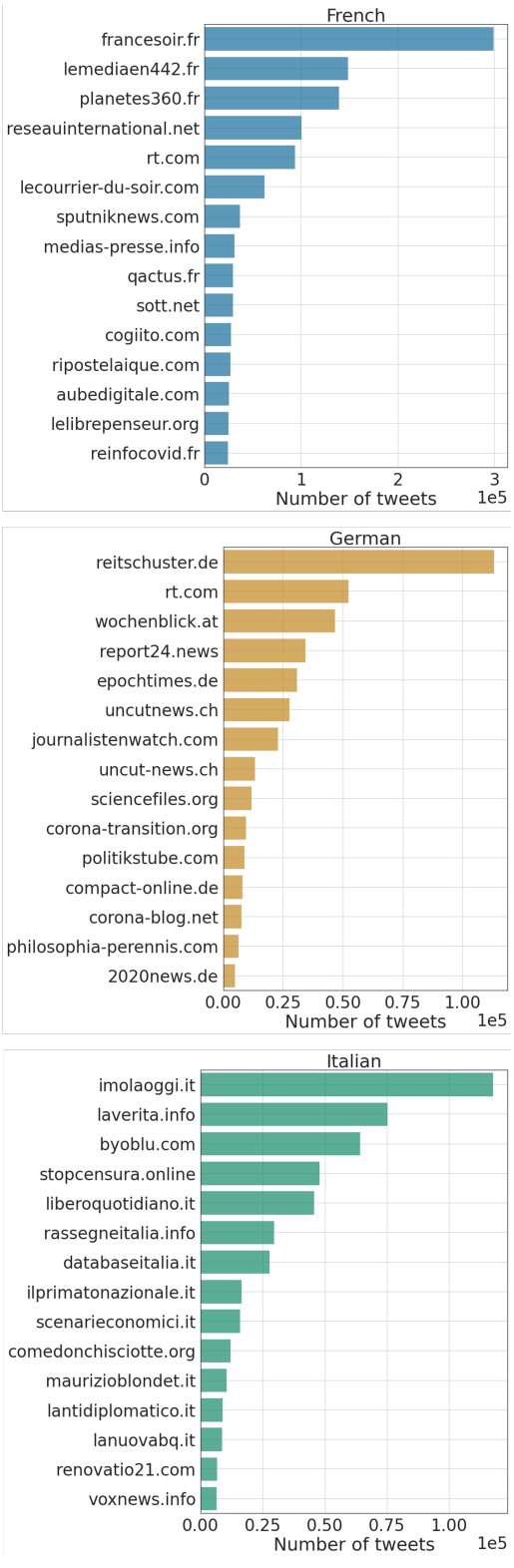

Figure 5: Top-10 most (re)tweeted low-credibility websites in different languages.

to Newsguard (it's rated 59.5 out of 100) and that has been flagged for sharing misinformation about the pandemic. In Italian: "imolaoggi.it" is a news website which has been repetitiously flagged for sharing hoaxes, misinformation and fake news. We leave further investigation of these websites for future work.

Geolocation

We used the methodology described in Mejova and Kourtellis (2021) to locate users in our dataset and estimate the geographical composition of the data collected for each language. This employs the GeoNames[15] location database to match the user-specified free-text location strings to a location. Not all users can be geolocated in this way, because many do not put a string in the "location" field. We report the following:

- French: over 750k users and 17.4 million tweets are geolocated. Around 55% users are geolocated to France and are responsible for 67% of the geolocated tweets. Second and third most frequent countries are United States ($\sim$ 7% tweets) and Canada ($\sim$ 4% tweets).
- German: over 270k users and 7.8 million tweets are geolocated. 66% of the users and tweets are geolocated in Germany. Second and third most frequent countries are Austria ($\sim$ 8% tweets) and Switzerland ($\sim$ 7.7% tweets).
- Italian: over 290k users and 7.5 million tweets are geolocated. Around 52% of the users are geolocated to Italy, and they shared over 80% of the geolocated tweets. Second and third most frequent countries are the United States ($\sim$ 4% tweets) and France ($\sim$ 3% tweets).

The approach is not completely accurate, since it is based on a simple string matching, but we can observe that indeed most of the accounts are geolocated in the main countries where each language is spoken, namely France, Germany, and Italy. For what concerns French, we do not get a large number of users geolocated in African countries, but a further investigation is needed to understand whether the geolocation technique is not working properly or Twitter is not very used in those countries.

Coordinated Activity

In this section we try to identify coordinated activity on the dataset by applying a coordination detection framework (Pacheco et al. 2021). While coordination may occur over many different possible dimensions, here we focus our attention on coordinated sharing of URLs. Other dimensions could be explored to identify other coordinated accounts, based for instance on shared hashtag and/or images.

Specifically, for each date in the period under analysis, we built a bipartite network of users and URLs they shared on native tweets (excl. retweets and quote retweets). Then, we projected it to users such that two users would be connected if they shared the same URL. Edges between users are thus weighted by the number of same URLs that they shared.

[15]The code for geolocation can be found at https://sites.google.com/site/yelenamejova/resources

To focus on the most suspicious users, we filtered out edges with a weight smaller than 10, and removed singleton nodes resulting from this procedure. Finally, we aggregated all daily networks such that edge weights correspond to the number of days in which we found a pair of users sharing the same URLs at least 10 times.

The resulting networks, one for each language, can be found in Figure 6. The network for French has 1,888 nodes and 28,951 edges, for German it has 157 nodes and 236 edges, and for Italian it has 392 nodes and 1,555 edges. The size of nodes corresponds to the percentage of links to low-credibility domains, as defined in the previous section, and edge are ranked by their weight, with thicker edges indicating a higher weight.

The Italian and French networks are dominated by a single large component. In contrast, the German one contains two large components. Of these two components, the one on the bottom left is densely connected with thicker edges while the one in the middle is sparser with thinner edges. This behavior makes the former more suspicious than the latter. Additionally, all of these components exhibit dissimilarities on uniformity or variety of low credibility sources shared. For example, the accounts found in the Italian network shared a lower percentage of these sources compared to those in France network.

Conclusions

We presented a large-scale dataset of Twitter messages related to vaccines in three different languages (French, German, and Italian), which allows to investigate the impact and the influence of online conversations about COVID-19 vaccines on social media.

We provided a few preliminary analyses of the dataset. We showed that throughout 2021 there were a few peaks of attention around the topic in correspondence of the beginning of vaccination programs, the AstraZeneca blood clots and the introduction of limitations for unvaccinated people. We showed that hashtags expressing positive and negative views about vaccines were highly shared in different periods depending on the language, and that online misinformation accounts for around 5% of the tweets shared in each language. We also showed that most of the users in our collection reside in three main countries: France, Germany, and Italy. We experimented with a coordinated activity framework highlighting the presence of clusters of users promoting anti-vaccination content in a coordinated fashion.

There a few limitations to our work. First, the procedure used to identify Twitter conversations about COVID-19 vaccines involved a manual evaluation to determine relevant keywords, and thus it might be unable to fully exclude irrelevant data and/or conversations around vaccines which are not COVID-19 specific (e.g. animals, MMR, etc). Still, it allows for further filtering and refinement at a later stage.

Second, Twitter users might not be a representative sample of the population, and their online activity might not reflect the general public opinion (Wojick and Hughes 2020). Besides, according to the 2021 Reuters Digital News Re-

Figure 6: Networks of coordinated accounts that shared the same URLs at least 10 times on a daily basis, based respectively on French (top), German (center) and Italian (bottom) tweets.

port[16], Twitter was used respectively by 17% of the respondents in France, 6% in Germany and 8% in Italy for any purpose. As a matter of fact, Facebook remains the most used social media platform (Boberg et al. 2020) in most countries, but it does not allow to collect relevant data.

Third, users cannot opt-out from our collection, and this might raise important ethical concerns about anonymity. Nevertheless, whenever a user deletes a tweet or account, the related content will be unavailable in the re-hydration process.

There is a number of potential usages for this dataset. We aim to explore the correlation between the prevalence of online misinformation about vaccines (Pierri et al. 2021a) and public health outcomes (e.g. COVID-19 vaccine uptake rates, hospitalizations, etc) in different countries. We also plan to further investigate the presence of suspicious accounts, such as bots and trolls, and provide evidence of coordinated campaigns promoting anti-vaccine messages (Pacheco et al. 2021). Finally, we plan to build models to describe how online vaccine misinformation and anti-vaccine sentiment spread in different countries.

Acknowledgments

This work has been partially supported by the PRIN grant HOPE (FP6, Italian Ministry of Education), and the EU H2020 research and innovation programme, COVID-19 call, under grant agreement No. 101016233 "PERISCOPE" (https://periscopeproject.eu/). We are grateful to Lorenzo Corti, Andrea Tocchetti, Silvio Pavanetto, Pascal Garel, Moritz Laurer, and Anita Gottlob for helping in the selection of relevant keywords, gold hashtags and for manually annotating tweets. Newsguard labels correspond to ratings released in September 2021.

References

Banda, J. M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, E.; Tutubalina, E.; and Chowell, G. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3): 315–324.

Boberg, S.; Quandt, T.; Schatto-Eckrodt, T.; and Frischlich, L. 2020. Pandemic populism: Facebook pages of alternative news media and the corona crisis–A computational content analysis. *arXiv preprint arXiv:2004.02566*.

Bovet, A.; and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1): 7.

Brena, G.; Brambilla, M.; Ceri, S.; Di Giovanni, M.; Pierri, F.; and Ramponi, G. 2019. News sharing user behaviour on twitter: A comprehensive data collection of news articles and social interactions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 592–597.

Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American journal of public health*, 108(10): 1378–1384.

Burki, T. 2019. Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6): e258–e259.

Caldarelli, G.; De Nicola, R.; Petrocchi, M.; Pratelli, M.; and Saracco, F. 2021. Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. *EPJ data science*, 10(1): 34.

Chen, E.; Lerman, K.; Ferrara, E.; et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2): e19273.

Cinelli, M.; Morales, G. D. F.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9).

Cossard, A.; Morales, G. D. F.; Kalimeri, K.; Mejova, Y.; Paolotti, D.; and Starnini, M. 2020. Falling into the echo chamber: the Italian vaccination debate on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 130–140.

DeVerna, M.; Pierri, F.; Truong, B.; Bollenbacher, J.; Axelrod, D.; Loynes, N.; Torres-Lugo, C.; Yang, K.-C.; Menczer, F.; and Bryden, J. 2021. CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines. *Proceedings of the International AAAI Conference on Web and Social Media*.

Di Giovanni, M.; Corti, L.; Pavanetto, S.; Pierri, F.; Tocchetti, A.; and Brambilla, M. 2021. A Content-based Approach for the Analysis and Classification of Vaccine-related Stances on Twitter: the Italian Scenario. *Workshop Proceedings of the International AAAI Conference on Web and Social Media*.

Gallotti, R.; Valle, F.; Castaldo, N.; Sacco, P.; and De Domenico, M. 2020. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, 4: 1285–1293.

Gargiulo, F.; Cafiero, F.; Guille-Escuret, P.; Seror, V.; and Ward, J. 2020. Asymmetric participation of defenders and critics of vaccines to debates on French-speaking Twitter. *Scientific Reports*, 10.

Guarino, S.; Pierri, F.; Di Giovanni, M.; and Celestini, A. 2021. Information disorders during the COVID-19 infodemic: The case of Italian Facebook. *Online Social Networks and Media*, 22: 100124.

Hayawi, K.; Shahriar, S.; Serhani, M. A.; Taleb, I.; and Mathew, S. S. 2021. ANTi-Vax: A Novel Twitter Dataset for COVID-19 Vaccine Misinformation Detection. *Public Health*.

Imran, M.; Qazi, U.; and Ofli, F. 2021. TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. *arXiv preprint arXiv:2110.03664 - Forthcoming in Data*.

Johnson, N. F.; Velásquez, N.; Restrepo, N. J.; Leahy, R.; Gabriel, N.; El Oud, S.; Zheng, M.; Manrique, P.; Wuchty,

[16]https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

S.; and Lupu, Y. 2020. The online competition between pro- and anti-vaccination views. *Nature*, 1–4.

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, 5(3): 337–348.

Lopez, C. E.; and Gallemore, C. 2021. An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining*, 11(1): 1–14.

Mathieu, E.; Ritchie, H.; Ortiz-Ospina, E.; Roser, M.; Hasell, J.; Appel, C.; Giattino, C.; and Rodés-Guirao, L. 2021. A global database of COVID-19 vaccinations. *Nature human behaviour*, 1–7.

Mejova, Y.; and Kourtellis, N. 2021. YouTubing at Home: Media Sharing Behavior Change as Proxy for Mobility Around COVID-19 Lockdowns. In *13th ACM Web Science Conference 2021*, WebSci '21, 272–281. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383301.

Muric, G.; Wu, Y.; and Ferrara, E. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health Surveill*, 7(11): e30642.

Orenstein, W. A.; and Ahmed, R. 2017. Simply put: Vaccination saves lives. *Proceedings of the National Academy of Sciences*, 114(16): 4031–4033.

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 455–466.

Pierri, F. 2020. The diffusion of mainstream and disinformation news on Twitter: the case of Italy and France. *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion)*.

Pierri, F.; Artoni, A.; and Ceri, S. 2020. Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections. *PloS one*, 15(1): e0227821.

Pierri, F.; Perry, B.; DeVerna, M. R.; Yang, K.-C.; Flammini, A.; Menczer, F.; and Bryden, J. 2021a. The impact of online misinformation on US COVID-19 vaccinations. *arXiv preprint arXiv:2104.10635*.

Pierri, F.; Piccardi, C.; and Ceri, S. 2020. A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter. *EPJ Data Science*, 9(35).

Pierri, F.; Tocchetti, A.; Corti, L.; Giovanni, M.; Pavanetto, S.; Brambilla, M.; and Ceri, S. 2021b. VaccinItaly: monitoring Italian conversations around vaccines on Twitter and Facebook. *Workshop Proceedings of the International AAAI Conference on Web and Social Media*.

Righetti, N. 2020. Health Politicization and Misinformation on Twitter. A Study of the Italian Twittersphere from Before, During and After the Law on Mandatory Vaccinations. *OSF Preprints, doi:10.31219/osf.io/6r95n*.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9: 4787.

Wojick, S.; and Hughes, A. 2020. Sizing Up Twitter Users. Pew Research Center, https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ (accessed January, 2021).

Yang, K.-C.; Pierri, F.; Hui, P.-M.; Axelrod, D.; Torres-Lugo, C.; Bryden, J.; and Menczer, F. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*. Special issue "Studying the COVID-19 Infodemic at Scale".

Zarocostas, J. 2020. How to fight an infodemic. *The Lancet*, 395(10225): 676.