

# COVID Vaccine Sentiment Dashboard based on Twitter Data

Ferenc Béres<sup>1,2</sup>, Rita Csoma<sup>1,3</sup>, Tamás Michaletzky<sup>1,2</sup>, András Benczúr<sup>1,\*</sup>

<sup>1</sup>ELKH Institute for Computer Science and Control (SZTAKI), Budapest, Hungary

<sup>2</sup>Eötvös Loránd University, Budapest, Hungary

<sup>3</sup>Budapest University of Technology and Economics, Budapest, Hungary

Received: 11 October 2021; Accepted: 22 December 2021

## Summary

We developed an interactive dashboard that collects Twitter information relevant to COVID vaccines and analyzes their sentiment based on time, geolocation and type of the information source. Vaccine skepticism is a controversial topic with a long history that became more important than ever with the Covid-19 pandemic. Only a year after the first international cases were registered, multiple vaccines were developed and passed clinical testing. Besides the challenges of development, testing and logistics, another factor in the fight against the pandemic are people who are hesitant to get vaccinated, or even state that they will refuse any vaccine offered to them. In the paper, we demonstrate the use of the dashboard to assess changes in sentiment towards vaccination and identify possible events that affect the public view.

**Keywords:** COVID vaccines, sentiment analysis, social media, interactive dashboard

## COVID vakcina szentiment dashboard Twitter adatok alapján

Béres Ferenc<sup>1,2</sup>, Csoma Rita<sup>1,3</sup>, Michaletzky Tamás<sup>1,2</sup>, Benczúr András<sup>1,\*</sup>

<sup>1</sup>ELKH Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI), Budapest, Magyarország

<sup>2</sup>Eötvös Loránd Tudományegyetem, Budapest, Magyarország

<sup>3</sup>Budapesti Műszaki és Gazdaságtudományi Egyetem, Budapest, Magyarország

## Összefoglalás

Kidolgoztunk egy interaktív dashboard alkalmazást, amely összegyűjti a COVID vakcinákkal kapcsolatos Twitter-kommunikációt, és elemzi a vakcinákkal kapcsolatos attitűd időbeli változását, a földrajzi hely és az információforrás típusa alapján. A vakcina-szkepticizmus régóta megosztó téma. Az oltások népszerűsítése, az oltásellenes hangok hatásának csökkentése minden eddiginél fontosabbá vált a COVID-19 világjárvánnyal. Alig egy évvel az első nemzetközi esetek regisztrálása után több oltóanyagot fejlesztettek ki, amelyek klinikai teszteken mentek keresztül. A fejlesztés, a tesztelés és a logisztika kihívásai mellett a járvány elleni küzdelem legfontosabb tényezője azon emberek meggyőzése lett, akik haboznak az oltás felvételével kapcsolatban, vagy akár kijelentik, hogy megtagadják a számukra felajánlott vakcinákat. A cikkben bemutatjuk a közösségimédia-elemzés használatát az oltással kapcsolatos érzések változásának felmérésére és a nyilvánosságot befolyásoló lehetséges események azonosítására.

2021. január 24. és július 31. között a Twitter publikus interfészén elérhető adatokat gyűjtöttünk a „vaccine”, „vaccination”, „vaccinated”, „vaxxer”, „vaxxers”, „#CovidVaccine”, „covid denier”, „pfizer”, „moderna”, „astra” és „zeneca”, „sinopharm”, „szputnyik” kulcsszavak használatával, néhány negatív szűrő mellett, hogy csökkentjük a témához nem illő tartalmak mennyiségét. A közvélemény felmérésének fő technikai eszköze a hangulatelemzés volt, amelyet egy nyílt forráskódú eszköztárral végeztünk, amely hat nyelven előre betanított modelleket tartalmazott. A tartalmakat földrajzi hely és a Twitter-fiók típusa alapján is megkülönböztettük.

A hangulatelemzés során egy adott szöveg szerzőjének véleményét természetes nyelvet feldolgozó eszközök segítségével a negatívól a pozitív véleményig terjedő hangulatpontszámmal értékeltük.

Összességében a Modernával kapcsolatban találtuk a legpozitívabb, a Sinopharmmal a legnegatívabb véleményeket, bár ezek között nagy a földrajzi különbség. Például Európa a legnegatívabb az AstraZenecával és az (angol

nyelvű) Ázsia a Sinopharmmal szemben. Az orvosszakértők véleménye a legpozitívabb, a nem a fősodorba tartozó médiaszerzők pedig a legnegatívabbak az összes vakcinával kapcsolatban. A különböző vakcinák tevékenységének földrajzi megoszlása szorosan követi a vakcinák megoszlását, például a keleti vakcinák esetében több a spanyol nyelvű és ázsiai tartalom.

Eszközünket az AstraZeneca és a Pfizer-BioNTech vakcinákhoz kapcsolódó események követésével is bemutattuk, a kommunikáció mennyisége és hangulata alapján. Sikertült azonosítani azokat az eseményeket, amelyek az üzenetek számának csúcspontját vagy a hangulatváltozást okozták.

**Kulcsszavak:** COVID oltás, szentiment analízis, közösségi média, interaktív dashboard

## 1. Introduction

Vaccine skepticism is a controversial topic with a long history that became more important than ever with the Covid-19 pandemic. Only a year after the first international cases were registered, multiple vaccines were developed and passed clinical testing. Besides the challenges of development, testing and logistics, another factor in the fight against the pandemic are people who are hesitant to get vaccinated, or even state that they will refuse any vaccine offered to them. There are two groups of people commonly referred to as

- a) pro-vaxxer, those who support vaccinating people
- b) vax-skeptic, those who question vaccine efficacy or the need for general vaccination against Covid-19.

It is very difficult to tell exactly how many people share each of these views. It is even more challenging to understand all the reasoning why vax-skeptic opinions are getting more popular.

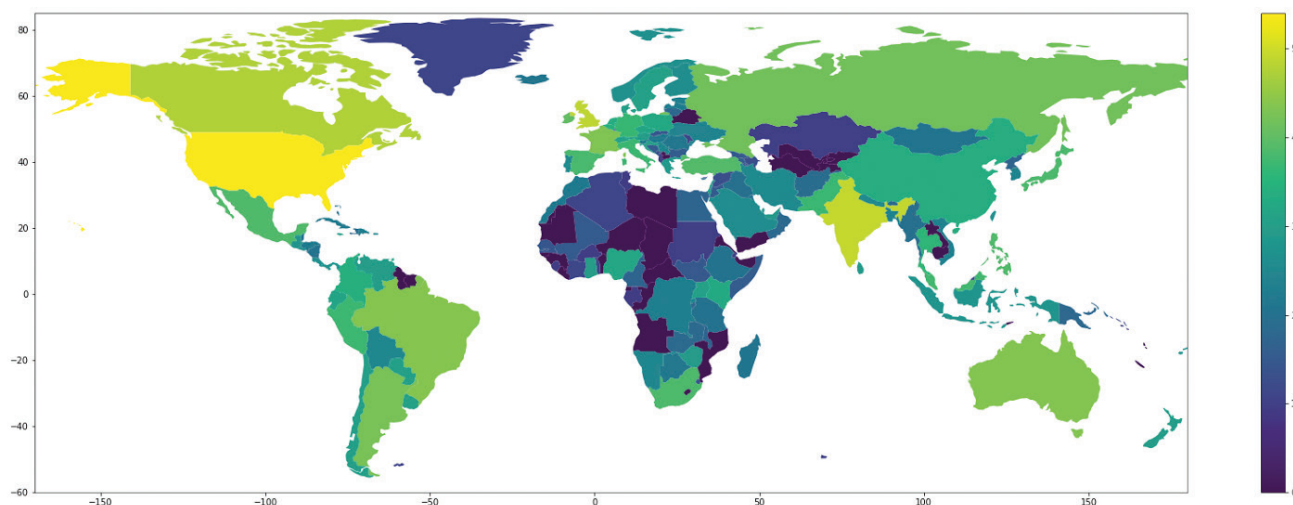
In this work, we monitor sentiment towards vaccines and assess pro-vaxxer and vax-skeptic content. After multiple data preprocessing steps, we used automated tools to assess the sentiment of communication, distinguishing vaccine types and geographic regions. We demonstrate how social media analysis can help identifying and explaining changes and differences towards different vaccines.

The prime goal of social media analysis towards vaccination is the detection and isolation of anti-vaxxer communities (Gaál et al. 2021; Mitra et al. 2016). By collecting social media content, one can assess the public opinion towards vaccination (Salathé-Shashank 2011) and even design communication to promote vaccination (Steffens et al. 2020). Very recently, experiments to identify Covid vaccine skeptic content (Ng-Carley 2021) and a data set (Muric et al. 2021) were also published. For the Central-Eastern-Europe region, where Facebook is the most popular social network platform, similar results appeared using Facebook data (Klimiuk et al. 2021); however, Facebook has no public data access API and hence its availability is strongly limited for research. As another alternative platform, research using data from Reddit has also appeared (Melton et al. 2021).

## 2. Data and methodology

From 24 January to 31 July, we collected data that anyone can view on Twitter by using the free Twitter API. By using the keywords

“vaccine”, “vaccination”, “vaccinated”, “vaxxer”, “vaxxers”, “#CovidVaccine”, “covid denier”, “pfizer”, “moderna”, “astra” and “zeneca”, “sinopharm”, “sputnik”,



**Figure 1** | Geographic distribution of tweets in the world. Colors indicate the tweet count on a 10-base logarithmic scale  
Source: Geographic distribution of our Twitter collection

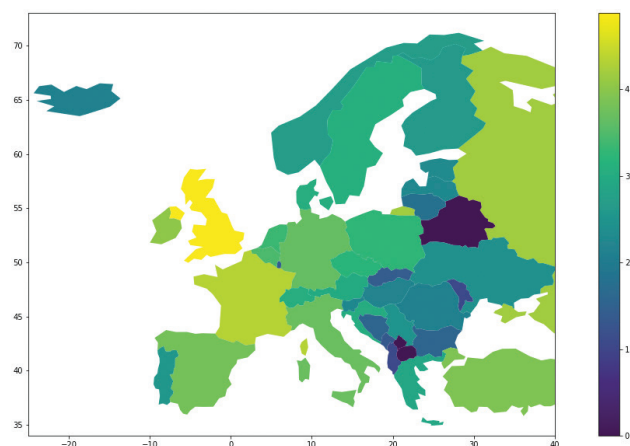
we collected 1,182,934 tweets, each with at least 50 likes. To eliminate drift towards topics of general politics such as US parties, we excluded the keywords “Trump”, “Biden”, “republican”, “democrat”. We further eliminated spam related to these keywords, especially for the search phrase “sputnik”. We dropped keywords with an insufficient amount of discussion in early 2021, including “johnson”, “novavax”, “covaxin”, “sanofi”, “can-sino”.

## 2.1 Geographic distribution

In *Figures 1* and *2*, we show the geographic distribution of the collected tweets, which align well with the general worldwide distribution of Twitter<sup>1</sup>, for example, low usage in Central Eastern Europe.

The extraction of the location from tweets is a noisy process. For each tweet, we first extracted the location string (e.g. “Washington, DC”, “London, UK”) from the posting user profile. If the location string is missing, we set the country based on the language of the tweet if it is spoken only in a single country (e.g. Italian). Then we assign geographic coordinates to the extracted cities and countries by sending queries to Wikipedia.

The geographical analysis in *Figures 1-2* is based on 59% of the collected tweets. We excluded the remaining tweets, since they have missing, invalid (e.g. Around the world, Mars, etc) or inconclusive (e.g. London-NYC-DC) location strings.



**Figure 2** | Geographic distribution of tweets in Europe. Colors indicate the base 10 logarithm of the tweet count.

Source: Our Twitter collection

<sup>1</sup> Number of active Twitter users in selected countries. <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries>. Visited 16.10.2021.; Covid-19 Twitter data geographic distribution. <https://data.humdata.org/dataset/covid-19-twitter-data-geographic-distribution>. Visited 16.10.2021.

## 2.2 Sentiment analysis

Sentiment analysis (or Opinion mining) is the task of finding the opinion of the author of a given text (*Feldman, 2013*). To do so, text is pre-processed using a variety of linguistic tools such as stemming, tokenization, part of speech tagging, entity extraction, and relation extraction. The main step is document analysis, which utilizes the linguistic resources to annotate the pre-processed documents with sentiment score ranging from negative to positive opinion.

In this work, we deployed a pre-trained multilingual sentiment analysis tool from Huggingface<sup>2</sup>, a leading NLP platform, without manually annotating any part of our data. Note that the accuracy of the sentiment detection can in general be improved by additional training steps on annotated data, but we had no resources for manual annotation. Huggingface was trained on product reviews in six different languages: English, Dutch, German, French, Spanish and Italian. In this work, we used this model to predict tweet sentiment on a scale from 1 to 5 stars. In *Table 1*, we show some examples to demonstrate how the selected model can differentiate between negative and positive content.

**Table 1** | Selected tweets from our collection that illustrate the 1-5 star sentiment rating

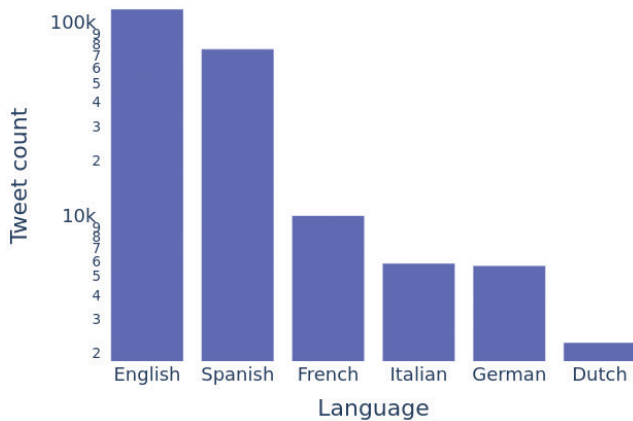
Label	Tweet example
1 star	Absolutely disgusting, Italian regional health authorities started running vaccination “open days” to get rid of unused astrazeneca doses to young people...
2 stars	Disappointing. I see no logical reason why someone who got two Pfizer jabs in London can travel freely to Amsterdam and back, but someone who got the same jabs in Amsterdam can’t travel to London and back...
3 stars	Okay, so I’ve had no side effects after the j&j vaccine and it’s almost been 24 hours
4 stars	I like this report much better than the study of the Pfizer vaccine in Israel that got lots of hype today. Why? This one examined efficacy for a longer period of time, and ...
5 stars	This is excellent. The US should be buying as many Pfizer, Moderna, and; j&j doses as they can produce and donating them to the rest of the world. It’s the right thing to do ...

Source: our Twitter collection

We used the following list of language processing tools:

1. Ekphrasis, a text processing tool geared towards text from social networks, such as Twitter or Facebook. It performs tokenization, word normalization, word segmentation (for splitting hashtags) and spelling correction: <https://github.com/cbaziotis/ekphrasis>
2. NLTK Python package is used for stemming: <https://www.nltk.org/howto/stem.html>

<sup>2</sup> The Huggingface multilingual sentiment analysis tool. <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>. Visited 16.10.2021.



**Figure 3** | The number of tweets collected for each language, on a logarithmic scale  
Source: Our measurement

- Language detection by Twitter: <https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages>
- Huggingface sentiment model: <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>

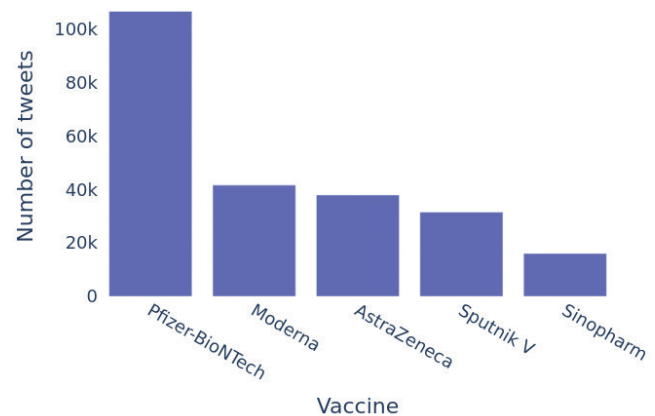
### 2.3 Data preprocessing

After we retrieved data through the Twitter API, we filtered the tweets so that they are associated with the selected vaccines and are written in languages for which the sentiment analysis module is available. After filtering, we obtained a collection of 221,720 tweets for our experiments.

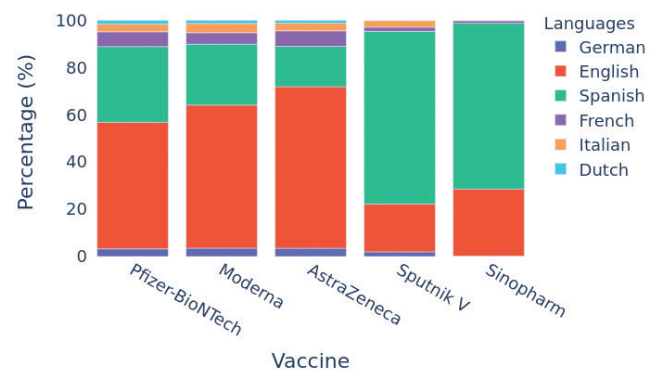
Due to language restrictions imposed by the selected sentiment model, we exclude tweets that were posted in other than the six supported languages. For example, our experiments do not include Portuguese, Turkish, Hindi, and Japanese tweets that were otherwise popular in our data collection. *Figure 3* shows that English and Spanish are the two dominant languages in the remaining data, while there are significantly less tweets for French, Italian, German, and Dutch.

In this work, we only analyze tweets that we could associate to the first five Covid-19 vaccines that were granted emergency authorization, see *Figure 4*. As Twitter is more popular in North America and Europe (see *Figure 1*), we managed to collect more tweets for western vaccines (Pfizer, Moderna, AstraZeneca) than for eastern vaccines (Sputnik, Sinopharm), which were mostly used in Asia and Latin-America. In *Figure 5*, we also observe that the fraction of Spanish content for eastern vaccines is significantly higher than for western vaccines. Both effects can be explained by the geographical distribution of Covid-19 vaccine supplies<sup>3</sup>.

<sup>3</sup> Covid vaccination tracker. <https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html>. Visited 16.10.2021.



**Figure 4** | The number of tweets collected for each vaccine  
Source: Our measurement



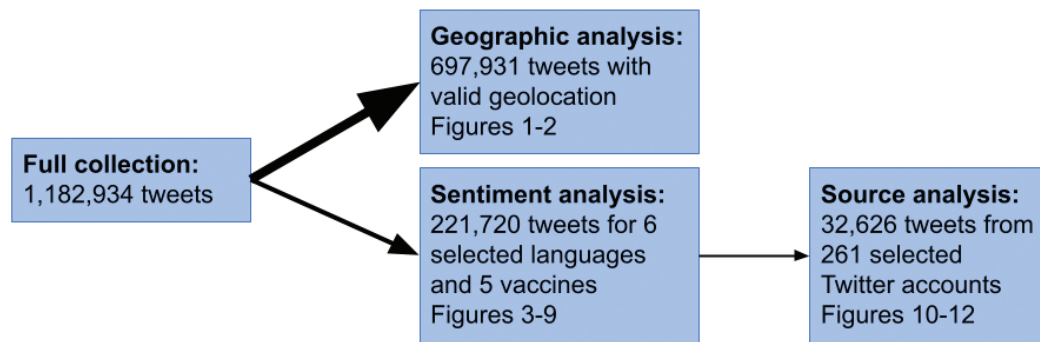
**Figure 5** | Language distribution of tweets for different vaccines  
Source: Our measurement

As summarized in *Figure 6*, we cleaned the full collection of over 1M tweets in three steps, for three different purposes. The geolocation analysis in Section 2.1 was based on the roughly 60% of the collection with valid location tag. Text analysis including language and sentiment considered less than 20% of the collection in the selected six languages. Finally, in Section 3, we will use roughly 3% of the collection from 261 selected and manually categorized source accounts.

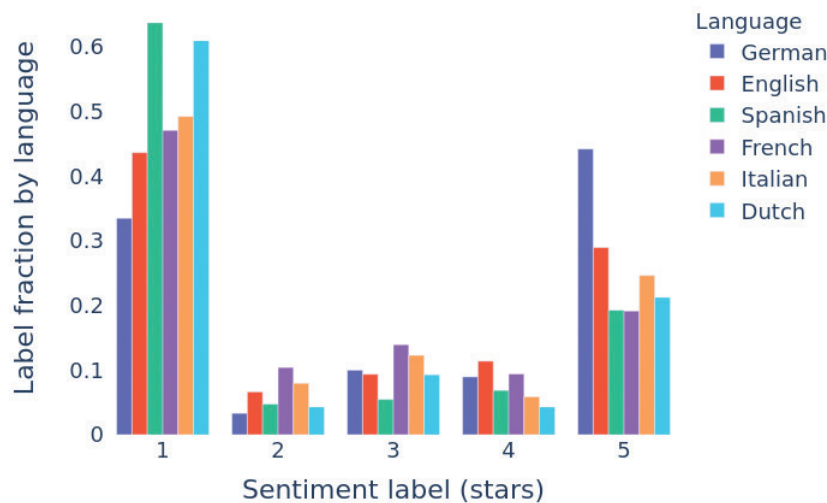
We also note that for 18,963 tweets (2.72% of those with explicit geolocation), we determined the location based on the language of the text. Part of these tweets could potentially come from minorities abroad, but due to the small size, we believe these tweets will not affect the conclusions.

### 2.4 Language normalization

After applying sentiment analysis to the six available languages, we observed that the average sentiment differs for different languages, see *Figure 7*. The predicted sentiment score for Spanish and Dutch content is biased towards negative, while German tweets have a positive



**Figure 6** | Major data filtering steps that we used throughout our work  
Source: Our measurement



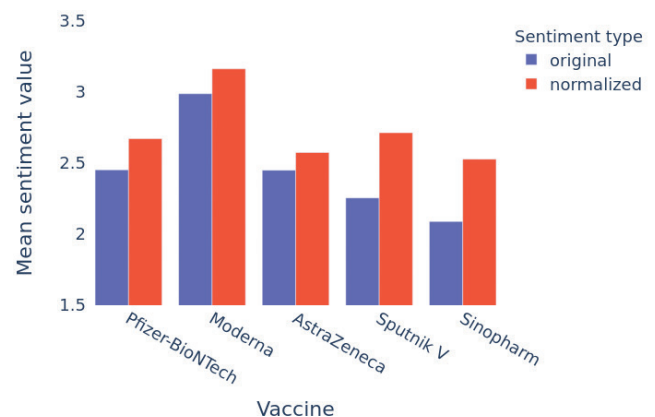
**Figure 7** | Sentiment label (1-5 star) distribution by languages  
Source: Our measurement

sentiment bias. This phenomenon can probably be partially explained by the difference in prediction accuracy: that much less Spanish (50k) and Dutch (80k) reviews were used to train the sentiment model<sup>4</sup> than English (150k). Another, more complex explanation could be the difference in the terminology used for products versus vaccines in the six languages. Since we consider the English sentiment as most accurate, and also the most popular language in our collection (Figure 3), we use English as a reference point for language sentiment normalization.

We normalized the sentiment score of languages other than English to have equal average sentiment over our collection. We modified the sentiment score by language dependent constants as follows: Spanish: +0.6219, Dutch: +0.5484, French: +0.3231, Italian: +0.2669, German: -0.5163.

With normalization, we achieved a more balanced average sentiment prediction across different vaccines, as

shown in Figure 8. It is interesting to see that the predicted average sentiment for Moderna-related tweets is significantly more positive than for other vaccines. On the other hand, Sinopharm was rated the least positive by Twitter users.



**Figure 8** | Original and language normalized mean vaccine sentiment  
Source: Our measurement

<sup>4</sup> The Huggingface multilingual sentiment analysis tool. <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>. Visited 16.10.2021.



### 3. Source analysis

Next, we analyze the amount of communication and the sentiment changes regarding different vaccines based on the types of the information sources. To do so, we selected the most active 261 accounts that had at least 20 tweets in the selected Twitter stream. We categorized the selected Twitter accounts by considering the meta-data provided by the account owners.

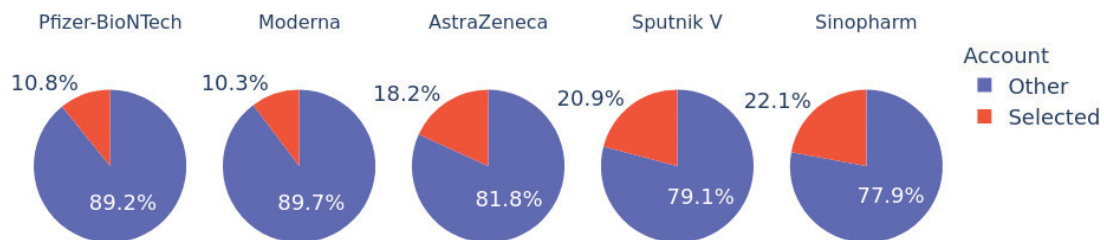
A large fraction of the selected 261 accounts correspond to mainstream media. Since online news with respect to different Covid-19 vaccines dominate the content, we further grouped mainstream media accounts by continent based on the geographic location of the publisher organization. For example, we assigned every BBC-related Twitter account to Europe. We note that Australian news is assigned to Asia while Africa is excluded from the experiments due to data scarcity.

Beyond mainstream media, we also find very active accounts with Medicine-related professions (e.g. doctor, MD, epidemiologist, microbiologist, etc.). Finally, the remaining active accounts all have many followers and share their own personal view. Here we find journalists,

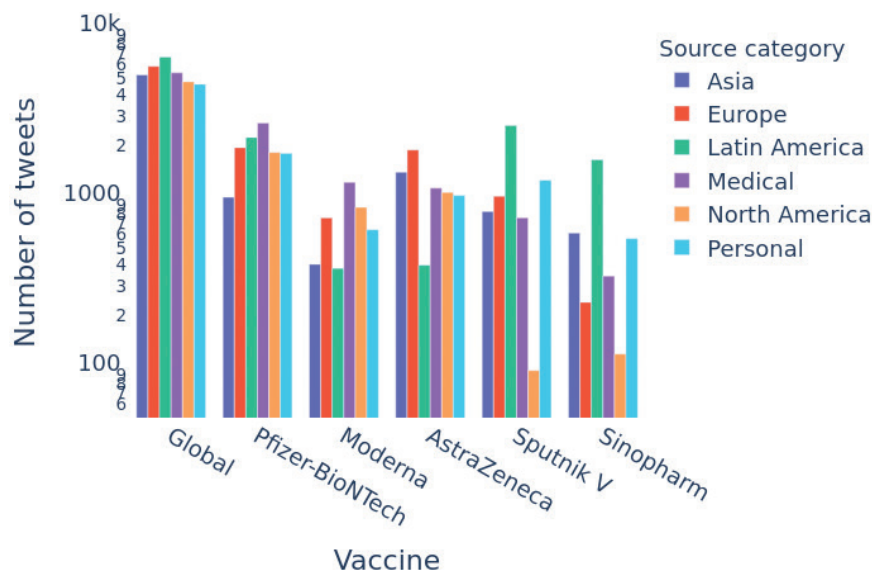
book authors, editors, and writers. We label this last category as authors.

In *Figure 9*, we show the fraction of tweets posted by the 261 selected accounts for each Covid-19 vaccine. It is interesting to see that Pfizer and Moderna acquired much more attention from the other, low activity accounts. This behavior can be explained by the fact that Pfizer and Moderna were two major vaccines in the Western Hemisphere where people were eager to share their personal experience related to Covid-19 vaccination. For example, how they feel or what kind of side effects they have after each dose. It seems that Sputnik V and Sinopharm received much less attention from the general public on Twitter. Finally, the reason for the high source ratio of AstraZeneca is related to the blood clot fear that was a controversial topic over several weeks in March and it was eagerly covered by mainstream media sources.

In the upcoming sections, we analyze 32,626 tweets posted by the 261 selected accounts. First we measure the activity grouped by source type and vaccine in *Figure 10*. We summarize some main findings of this figure:



**Figure 9** | The fraction of tweets posted by mainstream media, medical experts, and popular personal view related Twitter accounts  
Source: Our measurement



**Figure 10** | The number of tweets by different source categories and vaccines. In total, we analyze 32,626 tweets that were posted by the 261 accounts selected for the source analysis. The y axis is on logarithmic scale  
Source: Our measurement

- Latin America is slightly more active than the other categories, but the data is fairly balanced.
- Interestingly, the medical community was the most active category for Pfizer and Moderna. In our interpretation, these accounts were addressing a lot of misinformation and studies related to the efficiency of the two mRNA based Covid-19 vaccines.
- AstraZeneca appeared most actively in the European media, mostly related to news on the side effects. Note that the side effects were covered extensively only in Europe.
- Corresponding to their usage, the majority of eastern vaccine-related tweets are covered by Latin American media, while their activity is much less for Moderna and AstraZeneca.

### 3.1 Data dynamics: AstraZeneca case study

In our first case study, we analyze the number of AstraZeneca (AZ) related tweets over time. *Figure 11* highlights the events that induced increased activity on Twitter for three different source account categories. We rigorously assessed each event referenced by letters in the figure. The main events explaining the increased activity in the given periods are the following:

- a.) **February 8:** South Africa (SA) stops administering AZ after they found it is less effective for the SA variant.
- b.) **March 5–6:** EU and Italy block AZ shipment destined for Australia on March 5. One day later, Australia asked the EU to review its decision. Then, the EU tries to access AZ produced in the US.
- c.) **March 15–20:** Temporary AZ suspension in several countries due to blood clot fear that causes an enormous spike in Twitter activity:

- **March 15:** Suspension in Germany, Spain, France, Italy, Netherlands, Indonesia
- **March 16:** Suspension in Sweden. The EU health regulator states that there is no indication that AZ causes clots. Many doctors, politicians criticize European countries.
- **March 17:** Statements by EU and UK on vaccine export and delivery status. Boris Johnson announces that he will receive AZ. Multiple statements on AZ benefits outweigh the possible risks.
- **March 18:** Germany, France, Italy resume AZ roll-out after EU and UK drug regulators rally behind AZ.
- **March 19:** Scientists say they found the link to rare blood clotting. Multiple EU prime ministers take or will take the AZ vaccine.

d.) **April 7–8:** Multiple health regulators (UK, Spain, Philippines, etc.) suspend or advise to suspend AZ under certain age limits that differ for each country (ages: 60, 55, 50, 30).

e.) **April 28:** English study on Covid-19 transmission rate after taking up 2-dose from AZ.

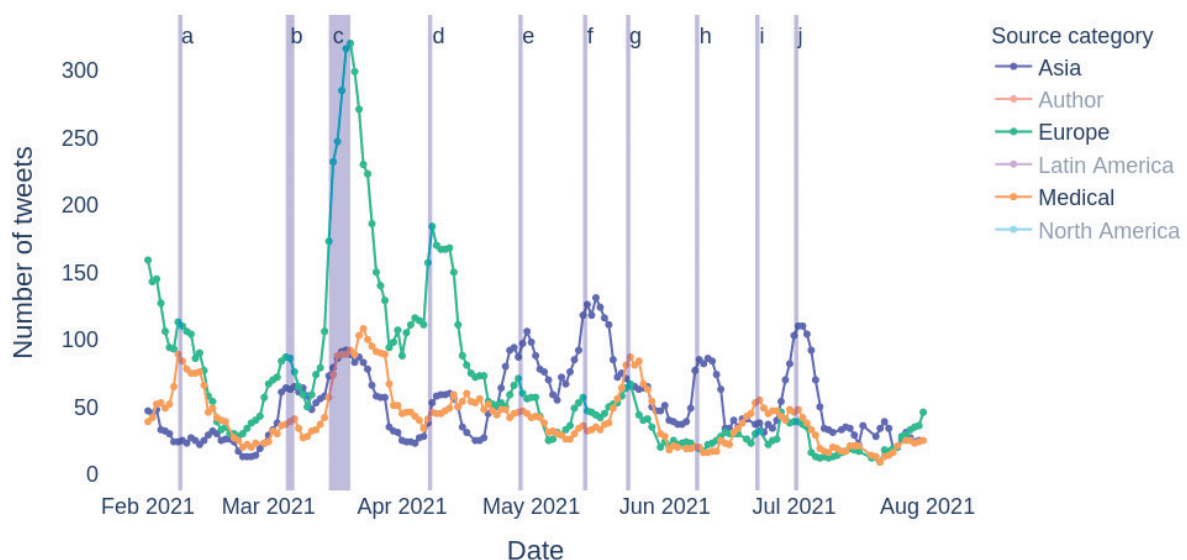
f.) **May 13:** Mixing Pfizer and AZ increased side effects. Italian study: 99% effectiveness against hospitalization after first AZ dose.

g.) **May 23:** Both Pfizer and AZ are effective against the Indian variant.

h.) **June 8:** Only Asia-related event: the government of India placed an order to buy AZ.

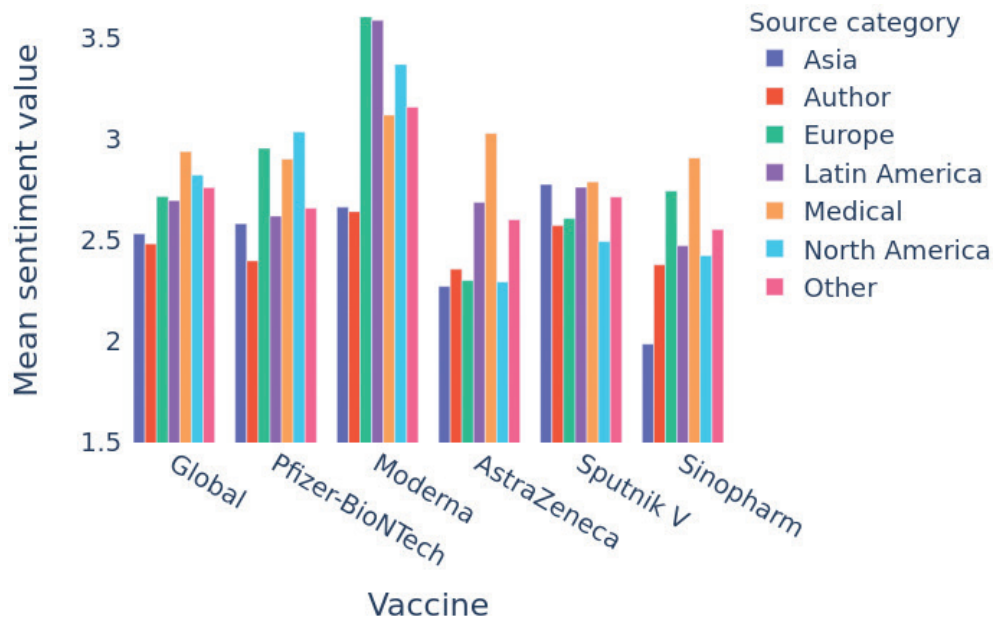
i.) **June 22:** Angela Merkel received Moderna second dose after AZ first dose.

j.) **July 1:** EU accepts an India-made version of AZ for traveling. Interestingly, this event was covered by mainstream media significantly more in Asia than in Europe.



**Figure 11** The number of AstraZeneca-related tweets over time within a 7-day rolling window. Major spikes corresponding to the narrative in Section 3.1 are highlighted by blue rectangles

Source: Our measurement



**Figure 12** | Mean sentiment by different source categories and vaccines  
Source: Our measurement

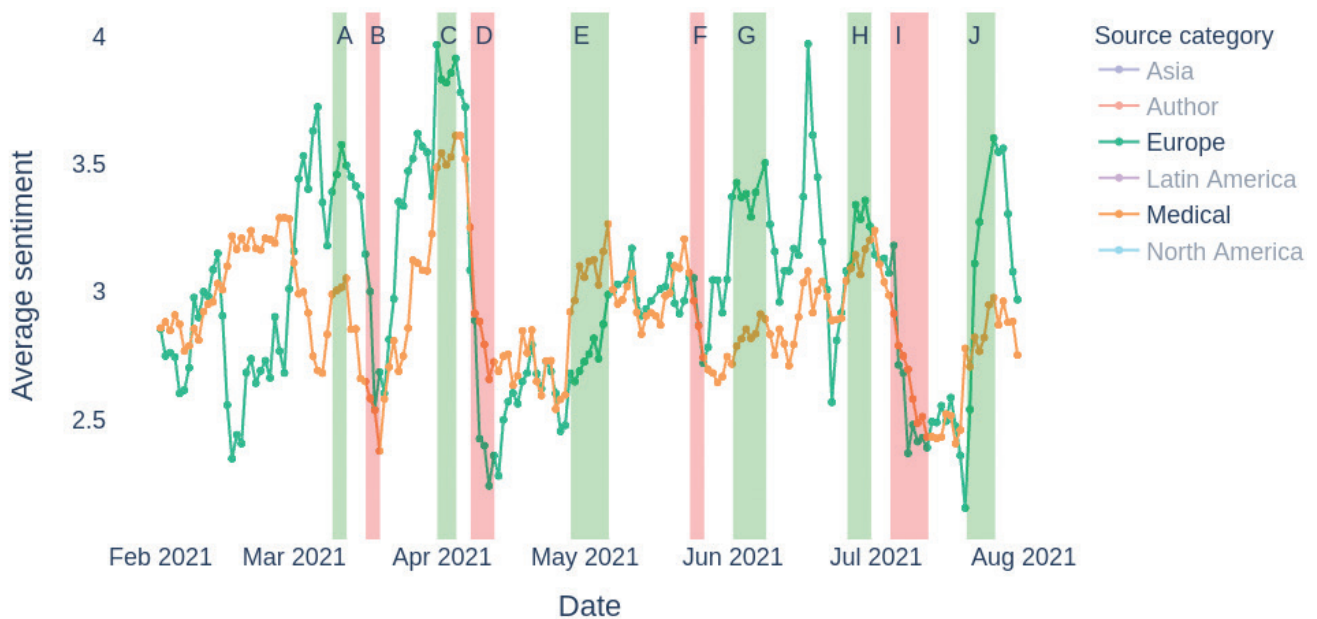
### 3.2 General vaccine sentiment

In this section, we continue by comparing the general sentiment between different source categories.

In *Figure 12*, we show the average sentiment for multiple vaccine-source pairs. In general, author accounts have the least positive opinion aggregated for every Covid-19 vaccine, see the 'global' column. On the other hand, medical accounts address the most positive tone.

Furthermore, the medical sentiment is well balanced across different vaccines, it is close to 3.0 in every case. It is also interesting to see that our analysis even managed to capture concerns related to Sputnik-V within the medical community, as this vaccine was the least positively rated by the sentiment model for medical accounts.

Despite the positive medical viewpoint on AstraZeneca, the general opinion for various mainstream media sources was quite negative. For example, for Europe, North



**Figure 13** | Average Pfizer vaccine sentiment within a 7-day rolling window. Major gains and losses in sentiment are highlighted by green and red rectangles, respectively  
Source: Our measurement



America, and author accounts it was the most negatively rated vaccine in our analysis. On the other hand, Moderna achieved the most positive sentiment for almost every category.

Interestingly, in Asia, the sentiment for Sputnik-V is as high as the Pfizer-BioNTech and Moderna, while AstraZeneca and Sinopharm receive much more negative sentiment than in other regions.

### 3.3 Sentiment dynamics: Pfizer case study

In our second case study, we assess the dynamic nature of vaccine sentiment on Twitter. In *Figure 13*, we show the average sentiment for the Pfizer vaccine within a 7-day rolling window. We rigorously assess each event referenced by capital letters in the figure. Using our interactive dashboard, we managed to identify the major events driving vaccine sentiment changes:

**a.) March 9–12:** Several studies were published on how Pfizer neutralizes the South African and Brazilian strain, as well as its effectiveness against asymptomatic infection.

**b.) March 16–19:** Significant sentiment drop due to vaccine export ban in the US combined with Pfizer and AstraZeneca supply problems between Europe and UK. Leaked information on political pressure on FDA and EMA related to Pfizer early authorization is also shared on Twitter.

**c.) March 31–April 4:** Very high sentiment values related to three major studies: 100% efficacy for teenagers (ages: 12–15); Pfizer works against the South African variant, protection up to 6 months after the second dose.

**d.) April 7–12:** The most significant drop in sentiment due to research on how the South African variant can break through Pfizer.

**e.) April 28–May 06:** General sentiment rebounds from low values due to various positive news: 94% effective against covid-19 hospitalization (ages: from 65); strong protection against multiple variants of concern; Canada is first to allow Pfizer for children (ages: 12–15); Pfizer starts to ship smaller packs of vaccine to reach more people; it donates vaccines to Olympic athletes; its quarterly sales exceeded expectations.

**f.) May 23–26:** A scandal related to Russia offering money for influencers to discourage their followers to receive Pfizer and Hong Kong's soon-to-expire vaccines cause a swift sentiment drop.

**g.) June 01–08:** Two major positive news: mRNA technology is being tested for cancer treatment; Pfizer starts clinical tests for children below 12.

**h.) June 25–30:** Encouraging research results: Pfizer and Moderna protection may persist for years; mixing Pfizer and AstraZeneca gives strong protection.

**i.) July 9–14:** Large drop in sentiment: 3rd booster dose is needed to maintain efficacy after 6–12 months; Israeli study reveals that Pfizer is less effective against delta variant than first thought (9X% -> 64%)

**j.) Jul 22–27:** Regaining positive sentiment due to multiple studies: 2-dose effectiveness against Delta-variant; optimal interval of 8–10 weeks between two Pfizer jabs. Finally, Pfizer and Moderna expands their studies for young children (ages: below 12)

We conclude the Pfizer case study by stating that the out-of-the-box multilingual sentiment model that we deployed for Twitter data proved to be very efficient in discovering vaccine sentiment shifts despite it being trained on an entirely different text domain (product reviews).

## 4. Conclusions

In this work, we demonstrated the use of a social media dashboard to analyze how events of vaccine testing, availability, side effects and their media coverage affect the public view on COVID vaccination. We collected tweets by using vaccine names as keywords, along some negative filters to reduce the amount of off-topic content. We selected tweets with high engagement and active users, along with discussion threads.

The main technical tool for assessing the public view was sentiment analysis, which we performed by an open source toolkit that had pre-trained models in six languages. We also distinguished content based on geolocation and Twitter account type.

Overall, we found most positive sentiment for Moderna and most negative for Sinopharm, although there is a high geographic difference in opinions. For example, Europe is most negative towards AstraZeneca and (English language) Asia for Sinopharm. The sentiment of medical experts are most positive and non-mainstream media authors the most negative in general for all vaccines. The geographic distribution of activity regarding different vaccines closely follows the distribution of the vaccines, for example more Spanish language and Asian content for eastern vaccines.

We also showcased our tool by following events corresponding to AstraZeneca and Pfizer-BioNTech vaccines, based on the amount and the sentiment of the communication, respectively. We were able to identify events that caused peaks in the number of messages or changes in sentiment.

In an ongoing future work, we evaluate Twitter content and user interaction network classification by combining text classifiers with several open source node embedding and community detection models.

## Acknowledgements

*The research was supported by the Ministry of Innovation and Technology NRD Office within the framework of the Hungarian Artificial Intelligence National Laboratory Program.*

## References

- Feldman, R. (2013) Techniques and applications for sentiment analysis. *Communications of the ACM*, Vol. 56. No. 4. pp. 82–89.
- Gaál, P., Joó, T., Palicz, T., Pollner, P., Schiszler, I., & Szócska, M. (2021) Adattudományi innováció az egészségügy környezeti kihívásainak kezelésében: a nagy adatállományok hasznosításának jelentősége és lehetőségei a járványkezelésben. *Scientia et Securitas*, Vol. 2. No. 1. pp. 2–11.
- Klimiuk, K., Czoska, A., Biernacka, K., & Balwicki, Ł. (2021) Vaccine misinformation on social media—topic-based content and sentiment analysis of Polish vaccine-deniers’ comments on Facebook. *Human Vaccines & Immunotherapeutics*, Vol. 17. No. 7. pp. 2026–2035.
- Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021) Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, Vol. 14. No. 10. pp. 1505–1512.
- Mitra, T., Counts, S., & Pennebaker, J. W. (2016) Understanding anti-vaccination attitudes in social media. *Tenth International AAAI Conference on Web and Social Media*, Vol. 10. No. 1. pp. 269–278.
- Muric, G., Wu, Y., & Ferrara, E. (2021) COVID-19 vaccine hesitancy on social media: Building a public twitter dataset of anti-vaccine content, vaccine misinformation and conspiracies. *arXiv preprint arXiv:2105.05134*.
- Ng, L. H. X., & Carley, K. (2021) Flipping stance: Social influence on bot’s and non bot’s COVID vaccine stance. *arXiv preprint arXiv:2106.11076*.
- Salathé, M., & Khandelwal, S. (2011) Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Computational Biology*, Vol. 7. No. 10. e1002199.
- Steffens, M. S., Dunn, A. G., Leask, J., & Wiley, K. E. (2020) Using social media for vaccination promotion: Practices and challenges. *Digital Health*, Vol. 6. DOI: <https://doi.org/10.1177/2055207620970785>

**Open Access statement.** This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated.