Co-funded by
the Health Programme
of the European Union

Grant Agreement No.: Start Date: End Date:	801495 01/08/2018 31/03/2022
Project title	**European Joint Action on Vaccination — EU-JAV**
WP number **Deliverable number** **Title**	WP8 D8.4 Design and development of visualisation tools. A vaccine confidence monitoring platform.
Responsible partner No. **Organisation** **Name** **E-mail address**	9 Istituto Superiore di Sanità, Italy Antonietta Filia antonietta.filia@iss.it
Nature R-report O-other (describe)	R
Dissemination Level **PU**-public **CO**-only for consortium members	PU
Delivery Month Planned	M44
Actual Delivery Date (dd/mm/yyyy)	31/03/2022

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Table of Contents

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Co-funded by
the Health Programme
of the European Union

Working Group

The activities of Task 8.3 were conducted by:

Francesco Gesualdo, Caterina Rizzo
Ospedale Pediatrico Bambin Gesù
Rome, Italy

Antonietta Filia (Task leader), Maria Cristina Rota
Department of Infectious Diseases
Istituto Superiore di Sanità, Rome, Italy

Acknowledgments

We would like to thank all EU-JAV partners who contributed to the activities of this Task, including participation in the surveys, validation of key words in English, Spanish and French, and validation of the algorithm for the stance analysis. We also thank Olivier Epaulard, EU-JAV coordinator, for useful discussions and comments on the report.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Co-funded by
the Health Programme
of the European Union

Executive Summary

In the context of Work Package 8 of the EU-Joint Action on Vaccination, the Italian National Health Institute and the Bambino Gesù Children's Hospital IRCCS have developed a social media and web monitoring platform (accessible at www.opbg.cloud) to monitor population sentiment and opinions towards vaccines and vaccination, and to identify the most influential on-line players on vaccine-related topics. This report describes the methodology used for the development of the platform, and the platform structure.

As preliminary work, we conducted two very brief surveys among EU-JAV partners to understand countries' experiences in monitoring the web for information on vaccines and their expectations and desires regarding social media monitoring. All the Member States envisaged the creation of a European platform that should contain information both at the European level and at the single country level.

The platform is composed of a data aggregator, a machine learning classifier and a data visualiser. The design of the platform followed a structured process, which included benchmarking of available products on the market, selection of data sources, creation of keyword filters based on a structured framework, reiterated processes of data cleaning, selection of the best way to view data, creation of a hashtags-based event detection system, creation of a community analysis system and influencer detection, training of a machine learning model for automatic classification of the stance towards vaccines expressed in the downloaded tweets (promotional/discouraging/neutral).

The goal of the platform is to inform health-care professionals, public health authorities, and policy makers about users' interest and opinions towards vaccines and the occurrence of vaccine hesitancy that could have a negative impact on vaccination uptake. It could also inform public health authorities in a timely manner of significant vaccine-related events especially those for which a rapid response is important, for example, to help decrease the spread of false information.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Background

A large part of the conversation on vaccines takes place on social media. Since 2015, the volume of articles published in the scientific literature on social media and vaccines has grown exponentially. As extensively described in the European Center for Disease Prevention and Control (ECDC) 2020 report "Systematic scoping review on social media monitoring methods and interventions related to vaccine hesitancy"[1], in recent years many studies have been published on the use of social media as a source of information on vaccines, on research based on social media monitoring projects, and reporting descriptions of interventions for vaccine promotion delivered through social media. A key recommendation of the ECDC report is that public health authorities use social media monitoring as a tool to detect changes in vaccine sentiment and to promptly respond to public health concerns.

In the context of WP8 Task 8.3 of the EU-Joint Action on Vaccination (JAV), the Italian National Health Institute (Istituto Superiore di Sanità - ISS) and the Bambino Gesù Children's Hospital (Rome, Italy) developed a social media and web monitoring platform to monitor population sentiment and opinions towards vaccines and vaccination, and to identify the most influential on-line players on vaccine-related topics. The main purpose of the platform was to create an in-house dashboard, based on a transparent methodology and on reliable data sources, to address vaccine hesitancy on the web and on social media.

Objective of the deliverable

To describe the methodology used for the development of a platform dedicated to monitoring the web and social media on vaccines and vaccine hesitancy.

Methodology used to develop the platform

Surveys among EU-JAV members

As preliminary work, we conducted two brief surveys among EU-JAV partners to understand countries' experiences in monitoring the web for information on vaccines, and their expectations and desires regarding the social media monitoring platform.

The first survey (Figure 1: Preliminary Workshop Questions), circulated in the pre-kick-off phase of the project in September 2018, mainly investigated JAV partners' previous experiences in web monitoring, media monitoring, digital communication and health literacy interventions. Moreover, it investigated the existence of web monitoring platforms dedicated to vaccines in EU/EEA Member States (MS).

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

The second survey (Figure 2: Questionnaire Task 8.3), was conducted after the kick-off of the project, in October 2018, and was aimed at gathering more technical information from MS, in order to define main features of the monitoring platform, such as accessibility, language and whether it had to be European or country-based.

The first survey showed that some media monitoring processes were already in place in different countries and highlighted the need for instruments to design a more universal and effective way of communicating with the target population.

France monitored vaccine confidence through annual surveys carried out by phone, with questions ranging from confidence towards vaccines in general down to specific vaccines. Results are published each year (http://invs.santepubliquefrance.fr/).

Slovenia had a media monitoring platform in place in Slovenian language (http://www.nijz.si/).

Slovakia involved public health specialists by email communication, through authorities and authorised web sites.

In the **Netherlands**, the National Institute for Public Health and the Environment (RIVM)'s communication department monitored different sources of information, such as Facebook, Twitter, online news platforms and other blogs, with the specific aim of addressing misperceptions, also using automatic systems for detecting posts with a negative vaccine sentiment. A private company provided them with tools and data. During the measles outbreak in 2013/2014, vaccination sentiment was monitored through manually coding messages' topics. Moreover, RIVM studied the correlation between the volume of measles related social media posts and the number of reported measles cases.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Figure 1. *Preliminary Workshop Questions*

1. Expectations of the outcome and realisation of WP8 and previous experiences in monitoring real time public opinions.

In your country, is there a system/platform in place for web monitoring of vaccine confidence or sentiment towards vaccines? Yes □ No □

- If yes, please indicate if it is open source and, if possible, include a link to the platform

2. Division of labour/possibilities to participate in Tasks 8.1, 8.2. and 8.3.

* 8.3: -Selection and validation of vaccine-related keywords (creation of an ontology)

Task 8.3:

Do you have any experience regarding vaccine communication or communication about any other health issue, in terms of:

- use of web services (e.g. Twitter, Health map, Google, Semrush) to obtain information about public opinion on health topics? **Yes □ No □**

  If yes, which services did you use and which health topics did you investigate?
- media monitoring? **Yes □ No□**
- digital communication? **Yes □ No □**
- health literacy initiatives for patients? **Yes □ No□**

Figure 2. *Questionnaire Task 8.3*

Survey 2:

Do you think that the system described in the workshop should be set up as:

a European platform Yes □ No □

a country-based platform? Yes □ No□

Would you be willing to have a monitoring system like the one described during the workshop set up in your country? Yes □ No □

Do you think that the platform should be open to the public or accessible only with credentials?

Open to the public Yes □ No □

Accessible only with credentials Yes □ No □

Is there any activity in which you would you like to be involved?

selection and validation of the vaccine-related keywords Yes □ No □

design of the web monitoring platform (i.e. features to be included, data to be visualized, etc.) Yes □ No □

training the natural language processing system (eg. classifying contents in terms of sentiment) Yes □ No □

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

All the Member States envisaged the creation of a European platform that should contain information both at the European level and at the single country level. This could help in being up to date with vaccine trends across Europe and react timely to the spread of fake news.

The platform should encourage collaboration between MS and timely reactions against misperceptions, and should facilitate the analyses to identify groups of fake news spreaders. Member States suggested that the platform should be open-access, while Slovakia would prefer access with credentials. Spain believed credentials might be necessary only in the preliminary stage. The Netherlands suggested that the platform should be European but also country and language-specific, and provide access to social media data.

Data sources

In November 2019, we accurately screened potential web and social media sources of data for the platform, considering data availability, cost, availability of an API (Application Programmer's Interface, an interface that allows for automatic data retrieval), and issues related to privacy and data confidentiality. We finally selected four data sources: Twitter, Reddit, Google Trends and Wikipedia.

Here follows a short description of the four sources.

Twitter

Twitter is a social network and microblogging public service. Currently, it is one of the most used social networks in the world, with 330 million active users per month, 145 million of which use the platform daily[2]. Six users out of ten are aged from 35 to 65 years; in Italy, the mean age of Twitter users is 32 years old. In March 2020, Italian Twitter users were 12.8 million, an increase of 24.4% from March 2019. France counts 16.9 million monthly users and 4.4 million unique daily visitors. In Spain, 7.5 million people have a Twitter account.

Twitter users interact by posting messages that are referred to as "tweets". Persons who are not registered on Twitter can read tweets but registration is required for any interaction to take place. Users create networks by following other users. Unlike other social media (e.g. Facebook), Twitter allows relationships between users even in one direction only (i.e. user A can follow user B even if user B does not follow user A). Users can type a tweet and publish it on their profile page; other users can read the tweet, like it, reply, comment, and/or retweet it. Users can also subscribe to track specific topics. A tweet can contain text up to 280 characters, and can include pictures, videos, internet links, polls, and mentions/comments of other tweets.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Upon registration, users can choose a username and personalise their profile by adding personal information, such as a profile image, a geographical location and a short biography. Location usually indicates where the user lives or works, while the biography is a short description of the user and often includes their interests and hobbies. As a default, all this information is public, so any other Twitter user can look at personal information and read the tweets of another user, with the exception of those users that choose to have a private profile and share data only with their followers. Moreover, tweets can be read in a completely anonymous way, without providing credentials. Finally, unlike other social media platforms, tweets are downloadable and can be acquired with metadata.

Reddit

Reddit is a social news aggregation, forum, and discussion website that has recently included livestream content. It is mostly used in the United States but its diffusion is increasing in Europe too. In October 2020, Reddit ranked as the 17th most-visited website in the world. It is structured in subject-dependent boards or "subreddits" that cover different topics (politics, science, music, cooking, image sharing, etc.). Users or registered members can publish content, such as images, text, links and videos. Other users can express a like or dislike on these contents, voting them "up" or "down". The more up-votes, the higher up the post is visualised in the subreddit. If the post reaches a high number of up-votes, it can be published on the website's front page, gaining a strong visibility. Articles and posts are also associated with "flairs", i.e. tags that describe article content. Reddit has a strong community orientation and aims at building conversations and discussions on specific articles posted by its users.

Google Trends

Google Trends is a Google-based tool that displays information on the relative frequency with which a word, a sentence or a topic is searched for on browsers and on the web.

Google Trends shows statistics about search queries on Google. When displaying results by "topic", the system automatically considers different terms and sentences related to the same concept, using automatic translation in different languages. Searches can be country and language specific. Trends are shown on a graph where the "popularity" of a keyword/sentence/topic over time is visualised. Search volumes are normalised by time and location, i.e. data are indexed to 100, where 100 is the maximum search interest for the selected keyword/sentence/topic, time period and location. Google Trends also offers a function that displays the most frequently used search queries related to the original search.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Wikipedia

The word wikipedia is a blend of the words "wiki" ("quick" in Hawaiian) and "encyclopedia". It is an online, multilingual, open-collaborative and free-content encyclopedia in which everyone can share knowledge. Contents are written by anonymous volunteers and are freely editable by users. It is one of the 15 most popular websites, as of August 2020[3]. Wikipedia contains more than 55 million articles and receives 1.7 billion unique visits per month. The *Wikipedia pageviews tool* shows the number of users that have visited a Wikipedia article during a given period of time.

Wikipedia pageviews of specific articles correlate with changes in stock market prices[4] or, more interestingly, with the spread of infectious diseases[5,6]. Since search engines have an impact on what is popular on Wikipedia, the volume of accesses to Wikipedia webpages can provide information on what people are searching for on the web, and can be used to monitor the public interest on certain topics[7].

Keyword Filters

Data collection is the first crucial step that needs to be carried out to track vaccine conversations on social networks. Extracting information that is both significant and specific from providers such as Twitter and Reddit can be challenging, due to the huge amount of data produced daily on these platforms. To filter these large amounts of information, interactions can be extracted which contain a specific set of keywords, which have to be manually identified and combined in filters.

In order to create consistent keyword filters in the three selected languages, we partially based our work on a structured framework previously published[8].

The first step was to gather a first set of keywords from Babelnet, a popular semantic network which connects concepts and named entities in a very large network of semantic relations (Babel synsets), made up of about 16 million entries. Each Babel synset represents a given meaning and contains all synonyms that express that particular meaning in a number of different languages. Through the BabelNet knowledge-graph, we generated a list of relevant keywords in different languages. We started from the node "vaccine" and subsequently extracted all the connected edges, which represent either a hypernym (i.e. a word indicating a broad category into which words with more specific meanings fall) or a hyponym (i.e. a word indicating a more specific category). For each selected node, we identified the specific synonyms in French, Spanish and Italian. Secondly, a group of experts, including EU-JAV partners, reviewed the identified keywords in order to select those unequivocally referring to vaccines. Thirdly, we started downloading tweets based on the selected keywords, and, after few months, we discarded those keywords that returned less than 80 entries per month. Moreover, we checked a sample of 200 downloaded tweets for each keyword for relevance, and we discarded the keywords that returned more

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

than 30% of tweets as being irrelevant. Along the months, with the COVID-19 pandemic, and, in particular, with the COVID-19 vaccination roll-out, we reviewed the keyword filters, adding newly identified vaccine-related keywords, e.g. those including the names of COVID-19 vaccine brands.

Here follows a list of the final filters obtained.

Italian: Vaccino OR Vaccini OR Vaccinazione OR Vaccinazioni OR Vaccinato OR Vaccinata OR Vaccinati OR Vaccinate OR Immunizzazione OR Immunizzato OR Immunizzata OR Immunizzati OR Immunizzate OR Novax

French: Vaccin OR Vaccins OR Vacciner OR Vacciné OR Vaccinée OR Immunisation OR Immuniser OR Immunisé OR Immunisée OR Novax

Spanish: Vacuna OR Vacunas OR Vacunación OR Vacunaciones OR Vacunada OR Vacunado OR Novax OR Antivacunas

Platform structure

Layers

The main feature of the online platform is the real-time monitoring of vaccine-related information flows from different social networks. The platform is available and ready-for-use 24/7 for EU-JAV members with credentials. A version with limited functionalities is available publicly.

The final dashboard is presented as a web page application composed of a back-end (where data are stored and elaborated) and a front-end (where data are presented) subsystems. The functions of the platform are to collect, analyse, aggregate and visualise data. These functions are achieved thanks to the platform's architecture, which is articulated in three different levels: the data layer, the analysis layer and the visualisation layer.

Data Layer

The Data layer collects and pre-processes raw data from the web, with the aim of storing clean information that is ready for the analyses.

Data collection consists in recording raw information from different platforms through the APIs, using pre-specified keyword filters (see above) and storing this information in a database management system (DBMS), namely MongoDB. To maintain an updated flow of information, the platform includes a script that is run every hour.

The pre-processing phase has the purpose of performing an initial cleaning of the data and of selecting consistent information, so that further analyses are easier to carry out. The information for every tweet is downloaded as a set of data, in a format named "JSON", including a mix of root attributes and child

objects. We selected only the attributes and objects consistent with the platform's purpose, including the following: tweet ID, creation date, text, username, hashtags, language, retweet status (and original tweet), reply status (i.e. if the tweet is a reply), URL. This allows to decrease the size of the record, still retaining all useful data. Moreover, we removed from the tweet's text all the non-alphabetic characters (e.g. emoticons).

Analysis Layer

The main purpose of the analysis layer is to analyse the pre-processed data, through the generation of graphs, which facilitate data interpretation, highlighting data structures and interconnections. On this layer, the following processes take place: tweet geolocalisation, influencer analysis and early signal detection (see specific sections).

Visualisation Layer

This final layer is dedicated to data visualisation. Information from the other layers is conveyed in a simple but representative manner. Results are aggregated in such a way that the final conclusion can be drawn as easily as possible. The user is able to customise data visualisations by country and by time frame.

Software Architecture Internals

To achieve a stable availability of the service, we set up a system for stratified backups, to minimise data losses. Also for the data layer, we implemented a "master-slave relationship" between the main database and backup database.

Tweet data downloaded through the API has a very different structure according to the type of tweet (tweet, retweet, reply, etc.).

For data storage, we used a "non-relational database" (NoSQL), namely MongoDB, mainly based on the following features:

- Each document (tweet) can have its own structure and not a fixed data structure, but a structure that is document dependent;
- Very flexible Query system;
- Possibility of managing great amounts of data.

All the architecture is based on the GNU-linux system.

For data analysis, we used the software Python, which offers a number of well-tested analytics libraries very consistent with our aims, covering numerical computing, data analysis, statistical analysis, visualisation and machine learning.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Data Collection

Twitter's Application Programming Interface (API) is a connectivity interface that enables users to download Twitter data. Twitter's API allows complex queries (e.g. downloading every tweet about a certain topic within the last twenty minutes, or downloading a specified user's non-retweeted tweets).

The Twitter API enables users to download, for each tweet, the following information:

- tweet text
- user
- retweet count (number of retweets)
- lang (language)
- id (a unique identifier for the tweet)
- created at (creation date of the tweet, as UTC-10)
- coordinates (longitude and latitude of the place where the tweet was posted from - available only for a limited number of tweets)
- in reply to user id (if the tweet is a reply, this field indicates the ID of the tweet it replies to)

All the activities available through Twitter APIs respect the Privacy Policy and the Terms of Service that every user has to sign at the moment of registration on the Twitter platform. Moreover, to gain access to the API, users and developers have to sign a Developer Agreement, in order to respect the platform's policy. For our purpose, we accessed a sample of Twitter data, using specific vaccine-related keyword filters (see specific section for details) and by specifying other features, such as geolocalisation (see specific section for details). The API provides only a sample corresponding to 1% of the total tweets published worldwide. This type of download is available through the API "Volume streams"[9].

Through the Pushshift API[10], we downloaded Reddit posts and comments selected through validated keyword filters, in the following subreddits: r/italy r/french r/spain r/england. On the platform, posts and comments with the highest score are visualised.

Through an API for Google Trends (https://github.com/GeneralMills/pytrends), we downloaded daily statistics for search queries concerning the vaccine topic. The platform displays the time-series for each retrieved keyword and a list of search queries relative to the vaccine topic in the selected timeframe.

Through the Wikipedia page views tool[11-12], we retrieved page views for vaccine-related articles, visualised as a daily time-series of page views for each article in selected languages.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Tweets Geolocation Inference

We included all tweets in Italian language as originating from Italy. Differently, for tweets in French and in Spanish, we had to set up a geo-localization process to exclude tweets in French not originating from France (e.g. originating from Algeria), and tweets in Spanish not originating from Spain (e.g. originating from Latin America).

We based the processes for tweets' geo-localization on a framework proposed by different authors[13].

Twitter gives users the possibility of enabling a geo-localization service on mobile devices. Nevertheless, only 1-3% of global tweets contain GPS coordinates (latitude and longitude). An alternative option to infer geo-location of a tweet is to use the *place field*, which is an attribute for each single tweet and can be filled by selecting one of the locations suggested by the platform. This place tag associates a tweet with a location referring to a city, an area, or a famous point-of-interest (e.g. a building or a restaurant). Another option for tweet geo-localization is the *user location*, provided in the user's profile. In this case, the indicated location might not be a valid location name. The network of followers and friends could also help determine the home location of a user. Alternatively, tweet content, i.e., the actual tweet message, can have one or more location mentions, which can be extracted for the location inference task.

For our platform's purposes, we decided to use the following tweet information for geo-localization: 1) geo-coordinates (latitude, longitude), 2) place field and 3) user location field.

The user location field is a free text field filled in by the user. Therefore, to exploit this information, we used the API Nominatim[14], which enables the user to identify a location's geo-coordinates from the location's name.

Event detection

The event detection feature aims at capturing the social or real-life event, which feeds the vaccine debate. Debates on social networks can be silent for long periods or explode as consequences of a specific event. Capturing early signals can enable us to understand what particular event causes the debate to revive.

For this purpose, our platform considers four different data sources: Twitter, Wikipedia, Google Trends and Reddit.

Data from the four data sources included in our platform (Twitter, Wikipedia, Google Trends and Reddit) are visualised in order to highlight trends and peaks of interest in vaccines. Twitter data are also exploited to identify the variation in hashtag volumes.

As for Twitter, through the Twitter API, our platform daily extracts tweets based on validated keyword filters. As a first step, the platform considers a quantitative analysis of the data, counting the number of tweets and retweets. Subsequently, the daily tweet count is composed in a time series, which is analysed. In particular, the interest lies in the peaks of the series. Thus, the evaluation of the median is calculated on the data. The monitoring platform considers the values over the median as an early signal for an event.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801493.

According to Xiao *et al.*[15], Twitter hashtags allow us to describe tweet contents and to cluster them into topics. For this reason, each day the monitoring platform extracts the 10 most common hashtags from this set of tweets, which can represent the topic and the contents of the tweets. Finally, the platform displays a tweets time-series, the median number of daily tweets and, for each day, the 10 most common hashtags.

As for Wikipedia, through the Wikipedia page views API, the platform extracts on a daily basis the page views of vaccine-related keywords corresponding to Wikipedia pages. This feature helps the user to understand if people are looking for information on vaccine-related topics. As for Twitter, also for Wikipedia, the page views time-series are shown on our platform.

As for Google Trends, the platform considers the time-series of vaccine-related keywords (as "vaccine topic") searched on the Google search engine.

Finally, the fourth source is Reddit. The monitoring platform filters for subreddit corresponding to the European country and then displays the articles' titles with higher scores, directly accessible through the platform.

For all data sources, our platform shows data by language and geo-localization for the three languages included in this pilot (Italian, French and Spanish).

Detection of influencers

In order to visualise the interactions between the users involved in the vaccine conversation on Twitter, the structure of a graph is analysed. Based on the kind of interaction between users that we aim at analyse, different kinds of graphs are available:

- **Retweet Graph**. Each user is considered as a node and an edge from user *u* to user *v* is added if *u* retweeted *v* at least once. This way, we are able to identify those users who retweet often (looking at the outlinks of the graph) and users whose tweets have been retweeted (looking at the inlinks of the graph). In other words, influential users, which are often retweeted, will have many links pointing to them.
- **Follower Graph**. Each user is considered as a node and an edge from user *u* to user *v* is added if *u* follows *v*. This graph enables us to detect influential users, since they typically have a large number of followers while they follow few users.
- **Mentions Graph**. The structure of this graph is very similar to the retweet graph, with the difference that an edge from user *u* to user *v* is added if *u* mentioned *v* at least once. Mentioning a user means that we are kind of replying to the mentioned user. Ideally, influential users should have many mentions since more users are interested in the content they share and would like to reply to them.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Based on the platform's objectives, we decided to base our influencer detection system on the **retweet graph**. Subsequently, we analysed the structure of the graph through a number of metrics called "centrality measures". The centrality measures are listed in the platform section dedicated to the influencer analysis. Here follows a detailed description of each centrality measure:

- **Degree centrality.** It measures, for each node, the number of edges incident upon it, i.e. the degree of node *v*. Given a graph G and a node *v* belonging to G, the degree centrality is defined as:

  C_d(v) = deg(v)

- **Betweenness centrality.** It represents the proportion of times a node can be considered as a "bridge" in the shortest path between a pair of nodes, over the total number of shortest paths between the same pair of nodes:

  C_b(v) = \sum_{s \div v \div t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}

  where\sigma_{s,t} is the total number of shortest paths from node *s* to node *t* and \sigma_{s,t}(v) is the number of those paths that intersect node v.

- **Closeness centrality.** It is defined as the reciprocal of the sum of the distances between *v* and all other nodes. Therefore, being close with respect to the graph means that the distance between *v* and all other nodes in the graph is minimal. The lower the closeness centrality, the lower the distance of *v* with respect to the graph.

  C_c(v) = \frac{1}{\sum{dist(v,t)}}

  In the "Closest" column, we reported the rankings of the closeness of a user with the other users of the graph. Closeness means that two users are directly connected through retweets (eg. user A retweets user B). Top ranking (=lower numbers) is assigned to users that are well positioned in the graph to **influence other users as fast as possible**.

- **Pagerank** is an algorithm to count the number and quality of links to a node, in order to estimate its importance. Ideally, the more important the node is, the more likely it is to receive links from other nodes, considering an edge from node *u* to *v* as a sign of agreement from *u* to *v*. The importance of a node is set according to both the number of incoming edges to the node and the importance of the corresponding source nodes. The algorithm works recursively and considers not only the metrics computed for the node *v* but also the metrics of all the nodes, which link to *v*. For this reason, through this algorithm it is also possible to detect nodes with higher popularity, meaning those nodes not only with a lot of links but also with important nodes pointing to them. In our specific case study, the output of the PageRank algorithm is a probability distribution that

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

represents the likelihood that a user randomly acting on Twitter (with retweets or clicking on mentioned users) will arrive at any particular user node. The higher this probability for a given node *v*, the higher the popularity or importance of that node *v*. In the "Pagerank" column, we reported the rankings of the weighted connections of a user with the other users. The weight refers to the retweeting activity between two users. Top ranking (=lower numbers) is assigned to a user who has the potential to influence users that are not directly connected with them.

Platform Access

In June 2020, the platform was made accessible with credentials to EU-JAV consortium members. An open version of the platform is accessible since July 2021 at the following link: www.opbg.cloud.

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.

Bibliography

1. European Centre for Disease Prevention and Control. *Systematic scoping review on social media monitoring methods and interventions relating to vaccine hesitancy.* (Publications Office, 2019).

2. 85 Twitter Statistics You Must Know: 2021/2022 Market Share Analysis & Data. Financesonline.com https://financesonline.com/twitter-statistics/ (2019).

3. Global top websites by monthly visits 2020. Statista https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/.

4. Quantifying Wikipedia Usage Patterns Before Stock Market Moves | Scientific Reports. https://www.nature.com/articles/srep01801.

5. Generous, N., Fairchild, G., Deshpande, A., Valle, S. Y. D. & Priedhorsky, R. Global Disease Monitoring and Forecasting with Wikipedia. PLOS Comput. Biol. 10, e1003892 (2014).

6. The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007633.

7. Tausczik Y, Faasse K, Pennebaker JW, Petrie KJ. Public anxiety and information seeking following the H1N1 outbreak: blogs, newspaper articles, and Wikipedia visits. Health Commun. 2012;27(2):179-85. doi: 10.1080/10410236.2011.571759. Epub 2011 Aug 9. PMID: 21827326.

8. Kim, Y., Huang, J. & Emery, S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. J. Med. Internet Res. 18, e41 (2016).

9. Volume streams introduction | Docs | Twitter Developer Platform. https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction.

10. PSAW: Python Pushshift.io API Wrapper (for comment/submission search) — PSAW 0.0.12 documentation. https://psaw.readthedocs.io/en/latest/.

11. Wikipedia: Pageview statistics. Wikipedia (2022).

12. REST API Documentation. https://wikimedia.org/api/rest_v1/.

13. Qazi U, Imran M, Ofli F. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. SIGSPATIAL Spec. 12, 6–15 (2020).

14. Reverse - Nominatim 4.0.1. https://nominatim.org/release-docs/latest/api/Reverse/.

15. Xiao, F., Noro, T. & Tokuda, T. Finding news-topic oriented influential twitter users based on topic related hashtag community detection. J. Web Eng. 13, 405–429 (2014).

EU-JAV has received co-funding from the European Union's Health Programme under Grant Agreement no. 801495.