Scientific Visualization, 2021, volume 13, number 4, pages 144 - 163, DOI: 10.26583/sv.13.4.11

Visual Analytics of Twitter and Social Media Dataflows: a Casestudy of COVID-19 Rumors

M.S. Ulizko[1,A,B], E.V. Antonov[2,A,B], M. A. Grigorieva[3,C], E.S. Tretyakov[4,B],
R.R. Tukumbetova[5,A,B], A.A. Artamonov[6,B]

[A] Plekhanov Russian University of Economics,
Stremyannyy Pereulok, 36, 115093 Moscow, Russia
[B] National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashira Hwy, 31, 115409 Moscow, Russia
[C] Lomonosov Moscow State University,
Leninskie Gory, 1, p.4, Moscow, 119991, Russian Federation

[1] ORCID: 0000-0003-2608-8330, mulizko@kaf65.ru
[2] ORCID: 0000-0003-1498-9131, eantonov@kaf65.ru
[3] ORCID: 0000-0002-8851-2187, maria.grigorieva@cern.ch
[4] ORCID: 0000-0002-1051-8562, etretyakov@kaf65.ru
[5] ORCID: 0000-0002-1976-1390, rrtukumbetova@kaf65.ru
[6] ORCID: 0000-0002-9140-5526, aartamonov@kaf65.ru

Abstract

One of the most significant and rapidly developing fields of data analysis is information flow management. In the course of the analysis targeted and stochastic dissemination patterns are studied. The solving of such problems is daunting due to the global growth of the amount of information and its availability for a wide range of users.

The paper presents a study of dissemination of information in open networks on the example of COVID-19. The study was conducted with the use of web scraping, methods of linguistic analysis and visual analytics. As sources of information variety of sources were used, such as the largest world and Russian information services, social networks and instant messengers. The paper considers statistical analysis of English media articles and posts form Twitter, dissemination of data flows between countries and information source. The developed methods can be scaled up to analyse information events of various topics.

Keywords: statistical analysis, graph analysis, geospatial analysis, named entity recognition, Web-technology, COVID 19, misinformation, twitter, mass media.

1. Introduction

In the digital world, Internet traffic is growing every year [1]. According to various projections, by the end of 2021 Internet traffic will exceed 3 zettabytes (ZB) [2, 3]. At the same time, communication in society also passes into the virtual world [4, 5]. The exchange of information via the Internet has significantly reduced the time for the delivery of information from the moment of the event to the moment it is received by the consumer / user. However, due to the large amount of data and its heterogeneity, such an environment has become a fertile ground for the dissemination of false information, and in some cases it is generated in larger volumes than true information. False information can form a false point of view, which can lead to the destabilization of society [6]. For instance, the coronavirus disease 2019 (COVID-19) pandemic shows that the spread of misinformation could result in poor physical and mental health outcomes among individuals [7]. The paper examines similar dataflows for COVID-19 disease considering the following sources of information: online media, Twitter and Telegram channels.

According to the authors, more than 14000 publications about the coronavirus indexed in WOS and Scopus in 2020 providing sharp rate of publication growth (1600% compared to the previous 5 years) [8, 9]. Authors performed the bibliographic analysis of scientific papers from 2000 to 2020 by existing visualization tools (for instance, VOSviewer [10, 11]) and noted the most significant terms in this field and emphasized international collaboration as a key for dealing with the COVID-19. Results show three main clusters for contributing papers: The USA, China and European countries. Meanwhile Belli et al [8] highlights the importance of open science: "Open science is the best method because it is an approach based on collaborative work, openness, and transparency in all stages including not only publication, but also data collection, peer review, and assessment."

Despite the existence of wide variety of software for monitoring coronavirus, a number of authors have proposed new visualization tools [12, 13, 14, 15]. Martínez Beltra ́n et al. [12] describe developed "web-based, user-friendly dashboard for interactive plotting" which is able to analyse both coronavirus and related data, such as maximum temperature, grocery and pharmacy, etc., mostly in Spain. As Marcílio-Jr et al. [13] writes: "Visualization-based strategies for monitoring the dissemination of diseases account for the fact that graphical representations can enhance the ability to identify data patterns and tendencies". So, authors proposed tool for drawing graphics, pie-charts, and especially maps to monitor the evolution of dissemination of coronavirus in Sao Paulo state. Another geospatial techniques are proposed by Mast et al. in their study [14]. Their analysis for the USA allows to reveal vaccinating distribution in two directions: over the time and over the place (cities, states). And finally, Chintala et al. develop software for workflow-based analysis over the world [15]. Their Python's application provides dynamic maps on cumulative daily confirmed COVID-19 cases for different countries.

All the papers below deal with coronavirus data rather than myths and rumors. Actually, there are much fewer publications that consider it true to spread false information on COVID-19 around the world via the Internet or mass-media. For instance, one study on the topic shows the statistical analyses to separate myths and facts [16]. Authors illustrate several graphs based on 13-item questionnaire of 125 participants. Pang et al. describes how governmental social media was used during COVID-19 pandemic [17]. Authors carry out word frequency and contents analyses for Macao governmental social media in Facebook and reveal that it can be useful to control rumors dissemination. Also Sond et al. (2021) show different types of rumors and how they can be corrected [18]. They perform their analysis of the data on Sina Weibo, the most popular microblogging site in China. Last but not least papers solve the issue of analyzing misinformation to some extent [19, 20, 21], emphasizing methods of text analysis rather than visualization tools. Some recent studies describe machine learning approaches to investigation of rumor dissemination [22, 23].

It ought to be noted that researchers rarely use visual tools for analysis in their works, except when they are looking at data within their own country [24, 25]. This paper tries to combine approaches from mentioned articles and apply them to main rumors over the world.

This paper presents methods of data processing and data visualization, which are conducted with the use of web technologies for building graphs and plotting data on the globe. The methods are examined and applied on the example of rumors about COVID-19. The described methods can be applied to analyse any data with a similar structure.

2. Methodology

2.1 Data collection and processing

With advances in information technology, researchers in various fields began to pay attention to the analysis of streaming data, which refers to data that is generated continuously from various sources. Also, as in the case of static data, the process of their processing can be represented as follows (Figure Figure 1). In this paper, visualization is used for performing analysis and showing the results.

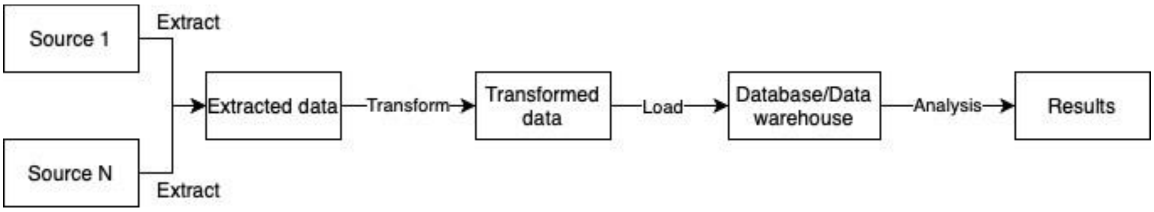


Figure 1. The process of data collection, processing and analysis

A study object was a publication in media. The sources of publications were social networks (Twitter, Facebook), instant messengers (Telegram channels), media (Russian News Agencies TASS and Interfax), governmental entities (.gov), commercial organizations (.com), Russian and foreign organizations (ru, .com, ...). In general, a publication had the following characteristics:

- Source;
- Title;
- Text;
- References to other sources (can be set implicitly in the text of the article);
- Publication date.

Considering the outlined characteristics it was decided to visualize information with the use of charts, maps, graphs. We also decided to build statistical report for better understanding of the collected data. The advantages of this method are the following:

1. Ability to define dissemination of information over time between sources, to distinguish initiating and disseminating subjects
2. Ability to identify key sources of information.
3. Ability to identify objects that are closely linked.

Thus, the generic algorithm of data collection and analysis can be represented as follows (Figure Figure 2).

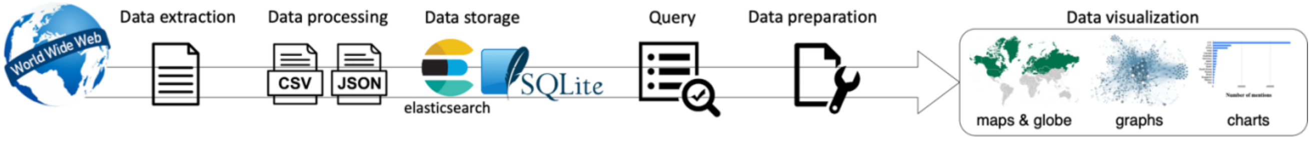


Figure 2. Algorithm of collection and analysis of data from the Internet information sources

Data collection consisted of three main stages: extraction, processing, storage. The first stage was data extraction from web pages of online information resources (carried out with the help of agent technologies [26]). The information collected from the Internet was weakly structured in terms of web markup and unstructured in terms of content, as textual information predominates. In view of the above, the next step was to structure the collected data according to the selected data model. Then, the processed data were uploaded to the selected repository for later use.

To analyse the data, first it was necessary to identify the information that was relevant to the topic from the total amount of data. This was done by making a query to the repository. The next step was to prepare the selected data, as different presentation formats are required to build the visualization, depending on the software method.

Creating a document repository involves performing the tasks of retrieving, processing, and data transferring. For the convenience of their further processing, SQLite 3, the database management system (DBMS), was used as an intermediate database. Then the processing of the data from SQLite 3 began: the data were read in small chunks, passed through a processor that performed a linguistic analysis of the texts, enriching the data with parameters such as keywords, extracted named entities (NER) and vector representations of the publication texts.

The case of the dissemination of coronavirus facts of varying degrees of veracity between countries is considered separately. The data source in this case is The CoronaVirusFacts/DatosCoronaVirus Alliance Database from the Poynter information resource [27]. As the database includes verified facts in over 70 countries and articles are published in at least 40 languages, there is no post-extraction processing step.

2.2 Linguistic analysis of medical texts

In order to obtain accurate research results, textual data were processed. In order to prepare textual data, which were in this case a media article on the topic of coronavirus, several steps were taken:

- extraction of named entities (Named Entity Recognition) from the texts;
- extraction of key phrases;
- retrieval of vector representations of article texts.

One of the most common NER application scenarios is the structuring of unstructured data (texts). Any text can be represented as a set of terms that can be stored in a database. The main types of entities are persons, countries (or regions), and organizations. However, nowadays there are also monetary units, events, numerals, and many others.

In this research the spaCy library [28] was used for text analysis and entity extraction. The following types of entities from the texts of media articles were extracted (Table 1).

Table 1. Entities extracted from texts with spaCy

Type of entity	**Desription**	**Examples**
PERSON	Names of persons	*Shailesh Kuber, Jackson, Shreyashi Sanyal*
NORP	Nationality, religious or political groups	*British, Chinese*
FAC	Airports, institutions, highways, bridges, etc.	*Atlanta International Airport, San Francisco International Airport*
ORG	Companies, agencies, institutions	*The Wuhan health commission, Weibo, Xinhua, Twitter, Imperial College London,*
GPE	Countries, towns, states	*Switzerland, Tehran, Miami, Kiev, India, Paris, U.S., Ukraine, Davos*
LOC	Locations of mountains, water areas (without geographical coordinates)	*Asia, Europe, Pacific, the Asia-Pacific region*
EVENT	Events	*Lunar New Year, The Spring Festival, Chinese New Year*

Since the task was to analyse texts, including those related to medicine, in addition to the standard NER we used a specially trained tool, BERN [29], for the extraction of biomedical entities (Table 2).

Table 2. Entities extracted from texts with BERN

Type of entity	**Desription**	**Examples**
Genes/Proteins	genes/proteins	*fda, hiv, ccr5, treg, hiv (r5) subtype, ncov, leronlimab antibody*
Diseases	diseases	*cough, respiratory illness, shortness of breath, infectious disease*
Drugs/Chemicals	drugs/chemicals	*chloride, vitriol, chromium, seprehvir, lidocaine, resiniferatoxin*
Species	species and roles	*human, person, patient, chicken, potato*

Keywords and phrases from texts were extracted with the use of Yake [30]. It is an automatic keyword extraction method and is based on the statistical characteristics of the text extracted from separate documents to select the most important keywords of the text.

spaCy performed the following operations for each publication from the database:

- tokenisation;
- searching for named entities (NER);
- searching for key phrases with the YAKE method;
- building vectors for the documents.

After that we ran the process of extraction of biomedical entities with BERN. Since BERN is very computationally intensive and its application as a remote web service is limited in terms of the number of queries, a randomly selected 10,000 publications were analysed.

2.3 Data presentation model

The Elasticsearch database was chosen as a data storage system. It can be considered both a NoSQL repository of JSON documents and a search engine based on the Lucene library. The structure of the database is almost identical to that of the intermediate SQLite 3 database shown below (Table Table 3).

Table 3. The structure of intermediate SQLite 3 database

Parameter name	***Description***	***Data type***
ID	internal identifier of the publication	int
date	publication date	date
author	author of the publication	text
domain	publisher name	text
title	publication title	text
content	text of the publication	text

topic_area	topic area: • general • finance • business • tech • medical	text
URL	URL of the publication	text
ner_events	event	text
ner_gpe	countries, towns, states	text
ner_loc	Locations of mountains, water areas (without geographical coordinates)	text
ner_norp	Nationality, religious or political groups	text
ner_org	companies, agencies, institutions	text
ner_person	Names of persons	text
ner_fac	airports, institutions, highways, bridges, etc.	text
yake_keywords	Keywords and phrases extracted with YAKE	text
bern_gene	genes	text
bern_species	animals	text
bern_disease	diseases	text
bern_drug	drugs	text
vector	Vector representation of the publication	text

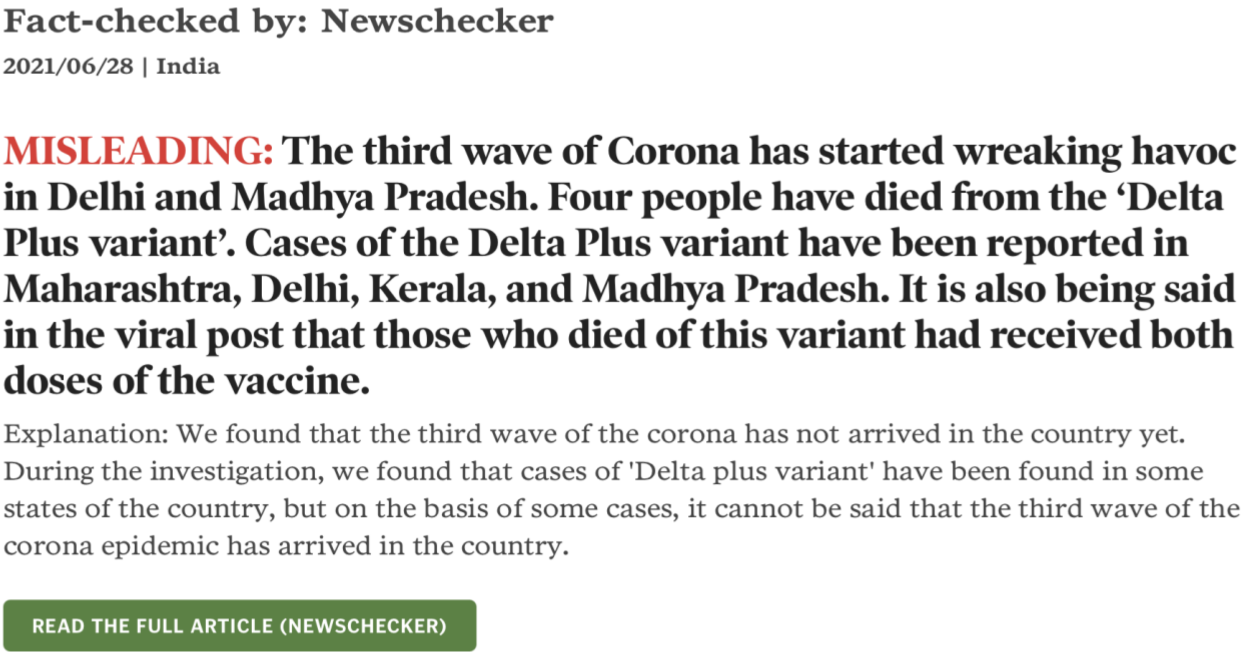
For the case of time propagation of information messages between sources, the "Links" field, derived from the SQLite 3 DBMS, was additionally used. The graph representation of the data was used to demonstrate the relations between sources [31].

Thus, data is represented in the form of a dynamically weighted directed graph, the nodes of which are information sources, edges are information messages, in which one source refers to another. Since information messages differ from each other by the time of posting, the graph is dynamic, that is, it can change its state over time, i.e. new nodes and edges appear over time (e.g. the graph is rebuilt at those moments when the information source posts a rumor). The weighting of the graph is applied to both edges and nodes: the weight of a node corresponds to the number of links to it, and the weight of an edge corresponds to the number of links between its initial and final nodes (all information messages are considered to be equal).

The graph consists of two types of edges (color legend of edges) since in the final samples an information message can be either a fake message or a contradiction to a fake message. To determine the type of edge, the text of an information message is analysed. It is compared with the thesaurus of words, which have a meaning of contradiction, to calculate the value of the criterion. If the value of the criterion exceeds a certain threshold, then the message is recognized as a contradiction.

The model which was built with the use of The CoronaVirusFacts/DatosCoronaVirus Alliance Database from the Poynter information resource looks as follows (an example of the raw data record is shown in Figure Figure 3):

- category/title of rumor,
- a description of information message,
- a country in which the rumor was spread,
- a source of a message (media, Facebook, etc.),
- a degree of credibility,
- explanation.



Fact-checked by: Newschecker

2021/06/28 | India

MISLEADING: The third wave of Corona has started wreaking havoc in Delhi and Madhya Pradesh. Four people have died from the 'Delta Plus variant'. Cases of the Delta Plus variant have been reported in Maharashtra, Delhi, Kerala, and Madhya Pradesh. It is also being said in the viral post that those who died of this variant had received both doses of the vaccine.

Explanation: We found that the third wave of the corona has not arrived in the country yet. During the investigation, we found that cases of 'Delta plus variant' have been found in some states of the country, but on the basis of some cases, it cannot be said that the third wave of the corona epidemic has arrived in the country.

Figure 3. An example of the record

This database has 50 types of veracity, but for this research the following was selected:

- False,
- Misleading,
- Mostly false,
- No evidence,
- Partially false.

3. Visualization and Analysis

3.1 Statistical characteristics of the frequency of mentions of different types of entities and keywords in publications

One of the possible ways to perform the analysis is to obtain quantitative characteristics of the frequency of mentioning of the following entity types: Organizations, Countries, Events, keywords and n-grams. For example, in the English publications such entities as The United States Securities and Exchange Commission, Reuters, Daily Express, CNN news agencies, Twitter, Facebook social networks were mentioned more often. An interesting fact was that the Second World War had been frequently mentioned in publications related to

COVID-19 (Figure Figure 4). However, the context, in which the World War was mentioned, were varied. It rangesfrom the cancellation of the Wimbledon tennis championship to the collapse of the world economy.

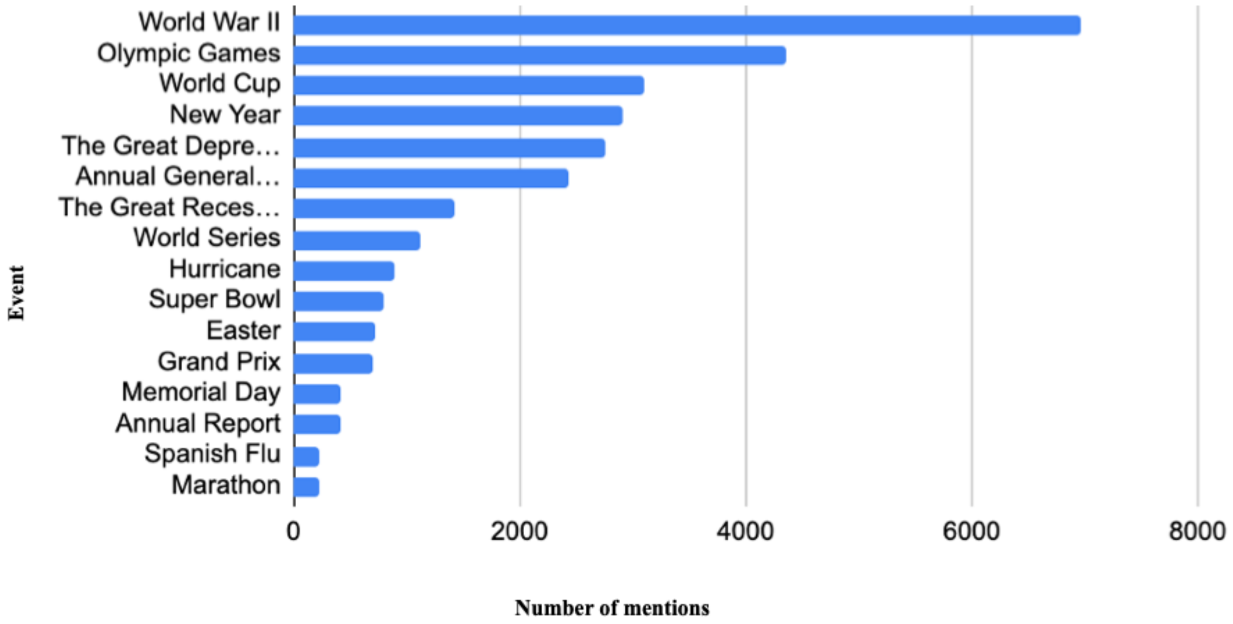


Figure 4. The frequency of mention of "Event" entities in English publications

Distributions of the number of mentions of drugs and various diseases were also constructed. Due to the complexity of organizing interaction with BERN and the long time required for query execution, a random sample of 10,000 publications was processed with its help (Figure *Figure* 5).

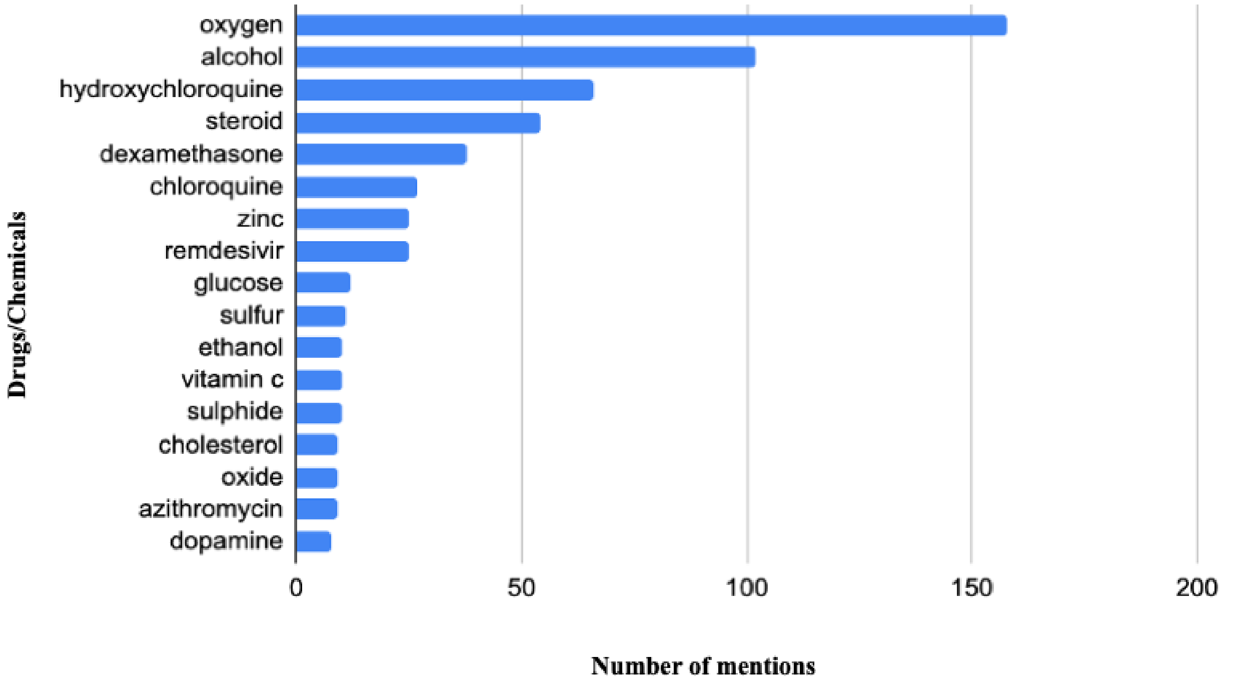


Figure 5. The frequency of mention of "Drug" entities in English publications

The frequent use of ‘oxygen’ is due to the fact that lack of medical oxygen was one of the major problems faced by patients and medical workers during the first wave of the coronavirus pandemic. The presence of ‘alcohol’ is justified by mention to sanitizers or, for example, by rumours of their denial that the coronavirus was controlled by taking highly concentrated alcohol. ‘Hydroxychloroquine’, ‘steroids’, ‘dexamethasone’, ‘chloroquine’ and ‘remsedivir’ are also frequently mentioned. These are the drugs that were used to treat coronavirus desease in the first wave.

The analysis of the frequency of mentioning of keywords and n-grams, which was obtained using the YAKE algorithm, revealed that the main context, apart from the terms coronavirus/pandemic/outbreak, was economic. It was evidenced by the presence of such terms as “company”, “stocks”, “economy/economic”, "Nasdaq", "wall street". It can also be noted that the months of April and June are the most frequently mentioned. The terms “virus”, “public health”, “patients”, “care” are also mentioned very frequently in the data set. It indicates the presence of large number of publications directly relevant to the pandemic.

3.2. Twitter analysis

The Panacea Lab at the University of Georgia, USA has published a dataset from Twitter posts related to the COVID-19 [32]. The tweets collected from Twitter are in all languages, but English, Spanish and French are the most common ones. In addition to the tweets themselves, daily hashtags, mentions and emoji were also collected. The set of all tweets and retweets contains ≈490 million messages, the cleaned set (without retweets) – ≈120 million. The diagram of number of tweets per day is shown below (Figure Figure 6).

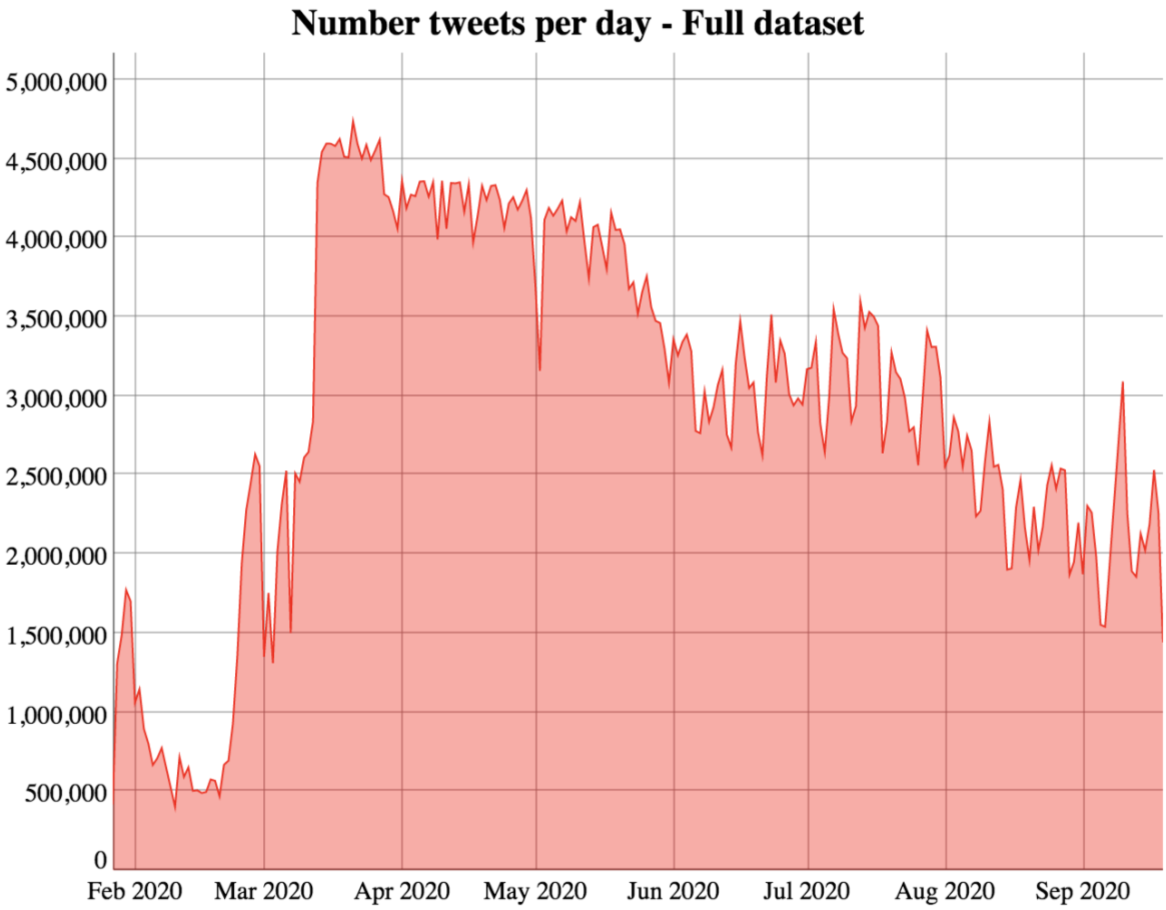
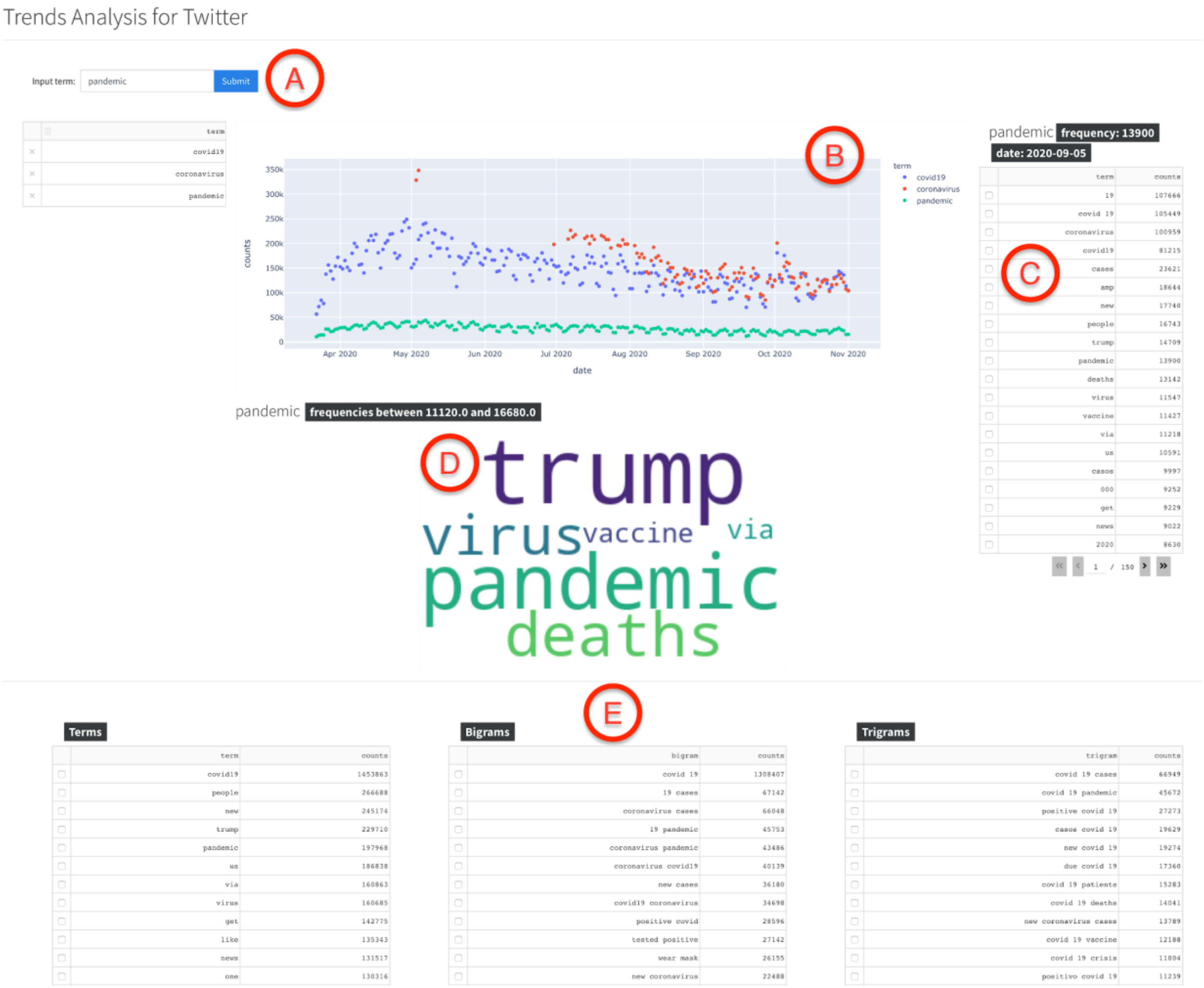


Figure 6. The diagram of number of tweets per day

For neurolinguistic programming (NLP) tasks, the resource provides 1,000 of the most commonly used terms, 1,000 of the most popular bigrams and 1,000 of the most popular trigrams.

Twitter has proven to be one of the main platforms for expressing public opinion. In our study, we use it to analyse trends in the frequency of use of terms and n-grams. For this purpose, all of the most used n-grams for each day since March 11, 2020 were collected, and a web application was developed for visual analysis, which allows displaying a dot plot of the frequency trend of a selected term over a certain time interval. The interface of the application is shown below (Figure 7). The application allows the user to enter the terms under study, bigrams or trigrams (or select them from lists of the most used terms) and to display trends in the frequency of their use over time on the dot plot. It was found that the trends of many terms are wave-like and contain extremes, which are the maximum number of mentions on certain dates.

The search for extremes in Twitter enables to determine the exact date when an event or news item attracted the most public attention. By clicking on an object on the graph it is possible to display the most frequently used terms and n-grams on the day of interest and to display a "word cloud" of the words closest in frequency of mention to the current word.



Trends Analysis for Twitter

Input term: pandemic

	term
×	covid19
×	coronavirus
×	pandemic

pandemic frequency: 13900

date: 2020-09-05

	term	counts
□	19	107666
□	covid 19	105449
□	coronavirus	100959
□	covid19	81215
□	cases	23621
□	amp	18644
□	new	17740
□	people	16743
□	trump	14709
□	pandemic	13900
□	deaths	13142
□	virus	11547
□	vaccine	11427
□	via	11218
□	us	10591
□	casos	9997
□	000	9252
□	get	9229
□	news	9022
□	2020	8630

1 / 150

pandemic frequencies between 11120.0 and 16680.0

Terms

	term	counts
□	covid19	1453863
□	people	266688
□	new	245174
□	trump	229710
□	pandemic	197968
□	us	186838
□	via	160863
□	virus	160685
□	get	142775
□	like	135343
□	news	131517
□	one	130316

Bigrams

	bigram	counts
□	covid 19	1308407
□	19 cases	67142
□	coronavirus cases	66048
□	19 pandemic	45753
□	coronavirus pandemic	43486
□	coronavirus covid19	40139
□	new cases	36180
□	covid19 coronavirus	34698
□	positive covid	28596
□	tested positive	27142
□	wear mask	26155
□	new coronavirus	22488

Trigrams

	trigram	counts
□	covid 19 cases	66949
□	covid 19 pandemic	45672
□	positive covid 19	27273
□	casos covid 19	19629
□	new covid 19	19274
□	due covid 19	17360
□	covid 19 patients	15283
□	covid 19 deaths	14041
□	new coronavirus cases	13789
□	covid 19 vaccine	12188
□	covid 19 crisis	11804
□	positivo covid 19	11239

Figure 7. Web interface of the application A – Entry field for terms and n-grams, B – Dot plot of the distribution of the number of mentions of selected n-grams over time, C – Most frequently used terms and n-grams for one day selected in the diagram, D – Word cloud with similar frequency of mention within 30% for the selected day, E – List of most frequently used terms, bigrams and trigrams for the whole period of time (since March 11, 2020).

The figure (Figure *Figure* 8) shows how the number of mentions of the term "COVID-19" increased sharply since April 2020. It peaked on April 29, 2020. After that the number of mentions gradually decreased. Such peaks can correlate with certain events in the world and are therefore a very useful indicator to highlight key points in a vast information field.

The method of semi-automated analysis was applied for the analysis of the English media for the days when the word or n-gram peaks had been detected in Twitter.

Thus, for April 29, 2020, 2,026 publications were found in the database of tweets. For each publication a vector of word embedding was stored, which enabled to apply different clustering methods to these vectors.

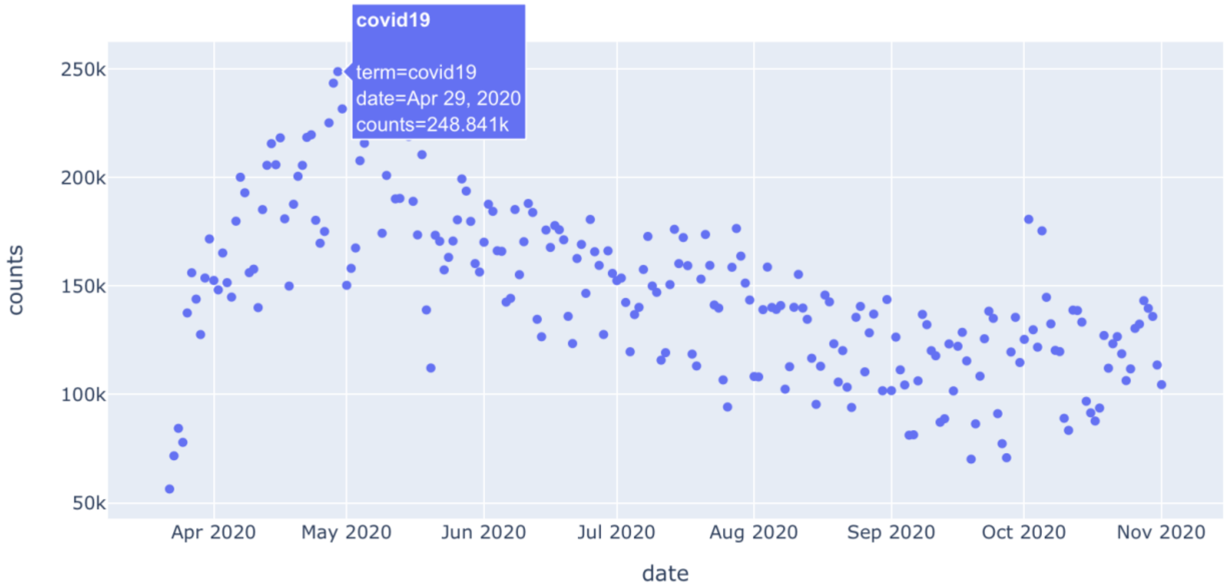


Figure 8. Trend in the frequency of mention of "COVID19" from April to October 2020

For the clusterization of publications, the HDBSCAN method [33] was used, which enabled to identify clusters based on the density distribution of distances between vectors. HDBSCAN, unlike the widely used K-means [34], does not require the number of clusters to be specified, which enables to detect anomalies efficiently [35]. With the use of HDBSCAN, 2,026 publications were distributed into 8 clusters. The locations of publications as data objects in relation to each other and their belonging to clusters (by colour) can be visually represented using t-SNE (t-distributed stochastic neighbor embedding) nonlinear dimensionality reduction method (Figure Figure 9).

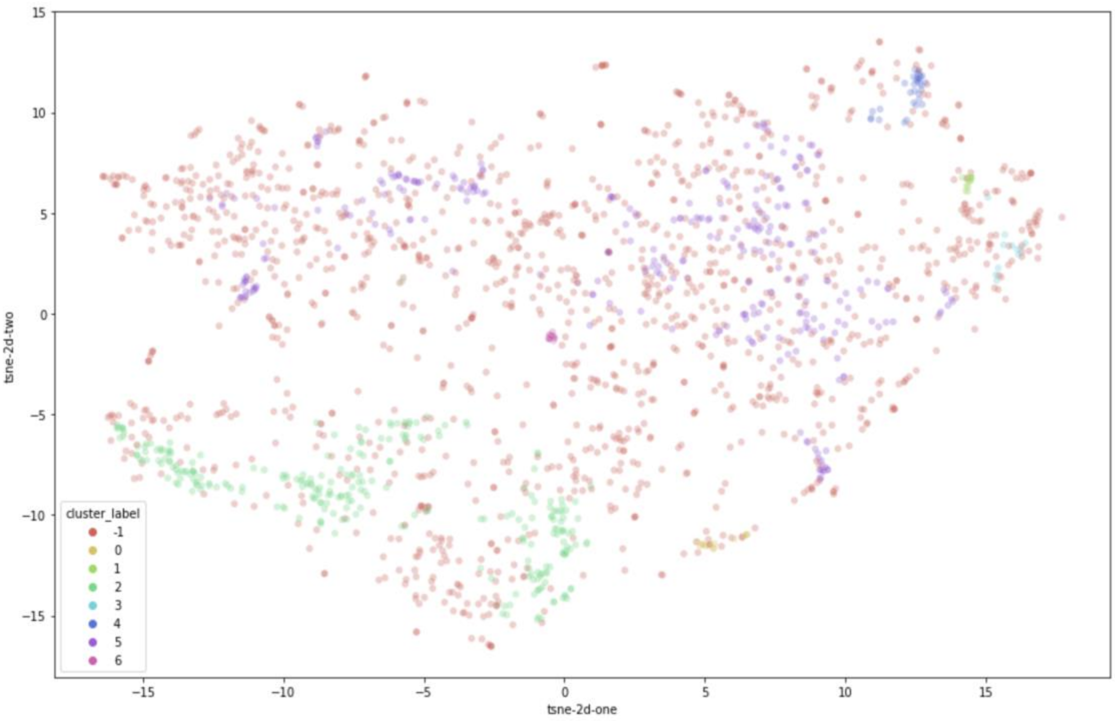


Figure 9. Distribution of publications by HDBSCAN clusters

About 50% of the publications are in the “-1” cluster, meaning that these publications are either isolated or do not overlap much with one another. They are anomalies.

Two large clusters (“2”, “5”) presumably represent the two most discussed topics. If we analyse their keywords, we can conclude that one topic is financial reports from various organisations, the second topic is relevant to the financial and economic situation in the USA, UK and China, directly related to COVID-19 and the declaration of the pandemic. The keywords and phrases extracted from all articles in the selected clusters are outlined below:

1. 'company financial performance', 'total operating expenses', 'net income adjusted', 'net cash provided', 'private equity partners', 'company annual report', 'rutherford and davidson', 'net operating income', 'net interest income', 'financial report company'.

2. 'health care workers', 'coronavirus testing trump', 'president mike pence', 'president', 'coronavirus', 'federal health officials', 'overlooking a horseshoe-shaped', 'united states', 'coronavirus pandemic', 'donald trump'.

It can be concluded that the most discussed topic was the financial situation, perhaps the day when many companies published their Q1 2020 financial reports and the media looked for relations between these reports and the coronavirus pandemic, assessing what impact it had on the financial markets. The titles of the first few publications from these clusters support this conclusion:

- *CB Insights: AI funding held steady in Q1, thanks to Waymo, but early-stage startups suffered;*
- *Airbus earnings q1 2020 as coronavirus crisis starts to bite;*
- *Anthem (ANTM) earnings Q1 2020, Deutsche Bank earnings Q1 2020;*
- *Facebook (FB) earnings Q1 2020;*
- *General Electric Q1 earnings 2020;*
- *Spotify (SPOT) earnings Q1 2020;*
- *Yum Brands (YUM) Q1 2020 earnings;*

- *Spirit of Texas Bancshares, Inc. Reports First Quarter 2020 Financial Results.*
- Considering other example, which is analysing mentions of 'hydroxychloroquine' using the developed web application on Twitter (Figure 10) in English publications (Figure Figure 11).

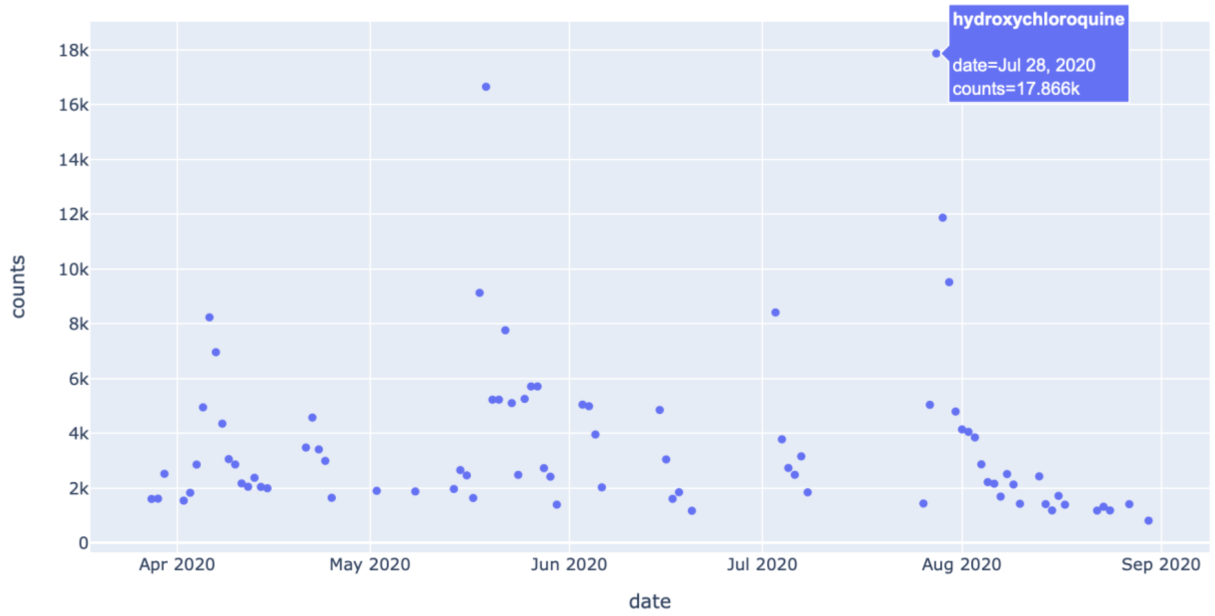


Figure 10. Distribution of the number of mentions of “hydroxychloroquine” on Twitter

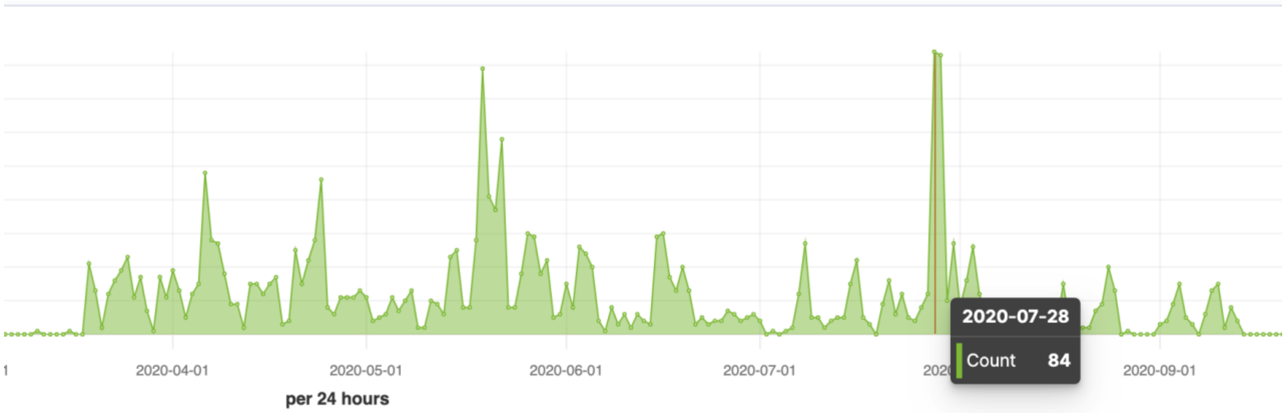


Figure 11. Distribution of the number of mentions of “hydroxychloroquine” in media publications

The distributions show two peaks, fully overlapping in the following dates: 29 May and 28-29 July 2020. The first peak is relevant to the White House confirmation of the fact that the US President, Donald Trump, took hydroxychloroquine. The second peak is related to publications about the effectiveness of hydroxychloroquine against the coronavirus infection.

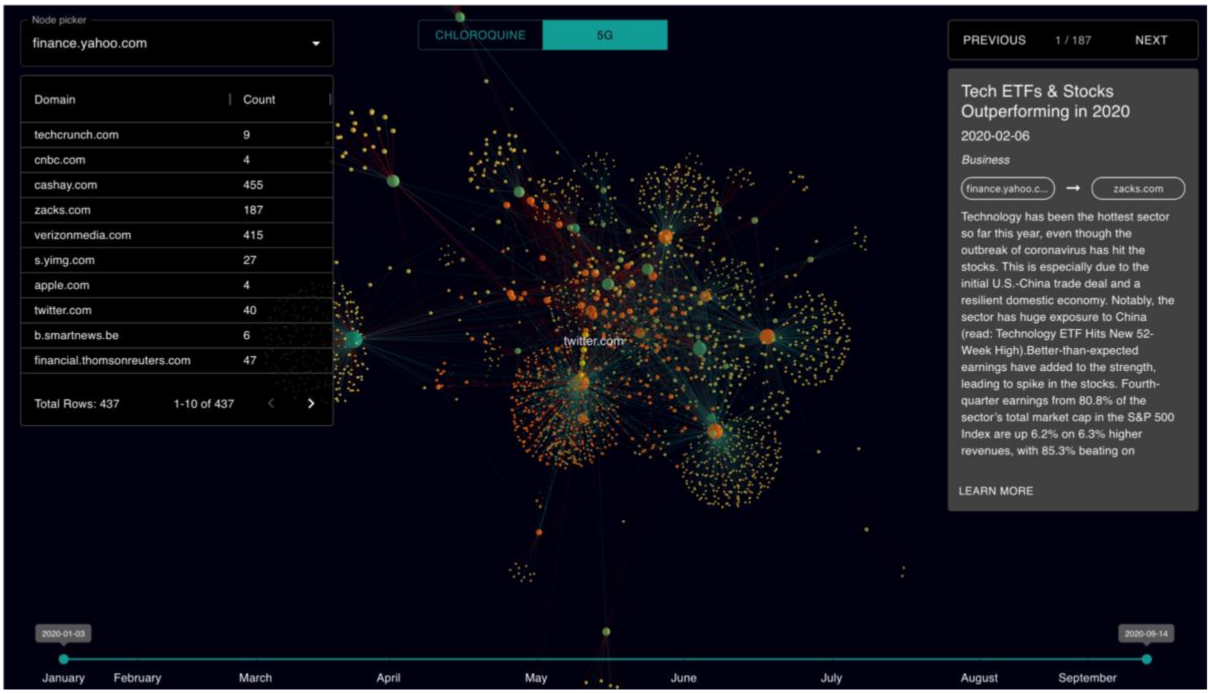
3.3 Dissemination of rumors and their contradictions between information sources

It is customary to build a graph with the help of such tools as Gephi, igraph, LargeViz, etc., or using additional libraries for programming languages. For dynamic and interactive work with graphs 3D Force-Directed Graph [36], a component for visualizing graphs was chosen, which uses ThreeJS/WebGL for rendering and force-directed graph drawing algorithms to build a graph.

To analyze information sources, the data is classified into separate rumors. The paper considers information messages on the topics '5G' and 'hydroxychloroquine'. The graph with '5G' rumors is represented below (Figure Figure 12).

The graph was built according to the following rules:

- nodes of the graph are information sources,
- if two nodes have an edge, which connects them, it means that the two sources are interconnected by an information message, in which one information source refers to another (henceforth such nodes will be called adjacent or connected),
- the graph is oriented. The direction of an edge is shown by the motion of the ball along the edge from the source of the information message to the link,
- a size of a node is proportional to the number of posts associated with this information source (both as a source and as a link),
- the graph is displayed using the force-directed graph drawing algorithm,
- an edge can belong to one of the two types: without contradiction (green edges) and with refutation (red edges),
- the nodes are color-coded: without selecting an active node, nodes have colors according to the color bar (Figure Figure 13); when an active node is selected, it and all nodes adjacent to it become orange.



Node picker

finance.yahoo.com

Domain	Count
techcrunch.com	9
cnbc.com	4
cashay.com	455
zacks.com	187
verizonmedia.com	415
s.yimg.com	27
apple.com	4
twitter.com	40
b.smartnews.be	6
financial.thomsonreuters.com	47

Total Rows: 437 1-10 of 437

Tech ETFs & Stocks Outperforming in 2020

2020-02-06

Business

finance.yahoo.c... → zacks.com

Technology has been the hottest sector so far this year, even though the outbreak of coronavirus has hit the stocks. This is especially due to the initial U.S.-China trade deal and a resilient domestic economy. Notably, the sector has huge exposure to China (read: Technology ETF Hits New 52-Week High).Better-than-expected earnings have added to the strength, leading to spike in the stocks. Fourth-quarter earnings from 80.8% of the sector's total market cap in the S&P 500 Index are up 6.2% on 6.3% higher revenues, with 85.3% beating on

Figure 12. Graph web interface

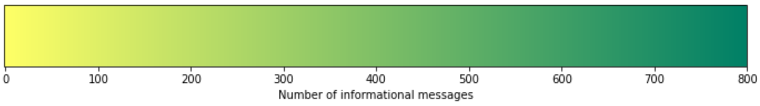
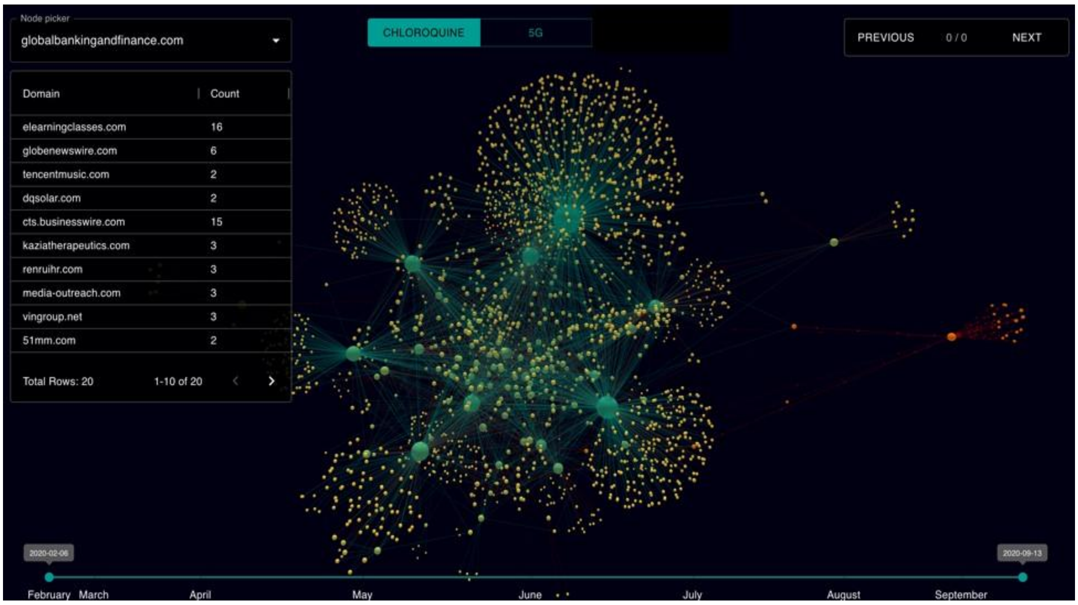


Figure 13. Color bar

The interface of the constructed web application makes it possible to analyze rumors on a selected topic by selecting individual nodes, choosing a time interval and considering information messages between two nodes connected by an edge.

This approach allows us to identify the primary sources of the dissemination of false information and the main distribution nodes. For instance, the rumor about chloroquine shows that “Global Banking & Finance Review” released only rumors with refutations (Figure Figure 14). On the other hand, the figure (Figure Figure 15) represents the possibilities of the graph. According to the graph, “MarketBeat” was the first resource which published information about connection between 5G and coronavirus, but with time its contribution to the total share of messages decreased. Also, analysis of the data using the graph revealed that most often information sources refer to Twitter (as evidenced by the maximum total input flow and the location of the node in the center of the graph), followed by “Verizon Media”. The graph also shows that the most active source of informational messages is "Yahoo! Finance", “The Conversation”, etc. For instance, in this time interval “The Conversation” released most of the messages in May and then reduced the intensity.

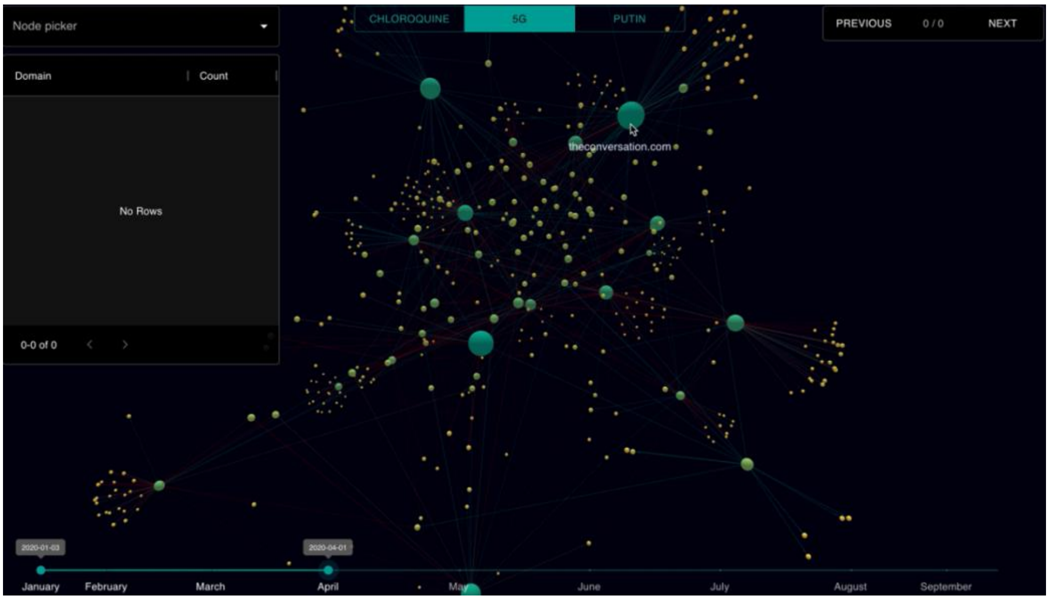


Node picker: globalbankingandfinance.com

Domain	Count
elearningclasses.com	16
globenewswire.com	6
tencentmusic.com	2
dqsolar.com	2
cts.businesswire.com	15
kaziatherapeutics.com	3
renruihr.com	3
media-outreach.com	3
vingroup.net	3
51mm.com	2

Total Rows: 20   1-10 of 20

Figure 14. The “hydroxychloroquine” rumour



Node picker

Domain	Count
No Rows	

0-0 of 0

Figure 15. Example of the analysis

3.4 Dissemination of facts of varying degrees of credibility between countries

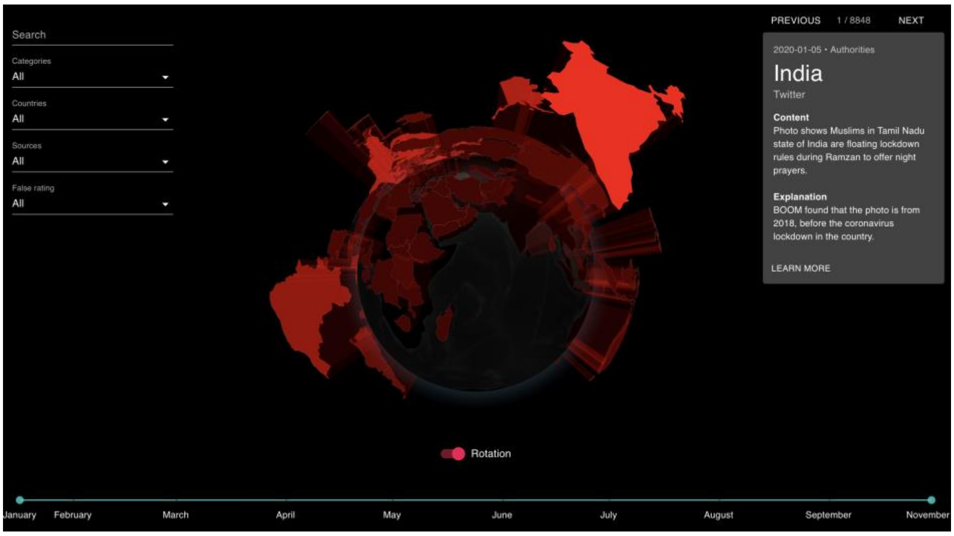
To analyze the dissemination of facts of various degrees of credibility between countries, a proper method is to plot them on the world globe. The authors used the Globe.GL web component [37], which enables applying data visualization layers to a three-dimensional globe. The final form of the web application is demonstrated in Figure Figure 16.

The globe has an intuitive interface: the fewer information messages related to the country, the darker the country looks and the closer it is to the globe; and vice versa: the more information messages related to the country, the more reddish the country and the higher it is raised above the globe. The web interface enables analyzing data by creating a query. It may include:

- a word or phrase;
- a type of rumors (7 main rumors and all the rest);
- a country to which a rumor belongs;
- a source of an information message;
- a degree of credibility of a rumor;
- time range.

The search query tool together with the user-friendly interface enables carrying out sufficient research. The analysis revealed that most of the rumors about the coronavirus came from India and the United States, while information is spreading in neighboring countries.

This approach allows us to identify the hidden relationships between countries. For example, when analyzing rumors connected with the United States, it can be seen that there are similar rumors in Spain (35 common rumors), France (20 common rumors), Canada (12 common rumors) and Ukraine (9 common rumors). This suggests that there are close connections between these countries.



Search

Categories: All

Countries: All

Sources: All

False rating: All

2020-01-05 • Authorities

India

Twitter

Content
Photo shows Muslims in Tamil Nadu state of India are floating lockdown rules during Ramzan to offer night prayers.

Explanation
BOOM found that the photo is from 2018, before the coronavirus lockdown in the country.

Figure 16. Globus web interface

4. Conclusion

Even though various objects from physical installations to social networks and Internet media can be a source of streaming data, the analysis of such sources can be carried out using similar methods and tools. The paper considers the task of visualizing the dissemination of rumors about coronavirus disease in several ways. The key feature of this study is the statistical analysis of the dynamic system, so with the help of visualization, the spread of information and assessment of the intensity of information dissemination is explicitly shown. The use of modern text-processing tools increases the accuracy of the research.

Statistical analysis identifies the most frequently mentioned entities of different types (Persons, Diseases, etc.), which enables to identify the most relevant entities and keywords related to the disease COVID-19. For example, the frequent use of economic terms such as "stocks" and "wall street" indicates the impact of the disease on the economic situation.

The considered methods of textual information analysis enable automated detection of various entities, such as names of medicines, persons, organizations, etc, in the automated mode. The used visualisation methods in the developed web-application for Twitter analysis enables to detect events that cause public resonance and are discussed by the society. It is possible to clearly see the dynamics of interest in topics. With the use of both web application and statistical analysis tools it is possible to draw conclusions about the dissemination of topics both on Twitter and in media.

A dynamic graph was built to analyze the dissemination of rumors on 5G and hydroxychloroquine in the world media. The graph enables to identify the most "important" nodes (information sources that produce a lot of information messages), considering the process of dissemination rumors over time and visually determine a cluster structure of objects.

To analyze the global situation about coronavirus disease according to rumors, a globe with a search engine was built, which demonstrates the spread of rumors between countries. With the help of this application, the countries in which the rumors appeared most often and their influence on the rest of the world have been identified.

According to the authors, the obtained data model together with the described tools may serve as a good basis for analyzing streaming data of various nature.

The developed visualisation tools can be applied for studying other topics and will become the basis for information management work in information networks.

References

[1] Cisco, Cisco Annual Internet Report (2018–2023) White Paper, 2018. URL: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internetreport/white-paper-c11-741490.html.

[2] Cisco, VNI Complete Forecast Highlights, 2017. URL: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecasthighlights/pdf/Global_2022_Forecast_Highlights.pdf.

[3] The World Bank, World Development Report 2021, 2021. URL: https://wdr2021.worldbank.org/stories/crossing-borders.

[4] Shiau, W.-L., Dwivedi, Y.K., Yang, H.S., Co-citation and cluster analyses of extant literature on social networks. International Journal of Information Management 37(5), 390-399. doi: 10.1016/j.ijinfomgt.2017.04.007.

[5] Shiau, W.-L., Dwivedi, Y.K., Lai, H.-H., Examining the core knowledge on facebook, International Journal of Information Management 43, 52-63. doi: 10.1016/j.ijinfomgt.2018.06.006.

[6] Gruzd, A., De Domenico, M., Sacco, P.L., Briand, S., Studying the COVID-19 infodemic at scale, Big Data and Society 8(1). doi: 10.1177/20539517211021115.

[7] Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of rumors and misinformation on COVID-19 in social media. Journal of Preventive Medicine and Public Health, 53(3), 171–174. https://doi.org/10.3961/jpmph.20.094.

[8] Belli, S., Mugnaini, R., Baltà, J., Abadal, E. (2020) Coronavirus mapping in scientific publications: When science advances rapidly and collectively, is access to this knowledge open to society? Scientometrics, 124(3), p. 2661-2685.

[9] Pal, J.K. (2021) Visualizing the knowledge outburst in global research on COVID-19. Scientometrics.

[10] Khakimova, A.Kh., Zolotarev, O.V., Berberova, M.A., Coronavirus infection study: Bibliometric analysis of publications on COVID-19 using PubMed and Dimensions databases, Scientific Visualization 12(5), 112-129. doi: 10.26583/SV.12.5.10.

[11] VOSViewer, VOSViewer, 2021. URL: https://www.vosviewer.com.

[12] Martínez Beltrán, E.T., Quiles Pérez, M., Pastor-Galindo et al. (2021) COnVIDa: COVID-19 multidisciplinary data collection and dashboard. Journal of Biomedical Informatics, 117, 103760.

[13] Marcílio-Jr, W.E., Eler, D.M., Garcia, R.E., Correia, R.C.M., Rodrigues, R.M.B. (2021). Visual analytics of COVID-19 dissemination in São Paulo state, Brazil. Journal of Biomedical Informatics, 117, 103753.

[14] Mast, T.C., Heyman, D., Dasbach, E. et al. (2021) Planning for monitoring the introduction and effectiveness of new vaccines using real-word data and geospatial visualization: An example using rotavirus vaccines with potential application to SARS-CoV-2. Vaccine: X, 7, 100084.

[15] Chintala, S., Dutta, R., Tadmor, D. (2021) COVID-19 spatiotemporal research with workflowbased data analysis. Infection, Genetics and Evolution, 88, 104701.

[16] Konar, K., Kabli, N. (2021) A statistical analysis on Covid-2019 to distinguish between myths and facts with data visualization. IOP Conference Series: Materials Science and Engineering, 1022(1), 012043.

[17] Pang, P.C.-I., Cai, Q., Jiang, W., Chan, K.S. (2021) Engagement of government social media on facebook during the COVID-19 pandemic in Macao. International Journal of Environmental Research and Public Health, 18(7),3508.

[18] Song, Y., Kwon, K.H., Lu, Y., Fan, Y., Li, B. (2021) The "Parallel Pandemic" in the Context of China: The Spread of Rumors and Rumor-Corrections During COVID-19 in Chinese Social Media. American Behavioral Scientist.

[19] Shahi, G.K., Nandini, D., FakeCovid--A multilingual cross-domain fact check news dataset for COVID-19, arXiv preprint arXiv:2006.11343. doi: 10.36190/2020.14.

[20] Patwa, P., Sharma, S., Pykl, S., et al., Fighting an Infodemic: COVID-19 Fake News Dataset, Communications in Computer and Information Science 1402 CCIS, 21-29. doi: 10.1007/978-3030-73696-5_3.

[21] Mookdarsanit, P., Mookdarsanit, L., The covid-19 fake news detection in thai social texts, Bulletin of Electrical Engineering and Informatics 10(2), 988-998. doi: 10.11591/eei.v10i2.2745.

[22] Bhowmik S., Prosun P.R.K., Alam K.S. (2022) A Novel Three-Level Voting Model for Detecting Misleading Information on COVID-19. In: Mandal J.K., De D. (eds) Advanced Techniques for IoT Applications. EAIT 2021. Lecture Notes in Networks and Systems, vol 292. Springer, Singapore. https://doi.org/10.1007/978-981-16-4435-1_36.

[23] Al-Sarem, M.; Alsaeedi, A.; Saeed, F.; Boulila, W.; AmeerBakhsh, O. A Novel Hybrid Deep Learning Model for Detecting COVID-19-Related Rumors on Social Media Based on LSTM and Concatenated Parallel CNNs. *Appl. Sci.* **2021**, *11*, 7940. https://doi.org/10.3390/app11177940.

[24] Biancovilli, P., Makszin, L. & Jurberg, C. Misinformation on social networks during the novel coronavirus pandemic: a quali-quantitative case study of Brazil. *BMC Public Health***21,** 1200 (2021). https://doi.org/10.1186/s12889-021-11165-1.

[25] Ceron, W., Gruszynski Sanseverino, G., de-Lima-Santos, MF. *et al.* COVID-19 fake news diffusion across Latin America. *Soc. Netw. Anal. Min.* **11,** 47 (2021). https://doi.org/10.1007/s13278-021-00753-z.

[26] Ulizko, M., Pronicheva, L., Artamonov, A., Tukumbetova, R., Tretyakov, E., Complex Objects Identification and Analysis Mechanisms, Advances in Intelligent Systems and Computing 1310, 517-526. doi: 10.1007/978-3-030-65596-9_63.

[27] Poynter, The CoronaVirusFacts/DatosCoronaVirus Alliance Database, 2020. URL: https://www.poynter.org/ifcn-covid-19-misinformation.

[28] SpaCy, Industrial-Strength Natural Language Processing; 2021. URL: https://spacy.io.

[29] Github, BERN; 2021. URL: https://github.com/dmis-lab/bern.

[30] Github, Yake; 2021. URL: https://github.com/LIAAD/yake.

[31] Ulizko, M.S., Antonov, E.V., Artamonov, A.A., Tukumbetova, R.R., Visualization of graph-based representations for analyzing related multidimensional objects, Scientific Visualization 12(4), 133142. doi: 10.26583/sv.12.4.12.

[32] Github, Covid19_twitter; 2020. URL: https://github.com/thepanacealab/covid19_twitter.

[33] HDBSCAN, How HDBSCAN works; 2021. URL: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

[34] Skilit-Learn, KMeans; 2021. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[35] Grigorieva, Maria and Grin, Dmitry. Clustering error messages produced by distributed computing infrastructure during the processing of high energy physics data // International Journal of Modern Physics Vol. 36, No. 10, 2150070 (2021).

[36] Github, 3D Force-Directed Graph, 2020. URL: https://github.com/vasturiano/3d-force-graph.

[37] Github, Globe.GL, 2020. URL: https://github.com/vasturiano/globe.gl.