Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2022 SCPE. Volume 23, Issues 1, pp. 9–21, DOI 10.12694/scpe.v23i1.1957

EXPLORING USABILITY OF REDDIT IN DATA SCIENCE AND KNOWLEDGE PROCESSING

JAN SAWICKI[*], MARIA GANZHA[†], MARCIN PAPRZYCKI[‡] AND AMELIA BĂDICĂ[§]

Abstract. This contribution argues that Reddit, as a massive, categorized, open-access dataset, is a useful data source, for "almost any topic". Hence, it can be used in data science, e.g. for knowledge exploration. This statement is backed-up with presented analysis, based on 180 manually annotated papers, related to Reddit itself, and data acquired from popular databases of scientific papers. Finally, an open source tool is introduced, which provides an easy access to Reddit resources, and an exploratory data analysis of how Reddit covers selected topics. These functions can be used as a prelude analysis to a broader exploration of Reddit's applicability.

Key words: Reddit, online forum, dataset, text mining, information retrieval, data analytics, knowledge processing

AMS subject classifications. 68-02, 68U15, 68U99, 68U01, 68T50, 91Fxx

1. Introduction.

Recently, social networks and content sharing networks became popular repositories of data, used for information and knowledge processing (especially for information retrieval). The aim of this work is to explore the usability of Reddit as a data source. In this context, we present a review of scientific literature about Reddit itself, its presence in scientific databases, and elaborate its "topical coverage". Moreover, for the latter study, a specialized tool (*Reddit-TUDFE*)' is introduced, which allows for fast check of Reddit coverage of a selected topic. The key contributions of this work are answers to the following research question (RQs):

- **RQ1**: What are the most popular methods to acquire Reddit data? (do they allow capturing graph networks[1])
- **RQ2**: What problems are the most researched when using Reddit as a dataset?
- **RQ3**: How does Reddit usage in data science change over time? Is it declining or is it increasing?
- **RQ4**: Are there any popular topics that are not (substantially) covered on Reddit?
- **RQ5**: Is Reddit used as a single dataset, or with datasets from other online platforms?[2]

These questions are essential for further planned research and positive answer would mean that Reddit is a proper choice for proceeding with the project of information retrieval about popular trends, using graph databases and complex networks. Moreover, positive answers would indicate that Reddit may be a competitor (or a companion) to explorations based on more popular data sources, like Twitter.

2. What is Reddit.

Let us start from a brief description of Reddit. It is a web content rating and discussion website [31]. It was created in 2005 and is ranked as the 17th most visited website in the world, with over 430 million monthly active users[3] and total of over 13 billion posts and comments[4]. The structure of Reddit is illustrated in Figure 2.1.

Reddit is divided into thematic subfora (so called, *subreddits*) dynamically created by its users. Therefore, the topic structure is systematically evolving, in response to user needs. Each subreddit has its *moderators*

[*] Warsaw University of Technology, Department of Mathematics and Information Sciences, (jan.sawicki2.dokt@pw.edu.pl).

[†] Warsaw University of Technology, Department of Mathematics and Information Sciences (m.ganzha@mini.pw.edu.pl).

[‡] Systems Research Institute Polish Academy of Sciences (marcin.paprzycki@ibspan.waw.pl).

[§] University of Craiova, Department of Statistics and Business Informatics (amelia.badica@edu.ucv.ro).

[1] Here, graph networks are of special interests, because it can be observed that large number of methods of data extraction and analysis are focused on application of graph theory.

[2] Answer to this question is crucial to establish (suggest) additional datasets, which could/should be used with Reddit.

[3] https://www.statista.com/topics/5672/reddit/#dossierSummary

[4] https://www.redditinc.com/

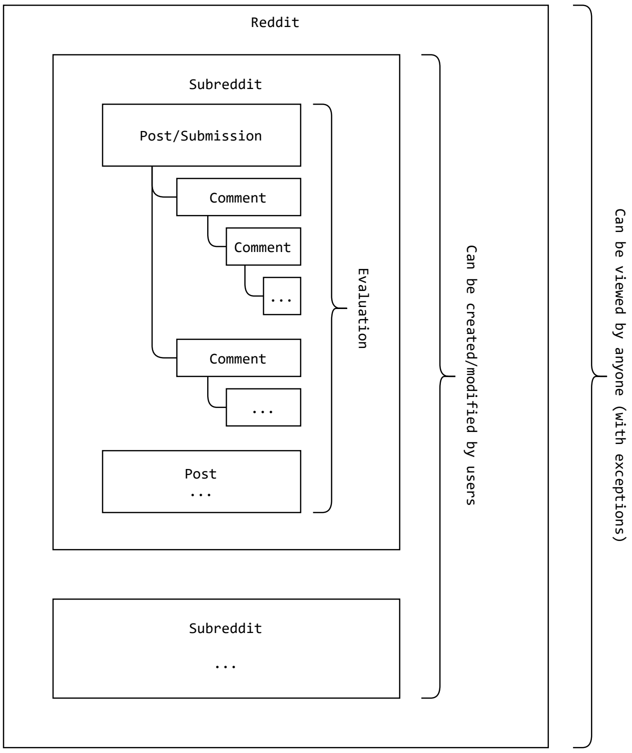

Fig. 2.1. *Reddit structure*

who may supervise *submissions* and *comments.* Comments are linked to submissions, or to earlier comments, forming a tree-like structure.

2.1. Content access rules and restrictions. Most of the subreddits are public (for registered and non-registred users). There are some exceptions based, for instance, on karma points (i.e. user's score), comments, gold (i.e. Reddit's currency that can be purchased with real money), moderator status, time on Reddit, username and others. For instance, such restriction can be be applied to even a Harry Potter house preference (e.g. r/gryffindor)[5]. Here, let us note that the Reddit topic explorations tool (introduced in Section 5), is based only on access to publicly available data.

2.2. Accessibility – Reddit API vs. Pushshift API. Not only is the data on Reddit publicly accessible (with the exception of private communities), it is also made available via the official Reddit API[6]. However, in the course of literature review, it was found that most researchers do not actually use it. Over 90% of analyzed papers either use ready datasets scraped earlier from Reddit and posted online (possibly in an annotated form), or they choose the Pushshift API [4]. None of the analysed papers stated the explicit reason for this choice (very few even mention how their datasets have been retrieved). However, practically testing capabilities of Reddit API and Pushshift API shows that the key factor could have been that Reddit API does not allow easy retrieval of historical data, while Pushshift API does. Hence, when developing the Reddit data exploration tool, the Pushshift API was used.

3. Data acquisition and processing. To explore Reddit, as seen by the scientists, a dataset of all, most recent, papers available on arXiv has been assembled – a total of 180 papers. All of them were related to Reddit

[5] https://www.reddit.com/r/ListOfSubreddits/wiki/privates
[6] https://www.reddit.com/dev/api/

and submitted to arXiv between 01-01-2019 and 01-03-2021 (and retrieved on 30-03-2021[7]). This dataset has been processed both manually and automatically. First, collected papers have been manually annotated with four attribute sets: **topic** (a general area of research), **methods** (theoretical approach, e.g. neural network, text embedding), **dataset** and **technologies** (practical software, e.g. BERT [10]). Next, obtained results were merged using arXiv identification code and the publicly available data, i.e. the content (title and raw text) and the bibliometric metadata. This allowed extraction of information presented in Section 4. All collected content has been converted to a raw text file, using PDF Miner software [38]. Next, the key features of titles and texts have been cleaned and mined using the NLTK framework [26] (for sentiment and subjectivity), and TF-IDF [36] for vectorization (both frameworks are part of the scikit-klearn library [34]).

4. Analysis and findings. As a result of processing of collected data, we were able to formulate a number of observations. Let us summarize the most important ones.

4.1. Metadata and bibliometrics. First, let us consider a few noticeable bibliometric and authorship statistics, gathered using Semantic Scholar[8] and presented in Figures 4.1, 4.2 and 4.3.

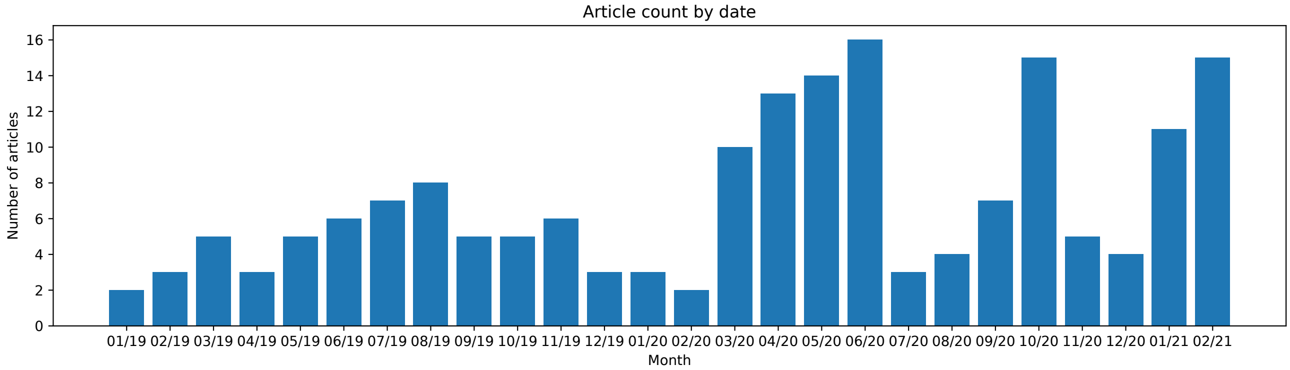

FIG. 4.1. *Article count by the month of the submission date*

As shown in Figure 4.1, there is a significant growth in the number of articles (related to Reddit) published after March 2020 (correlated with the outburst of the COVID-19 pandemic) and in October 2020 (correlated with notification dates for many scientific conferences [40]). The latter fact was also verified during manual processing of collected data. This suggests that Reddit was used to provide data related to COVID pandemic and that it is used as a data source for contributions to, broadly understood, data analytics related conferences.

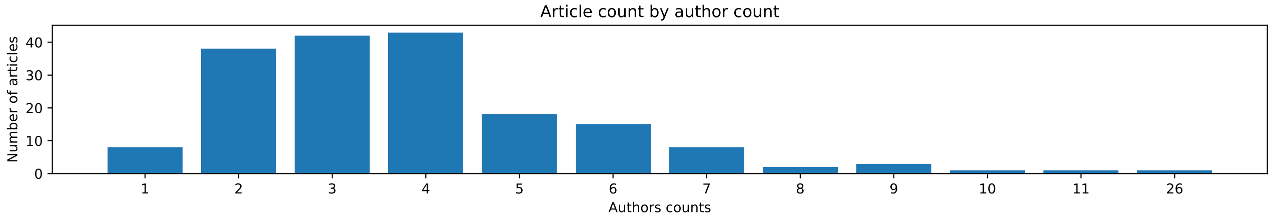

FIG. 4.2. *Number of authors for the selected papers*

Next, as seen in Figure 4.2, majority of papers were written by 2-4 authors, with one having 26 authors [13].

[7] https://arxiv.org/search/advanced?&terms-0-term=reddit&classification-computer_science=y&date-from_date=2019-01-01&date-to_date

[8] https://api.semanticscholar.org/

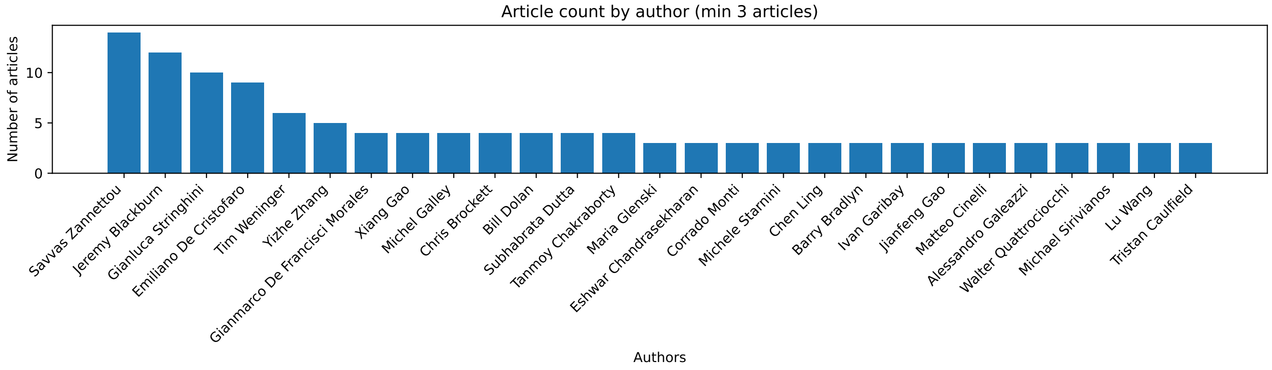

Fig. 4.3. *Article count by the author*

Finally, Figure 4.3, shows that the most prolific authors, of Reddit-related papers, were Savvas Zannettou (Max-Planck-Institute), Jeremy Blackburn (Binghamton University) and Gianluca Stringhini (Boston University). This seems to suggest that large number of scientific content, generated while studying Reddit posts, is delivered by a close circle of scientists.

4.2. Analysis of topic, methods and technology. Topics, methods and technologies are key to answer RQ1 and RQ2. These were extracted manually from the collected papers. They are summarized in Figures 4.4, 4.5 and 4.6.

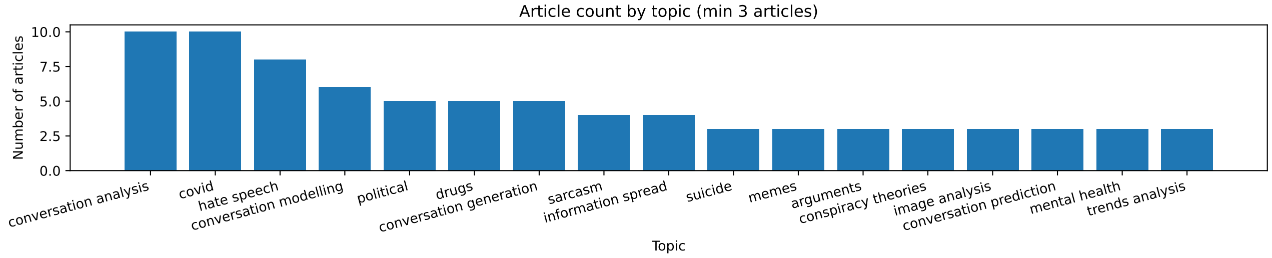

Fig. 4.4. *Article count by article topics manually annotated in all papers*

Figures 4.4 and 4.5 show clearly that the most popular research topic is *conversation*, which matches the fact that Reddit is a discussion forum. Due to the timing of this work (overlapping with the COVID-19 pandemic), the second most common topic is *COVID* (see Figure 4.4).

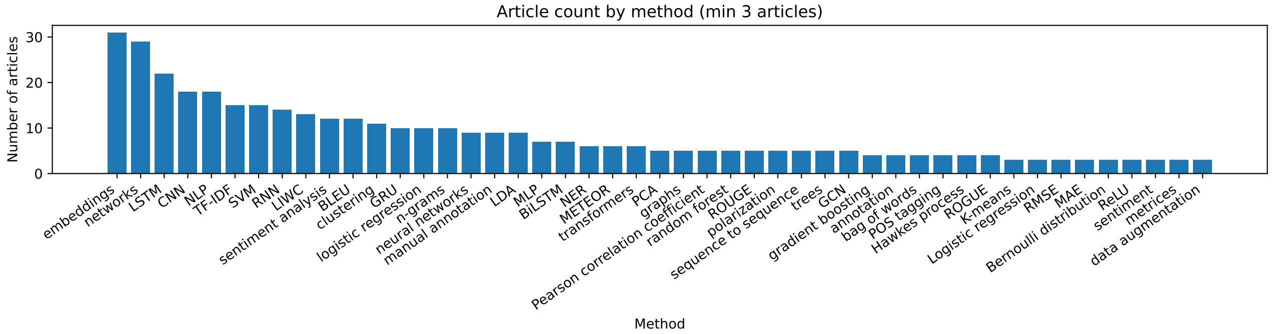

Fig. 4.5. *Article count by methods manually annotated in all papers*

Since Reddit consists mostly of text-based discussions, it is not surprising that the two most common methods, in Reddit-related research, are *text embeddings*, used in text processing, and *networks*, used for social network analysis. Note that, in the reported results, “network” (understood as a graph) and “neural network” are separate terms.

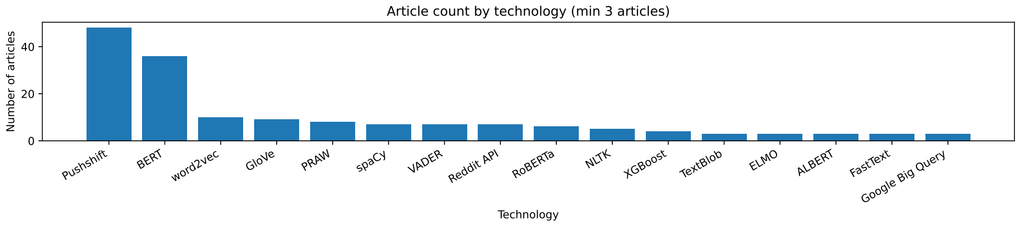

Fig. 4.6. *Article count by technologies manually annotated in all papers*

Regarding technologies (shown in Figure 4.6), over 45% of studies used Pushshift API [4] for Reddit data extraction, and over 35% applied BERT [10] embedding (and its variations) for the natural language processing.

Finally, topics and methods have been combined in a correlation heatmap (Figure 4.7).

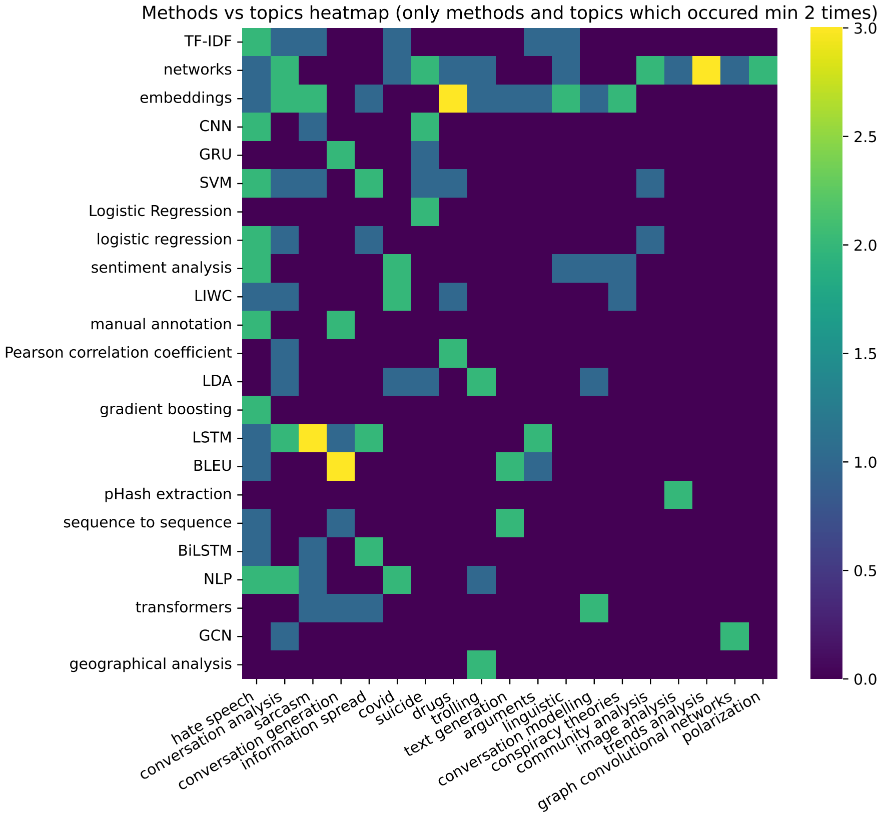

Fig. 4.7. *Research methods correlated with article topics*

Here, a few significant correlations have been established. However, they have to be considered keeping in mind that they materialize in the context of a specific dataset, created on from contributions reporting research that used Reddit as a data source. therefore, no claim is made that these observation can be immediately generalized beyond the dataset used in this work. However, based on general knowledge of the field, they seem to be in line with more general trends.

- Papers related to *drugs* typically use *word embeddings*. However, this can be related to the overall popularity of word embeddings in the research conducted in early 2020th (see, for instance, the citation count for [10]).
- *Networks* are typically applied in analysis of *trends*, e.g. topic popularity (this is a key finding for RQ1).
- Articles dealing with *sarcasm* often use *LSTM networks*.
- Research devoted to the *conversation generation* typically applies the *BLEU metric*.

4.2.1. Topics of knowledge and information processing. The topic of information and knowledge retrieval is one of the main aims of undertaken analysis. Hence, this category was checked specifically. Even though many works focus on information spreading in online communities [13, 41, 11, 12], there is hardly any focus purely on information/knowledge retrieval. There are precisely two papers (1% of the considered work) related to knowledge processing (specifically, knowledge graphs [6, 43]). Expanding arXiv search, to capture all articles including terms “knowledge” and “Reddit”, resulted in 4 records, none of which is related to knowledge capture. Pairing keyword “Reddit” with “information retrieval” or ’information processing” yielded 0 results. Therefore, top knowledge processing/management-related conferences were searched, but only one contribution [18], about knowledge and Reddit, has been found (published by the K-CAP conference in 2011[9]). This renders Reddit as a source that is definitely underexplored in terms of knowledge/information mining.

4.2.2. Use of Reddit combined with other datasets. Moving to the **RQ5**, it was discovered that among papers that use Reddit, over 30% also use Twitter, which is a data source that is very often used for sentiment analysis [24]). Other datasets that have been utilized together with Reddit are: Facebook, 4Chan, YouTube, and Gab. Each of them appears in less than 10% of papers, which used Reddit (details are shown in Figure 4.8). Datasets are rarely used in triplets, i.e. Reddit and two other datasets (the highest scoring triplets were Reddit, combined with Twitter and Facebook 6.6% of articles; Reddit, used together with Twitter and 4chan 6% (e.g. [41, 42]), and Reddit studied jointly with Twitter, YouTube 5% of contributions (e.g. [5])). Finally, a single paper considers combination of four datasets (i.e. Reddit, Twitter, Facebook, and Gab [7]).

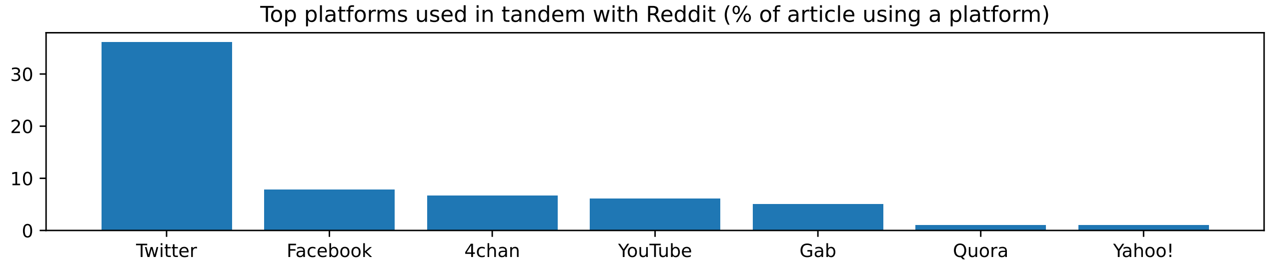

Fig. 4.8. *Online platforms used as data sources together with Reddit*

An interesting use case of Reddit usage in scientific environment has been found in “IEEE Top Programming Languages: Design, Methods, and Data Sources”[10]. This work shows a practical approach to a an interesting research question; here, what are the top programming languages. In this work Reddit is listed as one of the sources among others, such as Google Trends, Twitter, GitHub and Stack Overflow.

4.3. Linguistic analysis. During exploratory data analysis, various natural language processing techniques were applied. Among them, papers were also analysed linguistically. Specifically, sentiment analysis using NLTK framework [26] and SentimentAnalyzer[11] was applied. Observed polarization (depicted in Figure 4.9) indicates a negligible displacement towards the positive sentiment. This was expected, and is consistent with previous studies on scientific literature sentiment [19].

[9] https://www.k-cap.org/kcap11/index.html

[10] https://spectrum.ieee.org/ieee-top-programming-languages-design-methods-and-data-sources

[11] https://www.nltk.org/api/nltk.sentiment.sentiment_analyzer.html

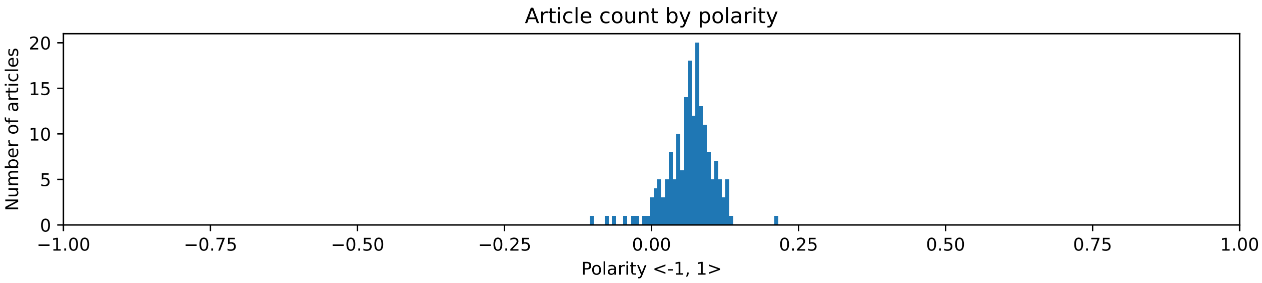

Fig. 4.9. *Histogram of number articles based on text polarity measures*

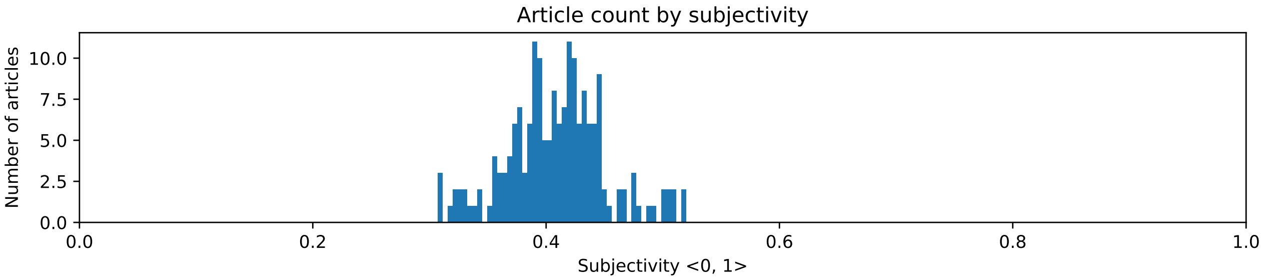

Fig. 4.10. *Histogram of number articles based on text subjectivity measures*

However, the subjectivity measure (summarised in Figure 4.10) raised concerns. Obviously, it has been claimed that scientific research may be subjective, as it needs to allow "leaps of faith" (see, [9]). Moreover, some philosophers [30, 29] argue that subjectivity is intrinsic for human nature. However, it is also claimed (and for good reasons) that the foundation of the scientific method [32] revolves around aiming at objectivity. Hence, results summarised in Figure 4.10, indicating high level of subjectivity, were somewhat concerning. To establish the reason for this finding, the most "subjective" texts were studied directly. As a result is was found that this is a false alarm. Specifically, apparent shift towards subjectivity was caused by inaccuracy of the classifier (*SentimentIntensityAnalyzer* from *nltk.sentiment*[12]). For further understanding, let us consider the selected sentences from the most subjective (according to the NLTK metric) articles.

- "However, this openness formed a platform for the polarization of opinions and controversial discussions" [22] (score: 0.95)
- "(...) also presented an extended version of the study discussing potential racial bias in offensive content datasets (...)" [2] (score: 1.0)
- "All datasets only contain activity between 01/2015 and 10/2018" [16] (score: 1.0)

Moreover, let us also consider how the calculated subjectivity measure changes with a simple modification of selected statements (i.e. by removing particular words):

- Statement before transformation (score: 0.63):
  "Controversially initiated and non-controversially initiated cascades, (a,b,c) are controversially initiated posts' cascades while (d,e,f) are non-controversial posts' cascades where the red dots represent a comment labeled as controversial by Reddit that is directed to the post's author while a green dot is a comment labeled controversial by Reddit that is directed to another comment." [22]
- The same statement after transformation (score: 0.15):
  "initiated and initiated cascades, (a,b,c) are initiated posts' cascades while (d,e,f) are posts' cascades where the red dots represent a comment labeled as by Reddit that is directed to the post's author while a green dot is a comment labeled by Reddit that is directed to another comment." [22]

[12] `https://www.nltk.org/api/nltk.sentiment.html`

This suggests that simply using the "subjective" (key)words (e.g. "controversial", "bias") in the text, regardless of their context, results in radically increased value of the variable that is to indicate subjectivity of the text. However, there are sentences that do not use such words, which have also received a high subjectivity score. Hence, further research would be required into the way that the NLTK metric works and why, sometimes, it is rather misleading. However, this is outside of scope of the current contribution.

4.4. Reddit-based literature in scholarly databases. Let us now address **RQ3** and **RQ4**. Even though they cannot be unequivocally answered, possible answers can be experimentally explored. To verify the change over time of the number of scholarly papers related to Reddit, between 2010 and 2021, 10 databases have been analysed and queried for the term "reddit". As shown in Figure 4.11 the number of found articles raises year to year (**RQ3**).

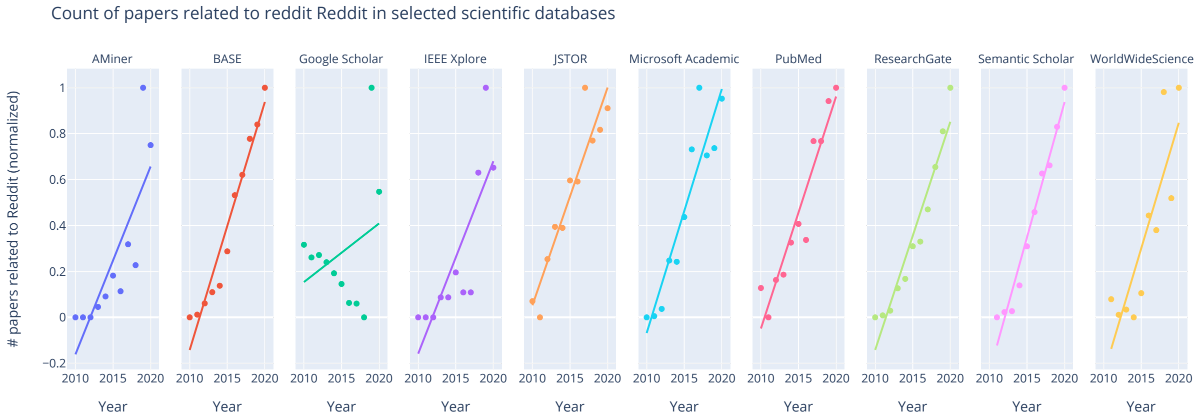

Fig. 4.11. *Non-cumulative count of papers related to Reddit in scientific databases in years 2010-2021.*

Table 4.1 shows how many articles, related to Reddit (e.g. using it as a data source, processing it, analysing the comments, etc.), have been indexed in scientific databases.

Table 4.1

Number of scientific papers (which appeared in 2011-2021) related to Reddit (e.g. using it as dataset, exploring its structure etc.) indexed in selected databases (all accessed on 09-06-2021).

Database	Total size	# papers
Google Scholar	330M [15]	980[13]
JSTOR	12M[14]	2555[15]
PubMed	32M[16]	243[17]
AMiner	230M [15]	840[18]
Bielefeld Academic Search Engine	270M[19]	5176[20]
Semantic Scholar	197M[21]	2280[22]
AMiner	320M[23]	784[24]
Microsoft Academic	170M [15]	902[25]
WorldWideScience	300M [15]	1250[26]
IEEE Xplore	8551[27]	162[28]
ResearchGate	135M[29]	3001[30]

Observations that can be made, on the basis of the results found in Table 4.1, are:

- the number of Reddit articles is quite small, yet representative,
- the number of Reddit papers is somewhat proportional in each database, so it can be stated that the literature is quite equally spread in the Internet.

4.4.1. Outlying results found in Google Scholar. The only database with trends inconsistent with others is Google Scholar (Figure 4.11). However, although it is one of the most widely known databases that indexes scientific publications [27, 17], it already received both praise and criticism [20, 39]. Main problems of Google Scholar, pointed out in the literature, are: (i) difficulties to estimate the actual size of the database [33, 21], (2) gender- and race-related bias in displaying contributions [23], (iii) favoring incremental work [23], (iv) favoring larger research communities [23], (v) limited indexing of files [20], (vi) incorrect bibliometrics (due to automated algorithms instead of skilled librarians) [28, 14], (vi) "uncertain quality of Google Scholar's performance" [14], (vii) "Google Scholar's inability or unwillingness to elaborate on what documents its system crawls" [14], and (viii) limitations of bibliometric analysis [28]. Moreover, Google Scholar declares inconsistently the number of results of a query, and the actual number of returned results (e.g. a query returns 1000 actual results, while it declares 58,600[31]). This finding may correspond to already reported Google Scholar inconsistencies [33, 21] and lack of transparency [14]. Therefore, Google Scholar can be treated as an outlier and disregarded in conclusions drawn from this experiment.

4.5. Google Trends. The next experiment explored presence of popular trends in Reddit. This was done based on Google Trends, an analytical website which provides information about popularity of search queries in Google search engine [32]. For all Global Google Trends 2020[33] their Reddit presence has been measured (see Table 4.2). Overall, 79% of top Google Trends have a dedicated subreddit, while *all of them* are widely discussed. Table 4.2 illustrates top three in each Google Trend category.

5. Reddit as "The Ultimate Dataset for Everything". To further study whether Reddit contains information about (almost) "any area", a tool for easy exploratory data analysis (EDA [8]) was designed. Specifically, *Reddit-TUDFE* allows quick search of any topic on Reddit, checking if/how it is represented, and how it is discussed. Specifically, *Reddit-TUDFE* delivers the following functions:

1. Uses Reddit API to search for best matching subreddit.
2. Downloads newest N posts from the subreddit, using Pushshift API and a combination of PRAW[36] and PSAW[37].
3. Performs basic text cleaning (tokenization with NLTK [26], removal of stopwords, punctuation, numbers).
4. Generates and displays post titles and content wordclouds[38].

The code follows state-of-the-art solutions for code sharing ([35]) and is publicly available on GitHub[39] as a Jupyter Notebook [37].

To illustrate the capabilities of the developed application, let us present few examples, in two groups, in Figures 5.2 and 5.1. The wordclouds are build from posts related to a subreddit dedicated (or closest) to the searched topic. *Reddit-TUDFE* allows to quickly check if, and how, a particular topic is covered. Note that similar examples can be derived for any other topic, while Reddit also shows potential in, for instance, building ontologies, or semantic graphs. However, this possibility is out of scope of this contribution.

In Figure 5.1:

- Left subfigure shows result for the phrase "music", a generic term, which is certainly discussed on Reddit. One may see particular genres: rock, pop, rap, relaxing, electronic, etc.
- Middle subfigure displays results for phrase "rock", a bit narrowed, but still vague music-related (sub)topic, which is also present in Reddit, including artists/bands like: Rolling Stones, AC/DC, Led Zeppeling, Queen, Pink etc.
- Right subfigure contains a strictly specific topic, i.e. the band "The Beatles", which is also widely

[31] https://scholar.google.com/scholar?start=990&q=reddit&hl=en&as_sdt=0,5&as_ylo=2020&as_yhi=2020 accessed on 11-09-2021

[32] https://trends.google.com/trends

[33] https://trends.google.com/trends/yis/2020/GLOBAL/

[36] https://github.com/praw-dev/praw

[37] https://github.com/dmarx/psaw

[38] https://github.com/amueller/word_cloud

[39] https://anonymous.4open.science/r/reddit-tudfe-B736/reddit_tudfe.ipynb

https://anonymous.4open.science/r/reddit-tudfe-B736/reddit_tudfe.ipynb

Table 4.2
Global Google Trends 2020[35] *(top 3 in each Google Trends category) and their appearance on Reddit ("subreddit" – there exists a dedicated subforum, "discussion" – the topic is present in (a) subreddit(s) of a broader topic)*

Google Trend	category	on Reddit	reference
Coronavirus	searches	subreddit	r/Coronavirus
Election results	searches	discussion	r/politics
Kobe Bryant	searches	subreddit	r/kobebryant
Tom Hanks	actors	subreddit	r/tomhanks
Joaquin Phoenix	actors	subreddit	r/joaquinphoenix
Amitabh Bachchan	actors	subreddit	r/india
Ryan Newman	athletes	subreddit	r/RyanNewman
Michael Jordan	athletes	subreddit	r/michaeljordan
Tyson Fury	athletes	subreddit	r/TysonFury
Parasite	movies	subreddit	r/parasite
1917	movies	subreddit	r/1917
Black Panther	movies	subreddit	r/blackpanther
Tiger King	tv shows	subreddit	r/TigerKing
Big Brother Brasil	tv shows	subreddit	r/BigBrotherBrasil
Money Heist	tv shows	subreddit	r/MoneyHeist
Joe Biden	people	subreddit	r/JoeBiden
Kim Jong Un	people	subreddit	r/kimjongun
Boris Johnson	people	subreddit	r/BorisJohnson
Coronavirus	news	subreddit	r/Coronavirus
Election results	news	discussion	r/politics
Iran	news	subreddit	r/iran
Among Us	games	subreddit	r/AmongUs
Fall Guys: Ultimate Knockout	games	subreddit	r/FallGuysGame
Valorant	games	subreddit	r/VALORANT
Dalgona coffee	recipes	discussion	r/cafe
Ekmek	recipes	discussion	r/Breadit
Sourdough bread	recipes	subreddit	r/SourdoughBread
Kobe Bryant	loss	subreddit	r/kobebryant
Naya Rivera	loss	subreddit	r/NayaRivera
Chadwick Boseman	loss	subreddit	r/ChadwickBoseman

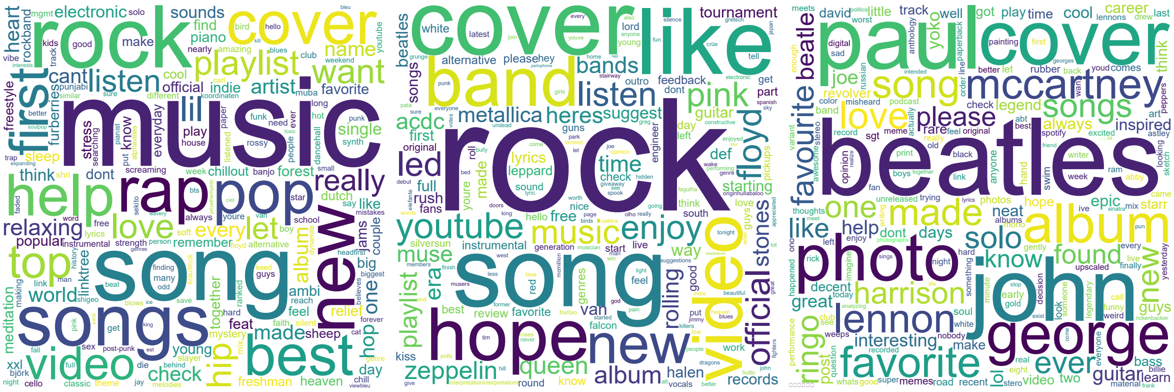

Fig. 5.1. *Exemplary wordclouds of 200 posts (before 01-09-2021) concerning (top to bottom): "music", "rock" and "The Beatles".*

covered on Reddit. Here one may see, among others, individual band members: John Lennon, Paul McCartney, Ringo Starr, and George Harrison.

Another example is summarized in Figure 5.2.

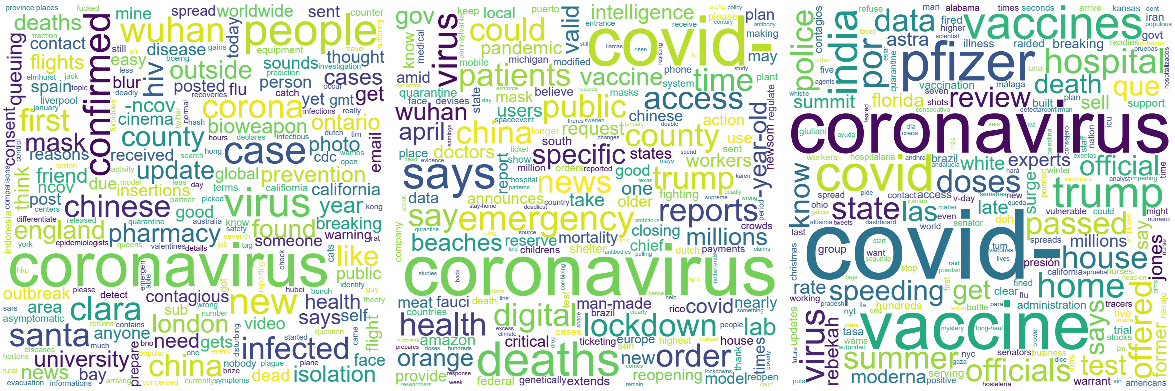

Fig. 5.2. *Word clouds of 100 posts title from subreddit r/Coronavirus*[41] *at different times during COVID-19 pandemic (left to right: 01-02-2020, 01-05-2020 and 08-12-2020)*

In this figure, one can notice a clear shift of focus on the subreddit r/Coronavirus at different times of COVID-19 pandemic [1] (phrase "coronavirus" is skipped). Figure 5.2 (left) (before 01-02-2020) shows that the main interest concerned the phrases a.o.: "confirmed" (the number of confirmed infections), "Wuhan" and "Chinese" (the geographical origins of the fist reported infections[25]). Figure 5.2 (middle) (before 01-05-2020) displays that the main phrases changed to: "deaths" (due to COVID-19 infection) and "lockdown" (the preventive measures against the spread of the virus). Figure 5.2 (right) (before 08-12-2020, i.e. near the first vaccine invention) shows the general interests in phrases like: "vaccine" and "pfizer" (the company to invent the vaccine [3]).

Note that analysing the evolution of thematic ecosystem is just one of possible applications of the *Reddit–TUDFE* tool. Most importantly, it quickly allows checking whether given topical domain contains live (evolving over time) information.

6. Concluding remarks. This work provides evidence that Reddit is a robust, but underutilized, resource for information retrieval and knowledge capture, in almost any field of interest. Based on performed exploratory analysis, the following answers to the research questions formulated at the beginning of this work can be stipulated:

- **RQ1**: Reddit offers publicly available data, which can be easily retrieved with Pushshift API.
- **RQ2**: Most popular techniques for Reddit information processing are: text embeddings, neural networks, and graph networks.
- **RQ3**: Reddit is trending in scientific research as more and more articles using it are published every year.
- **RQ4**: Reddit covers the majority (79%) of topics that appear in Global Google Trends, sustaining the claim that Reddit is a robust source of knowledge about "everything trendy".
- **RQ5**: Reddit is most commonly used in tandem with Twitter.

These conclusions render Reddit a perfect candidate for future research – especially the presence of graph networks among common research methods and high coverage of popular trends. Finally, this analysis and the *Reddit–TUDFE* tool provide solid foundation for future research on Reddit and its potential in information retrieval.

Acknowledgement. This work has been supported in part by the joint research project "Novel methods for development of distributed systems" under the agreement on scientific cooperation between the Polish

[41] https://www.reddit.com/r/Coronavirus/

Academy of Sciences and Romanian Academy.

REFERENCES

[1] A. Abd-Alrazaq, J. Schneider, B. Mifsud, T. Alam, M. Househ, M. Hamdi, and Z. Shah, *A comprehensive overview of the covid-19 literature: Machine learning-based bibliometric analysis*, Journal of medical Internet research, 23 (2021), p. e23703.

[2] K. Aggarwal, P. Bamdev, D. Mahata, R. R. Shah, P. Kumaraguru, et al., *Trawling for trolling: A dataset*, arXiv preprint arXiv:2008.00525, (2020).

[3] A. Badiani, J. Patel, K. Ziolkowski, and F. Nielsen, *Pfizer: The miracle vaccine for covid-19?*, Public Health in Practice, 1 (2020), p. 100061.

[4] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, *The pushshift reddit dataset*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 830–839.

[5] C. Buntain, R. Bonneau, J. Nagler, and J. A. Tucker, *Youtube recommendations and effects on sharing across online social platforms*, Proceedings of the ACM on Human-Computer Interaction, 5 (2021), pp. 1–26.

[6] L. Cao, H. Zhang, and L. Feng, *Building and using personal knowledge graph to improve suicidal ideation detection on social media*, IEEE Transactions on Multimedia, (2020).

[7] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, *Echo chambers on social media: A comparative analysis*, arXiv preprint arXiv:2004.09603, (2020).

[8] V. Cox, *Exploratory data analysis*, in Translating Statistics to Make Decisions, Springer, 2017, pp. 47–74.

[9] A. Curtis, *The science of subjectivity*, Geology, 40 (2012), pp. 95–96.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).

[11] A. Edelbo Lillie and E. Refsgaard Middelboe, *Danish stance classification and rumour resolution*, arXiv e-prints, (2019), pp. arXiv–1907.

[12] M. Fajcik, L. Burget, and P. Smrz, *But-fit at semeval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers*, arXiv preprint arXiv:1902.10126, (2019).

[13] I. Garibay, T. A. Oghaz, N. Yousefi, E. C. Mutlu, M. Schiappa, S. Scheinert, G. C. Anagnostopoulos, C. Bouwens, S. M. Fiore, A. Mantzaris, et al., *Deep agent: Studying the dynamics of information spread and evolution in social networks*, arXiv preprint arXiv:2003.11611, (2020).

[14] J. E. Gray, M. C. Hamilton, A. Hauser, M. M. Janz, J. P. Peters, and F. Taggart, *Scholarish: Google scholar and its value to the sciences*, Issues in Science and Technology Librarianship, 70 (2012).

[15] M. Gusenbauer, *Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases*, Scientometrics, 118 (2019), pp. 177–214.

[16] H. Habib, M. B. Musa, F. Zaffar, and R. Nithyanand, *To act or react: Investigating proactive strategies for online community moderation*, arXiv preprint arXiv:1906.11932, (2019).

[17] G. Halevi, H. Moed, and J. Bar-Ilan, *Suitability of google scholar as a source of scientific information and as a source of data for scientific evaluation—review of the literature*, Journal of informetrics, 11 (2017), pp. 823–834.

[18] J. Hastings, O. Kutz, and T. Mossakowski, *How to model the shapes of molecules? combining topology and ontology using heterogeneous specifications*, in In Proc. of the Deep Knowledge Representation Challenge Workshop (DKR-11), K-CAP-11, Citeseer, 2011.

[19] D. M. E.-D. M. Hussein, *Analyzing scientific papers based on sentiment analysis*, Information System Department Faculty of Computers and Information Cairo University, Egypt, (2016).

[20] P. Jacsó, *Google scholar: the pros and the cons*, Online information review, (2005).

[21] P. Jacsó, *Google scholar revisited*, Online information review, (2008).

[22] J. Jasser, I. Garibay, S. Scheinert, and A. V. Mantzaris, *Controversial information spreads faster and further in reddit*, arXiv preprint arXiv:2006.13991, (2020).

[23] F. R. Jensenius, M. Htun, D. J. Samuels, D. A. Singer, A. Lawrence, and M. Chwe, *Benefits and pitfalls of google scholar*, PS: Political Science and Politics, (2018).

[24] V. Kharde, P. Sonawane, et al., *Sentiment analysis of twitter data: a survey of techniques*, arXiv preprint arXiv:1601.06971, (2016).

[25] K. Leung, J. T. Wu, D. Liu, and G. M. Leung, *First-wave covid-19 transmissibility and severity in china outside hubei after control measures, and second-wave scenario planning: a modelling impact assessment*, The Lancet, 395 (2020), pp. 1382–1393.

[26] E. Loper and S. Bird, *Nltk: The natural language toolkit*, arXiv preprint cs/0205028, (2002).

[27] E. D. López-Cózar, E. Orduña-Malea, and A. Martín-Martín, *Google scholar as a data source for research assessment*, in Springer handbook of science and technology indicators, Springer, 2019, pp. 95–127.

[28] E. D. López-Cózar, E. Orduña-Malea, A. Martín-Martín, and J. M. Ayllón, *Google scholar: the big data bibliographic tool*, Research analytics: boosting university productivity and competitiveness through scientometrics, (2017), p. 59.

[29] F. MacKellar, *Subjectivity in qualitative research*, EDUC 867 WEBSITE, (2012).

[30] M. A. Mannan, *Science and subjectivity: Understanding objectivity of scientific knowledge*, Philosophy and Progress, (2016), pp. 43–72.

[31] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, *The anatomy of reddit: An overview of academic research*, in Dynamics on and of Complex Networks, Springer, 2017, pp. 183–204.

[32] I. Newton, *Philosophiae naturalis principia mathematica*, vol. 2, typis A. et JM Duncan, 1833.

[33] E. Orduña-Malea, J. M. Ayllón, A. Martín-Martín, and E. D. López-Cózar, *Methods for estimating the size of google scholar*, Scientometrics, 104 (2015), pp. 931–949.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[35] J. M. Perkel, *Why jupyter is data scientists' computational notebook of choice.*, Nature, 563 (2018), pp. 145–147.

[36] A. Rajaraman and J. D. Ullman, *Data Mining*, Cambridge University Press, 2011, p. 1–17, https://doi.org/10.1017/CBO9781139058452.002.

[37] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, *Using the jupyter notebook as a tool for open science: An empirical study*, in 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2017, pp. 1–2.

[38] Y. Shinyama, *Pdfminer: Python pdf parser and analyzer*, Retrieved on, 11 (2015).

[39] M. Shultz, *Comparing test searches in pubmed and google scholar*, Journal of the Medical Library Association: JMLA, 95 (2007), p. 442.

[40] G. Viglione, *How scientific conferences will survive the coronavirus shock.*, Nature, 582 (2020), pp. 166–168.

[41] S. Zannettou, *Towards understanding the information ecosystem through the lens of multiple web communities*, arXiv preprint arXiv:1911.10517, (2019).

[42] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, *Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web*, in Companion proceedings of the 2019 world wide web conference, 2019, pp. 218–226.

[43] H. Zhang, Z. Liu, C. Xiong, and Z. Liu, *Grounded conversation generation as guided traverses in commonsense knowledge graphs*, arXiv preprint arXiv:1911.02707, (2019).

Edited by: Dana Petcu
Received: Dec 22, 2021
Accepted: Apr 1, 2022