The presence of SARS-CoV-2 in wastewater treatment: A bibliometric analysis approach

Miguel A. Dias[1], Daniela P. Mesquita[2], Eugénio C. Ferreira3 and Anália Lourenço[4]

[1] Department of Informatics, University of Minho, 4710-057 Braga, Portugal
pg40968@alunos.uminho.pt
[2] BIOSYSTEMS, Department of Biological Enginering, University of Minho, 4710-057 Braga, Portugal
daniela@deb.uminho.pt
[3] BIOSYSTEMS, Department of Biological Enginering, University of Minho, 4710-057 Braga, Portugal
ecferreira@deb.uminho.pt
[4] BIOSYSTEMS, Department of Biological Enginering, University of Minho, 4710-057 Braga, Portugal
analia@ceb.uminho.pt

Abstract. In this preliminary work of the curricular unit of project in Bioinformatics, a summary of the state of the art on the theme of bibliometric analysis on wastewater covid-19 data is presented.
As the objective of this is to use available online software to give an overview of the tracking of covid-19 viral data from past articles as well as some of the existing packages available for the python language to process the data and use for natural language processing.

Keywords: Covid-19, SARS-COV2, qRT-PCR, RNA

1 Introduction

On December 2019, an outbreak caused by a new human pathogen, lately named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was detected in Wuhan, China. The virus caused a severe respiratory syndrome, generically defined as coronavirus disease (COVID-19). Since the transmission of SARS-CoV-2 was found to be due to its presence in the respiratory tract of infected people, the disease quickly spread globally thanks to its high rate of infectivity and the presence of a huge number of asymptomatic or minimally symptomatic patients. In March 2020, the World Health Organization declared a pandemic state. SARS-CoV-2 has been also detected in the feces and urine of infected people. Thus, the need of detecting and quantifying the number of viral particles and the pathogenicity in wastewater and sewage sludge produced by hospitals and houses with infected people, being a possible source of contamination, has been considered of utmost importance. Also, it has been reported the need of identifying the best strategies for treating wastewater and managing sewage sludge.

1.1 Objective

In the present project, a bibliometric analysis will be performed to evaluate the trends in SARSCOV-2 research since the outbreak. To undertake this analysis, the following information categories will be used to analyze publication patterns in the published literature: document type, languages, categories of articles, journals in previously published articles, the countries, and territories from which the publications emerged, the institutions that had published the articles, and h-index. The purpose of this study is to present a bibliometric overview of current SARS-COV-2 research and provides useful advice mainly on possible exposure risks.

2 State of the art

Covid-19 has been a hot topic in the news ever since the first cases of this disease, caused by a novel strain of severe acute respiratory corona virus (SARS-COV2), was reported in Wuhan around December 2019 (WHO, 2020). Since then, this pathogen as spread throughout the entire world and as of the time of the writing of this project is still considered a global pandemic by the World Health Organization.

This is a respiratory disease similar to an influenza infection with symptoms that include dry cough, fever, headaches and some less commonly seen symptoms like diarrhea, nausea, loss of taste and/or smell. The more severe version of this illness requires immediate medical assistance if the person is having shortness of breath or chest pain and is more likely to occur the older the individual is or if the person has some kind of chronic respiratory disease such as bronchitis or asthma. Other individuals at risk include those that are overweight or with a repressed immune system.

Although less deadly in comparison to 2003 SARS-COV strain (fatality rate around 1-3%, lower than 2003 SARS's 11% (Bialek et al., 2020)) what makes this virus so problematic is It's higher infectivity, mainly due to the fact that most people get the mild version of this disease while unknowingly spreading to other, potentially more vulnerable, individuals. According to (Petersen et al., 2020) this new coronavirus strain, in comparison to past influenza pandemics and the SARS 2003 strain, has a longer incubation period that ranges from 4 to 12 days as well as having the highest R0 value of all 4 viruses at the time of the article's writing. It can be spread by asymptomatic individuals and since they make a significant portion of around 30% (Gerrity et al., 2021; Nishiura et al., 2020) of the overall infected population they can be a cause of concern when It comes to dealing with future infections.

As the pandemic evolves many measures have been taken to reduce the propagation of this virus, mostly by enforcing everyone to use a mask, practice social distancing, city wide lockdowns for when the number of cases get to out of control, and more recently vaccination. One year later since covid-19 struck the world and even after all these measures were imposed the pandemic still seems far from over. One of the many challenges that comes from fighting this disease is identifying the presence of SARS-

COV2 in a population and quarantine those infected. As already mentioned, asymptomatic individuals can still spread the disease, controlling the pandemic can't rely solely on waiting for clinical symptoms to appear.

Detection of viral RNA from a person uses a molecular diagnostic test from a nasal swab sample, the most commonly used analytical method for this purpose is usually a Reverse transcription polymerase chain reaction (RT-PCR) due to Its relatively quick testing results and high sensitivity (Shen et al., 2020)even when diagnosing early on during an infection incubation period (Tahamtan & Ardebili, 2020). However, there's Is still the issue of potential dubious results since false positives and false negatives can still occur. The concern over false results is an important thing to note and false negatives are particularly bad since individuals with these results may unknowingly relax when It comes to safety measures designed to stop the spread of covid or medical staff may be put into quarantine when they could be working if they get a false positive result (West et al., 2020). Therefore, there is an ever-growing need for faster, more accurate and earlier covid-19 viral load detection methods.

Water-based epidemiology is an example of a complimentary approach for detection of covid-19 that has been implemented worldwide in an effort to fight this disease (Ahmed et al., 2021). Recent studies have shown that not only can SARS-COV2 show up in the stool of infected individuals, it’s also present in a larger time window (9 to 16 days) in comparison to nasal swab testing (6 to 11 days) (Ling et al., 2020)which means an individual could have the disease even if a qRT-PCR test turns out negative. Measuring the level of the total viral RNA in feces is a way to monitor the overall infection rate among a given population, results from these tests have been used to detect the presence of covid-19 days before the first official case was reported in many countries around the world (Ahmed et al., 2021) proving to be very effective in predicting the spread of covid-19.

In order to carry out a study of this magnitude for the coverage of covid-19 throughout the world there’s a need to accumulate information from past research with proven reliable results. Looking through every article on the subject can be tiresome, automating this process with a bibliometric approach will undoubtedly facilitate this process.

2.1 Bibliometric analysis

Bibliometrics can be defined as the study of books, papers and other publications by applying statistical methods to evaluate patterns in previously published literature like the number of scientific articles relating to a subject, the authors of said work and where the articles were published (journals, institution and country of origin) or how many times they were cited by other authors (Jones, 2015) as well as the corresponding h-index value for a given scientist. This last parameter is a citation metric invented by Physicist Jorge Hirsch (Jones, 2015) for evaluating the reputability of a scientist's work

by sorting all the articles of a scientist by descending order and giving them a number, being that the article with the most citations is number one on the list followed by an incremental value until the last article with the least citations. The h-index value is the last article whose number is equal to or greater than the number of citations (Jones, 2015). The greater the h-value the higher the quality of a person's work is assumed to be.

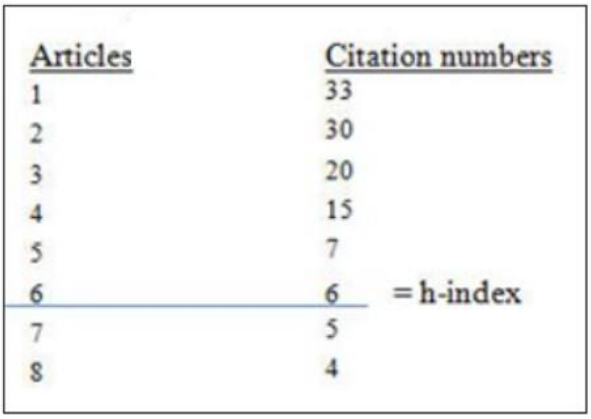

Figure 1: *Example of an author's h-index calculation taken from https://subjectguides.uwaterloo.ca/calculate-academic-footprint/YourHIndex. The value is 6 in this case since It was the last article with a number of citations equal to or greater than the number for the position of the article in the list.*

The h-index approach can be useful to filter out bad academic reports since It values authors with various publications and multiple citations, even if an author has a lot of articles if they don't get a lot of citations (even if one or two articles have a huge amount) It will still result in a relatively lower h-index value.

Using text mining techniques can also be great for obtaining more information on each article, like the sentiment analysis of an article (If It's positive, neutral, or negative relating to a subject) or how many times a given word was mentioned in an article or even to be able to figure out what the topic of a given text is. Text mining uses natural language processing for making a computer understand text by making a simulation of how a person can understand a natural speaking language (Hao et al., 2018).

3 Methods and Results

For this study, scopus.com was the main database used to extract a dataset containing articles relating to covid-19 presence in wastewater. Using the key words "covid-19" and "wastewater" the scopus search resulted in 1978 articles as of the time of the writing of this article. Adding more selective terms related to wastewater treatment resulted in fewer articles so to get the broadest range of articles possible only two words were used.

All the information related to this dataset as well as the analysis can be checked on the following repository: https://github.com/MiguelAndreDias/Projecto_Bioinformtica/tree/main

3.1 Search result analysis:

Using the "Analysis Search results" button on the scopus.com database after a search result, important bibliometric information was extracted from this dataset. Information related to the year and country the articles were published, the type of documents, the authors, the subject area of the work and the institutions that produced the work.

The main giveaway from this analysis is that most of the articles were published in the year 2021 (1395 articles, in comparison 2020 only produced 593), the top 3 countries with the most published work where the United States, China and India in that respective order and almost all the articles were written in English. The institution that produced the most literary work was the Chinese Academy of Sciences followed very closely by the Ministry of Education China.

When it comes to the subject area 25,7% of the articles or 508 articles were related to environmental Sciences as seen in figure 2. This is the main subject area of interest when it comes to dealing with wastewater-based studies in terms of presence of covid-19 in water. Going back to the scopus.com database with a more restrictive search with the keywords of "covid-19" and "wastewater-based epidemiology" It returned 386 articles which seems to go somewhat in accordance with the previous search results in the broader dataset.

As an added information It was possible to extract the most cited articles being that the most cited article "First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community" was written by the author with the placed third with the most publications, Ahmed W.

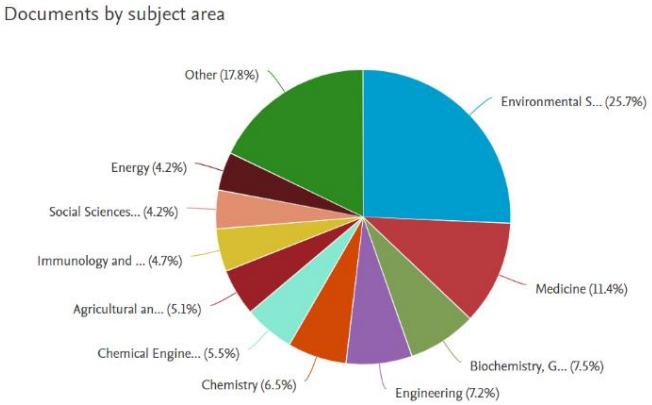

Figure 2: *Piechart of the subject area for the search results from scopus.com*

3.2 Vosviewer

Using the vosviewer software It's possible to create powerful bibliometric networks that can be used to visualize the relationship between words and It's also possible to trace relationships between researchers with similar work or that worked together. These models can be done by checking bibliographic information about each article and its respective authors.

The co-authorship option in the create new map in the software, as Its name implies, groups authors by how similar their articles are. For this analysis the minimum number of documents an author had to have its name in was 10 and, of the 138 authors that remained after this filter, only 29 researchers were chosen since they had the greatest co-authorship link strength. The results of this network can be seen on figure 3.

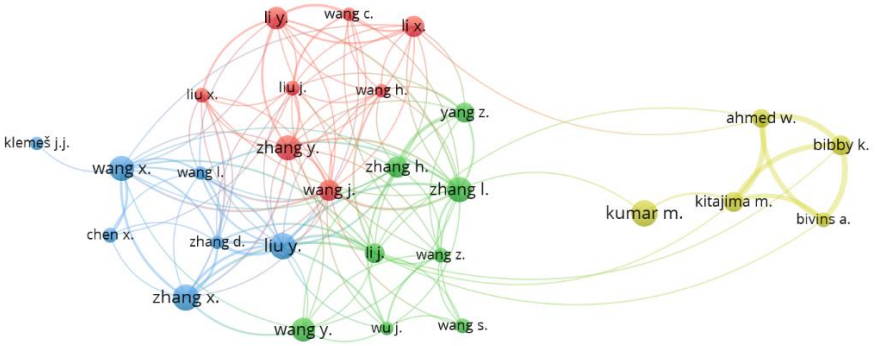

Figure 3: *Co-authorship model. Clusters of same color mean the authors have similar work and bigger nodes mean that the authors were cited more times.*

Each author has a node associated and the size of this node indicates the number of times the author has been cited, being larger if number of citations is higher. The nodes are grouped by clusters of different colors, authors with similar work or that were co-authors are grouped together more closely and have the same color. In the yellow-colored cluster, all the authors made articles related to wastewater covid-19 detection and that can be checked on the scopus search result top 10 authors.

In figure 4 there's a visual representation of all the words most commonly seen in the all the articles. This time a co-occurrence model was created and words that were seen more times in the same articles were grouped closer together and bigger nodes mean that the words were used more times. The words were filtered by a minimum of occurrences of at least 15 times. Of the around 15000 words only 387 were put on the model. As to be expected words like covid-19, pandemic or coronavirus were the most used words on the map created in figure 4.

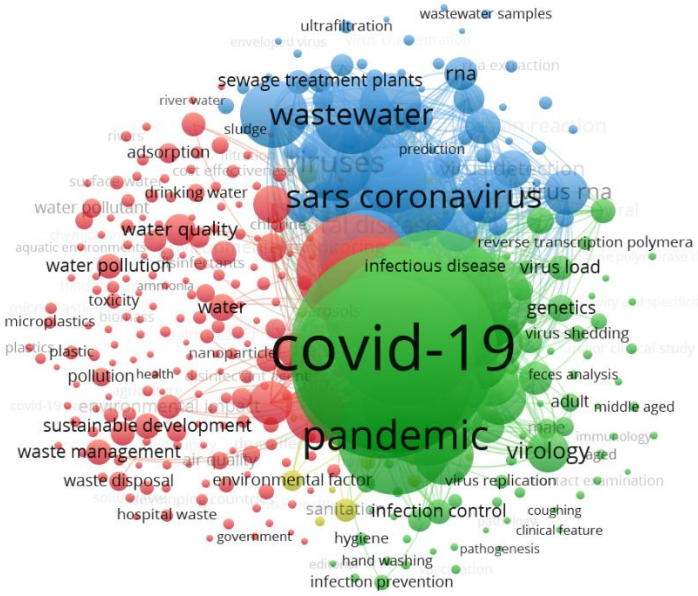

Figure 4: *Co-occurrence model. Words used in the same articles were grouped closer together or put in the same-colored clusters. Covid-19 was the word most used as can be seen by the size of the node.*

It's possible to see that the blue cluster contains words most closely related to wastewater treatment related to epidemiology. The red cluster also has words that relate to blue but It seems the topic is more about environment pollution instead of wastewater epidemiology while the green cluster is more about the virus itself and Its clinical effects on people or measurements needed to fight It.

3.3 Topic modeling

For this analysis all the abstracts related to each article were collected in a column in a single csv. Using the text data present in this column a statistical model was constructed using Latent Dirichlet Allocation (LDA), a very popular algorithm used when dealing with text data analysis. The gensim and sklearn python packages both offer very streamlined ways of constructing reliable statistical models which can be used for topic modeling.

With the gensim package the analysis started by loading the main dataset followed by preprocessing of the abstract data so that all the stopwords (commonly used adverbs, verbs or pronouns) were removed and words were reduced to their canonical form by a process called lemmatization. Only after this process was done the LDA model could

be constructed. To keep It relatively simple the LDA model was constructed by only clustering words to find 7 different topics. The results can be seen in figure 5.

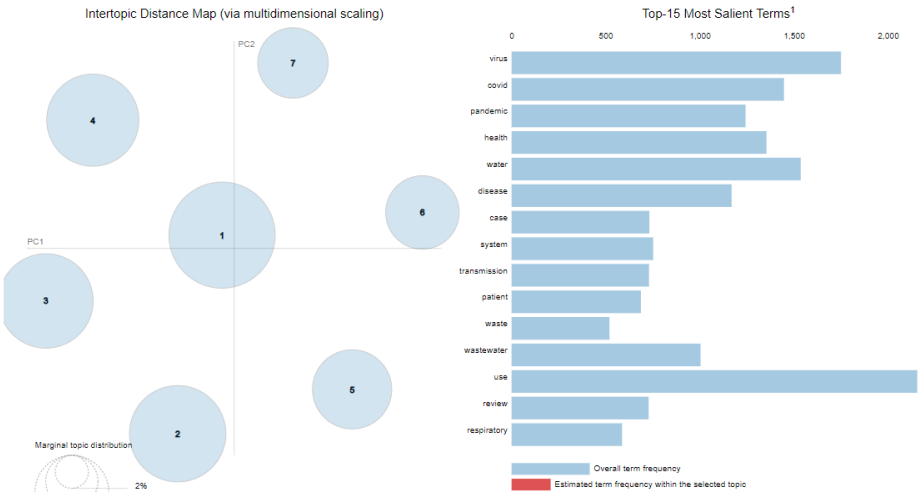

Figure 5*:Visual representation of the topic model produced with the gensim package. On the left there's a representation of each topic by cluster and on the right the 15 words most commonly seen across all topics.*

As to be expected the most salient terms are words related to virology, health or wastewater. Not all articles are about the study of covid-19 wastewater epidemiology, if we go over one of the clusters some words are more represented in each article abstract in comparison to others. In figure 6 it's possible to see the words that are more present for topic number 3.

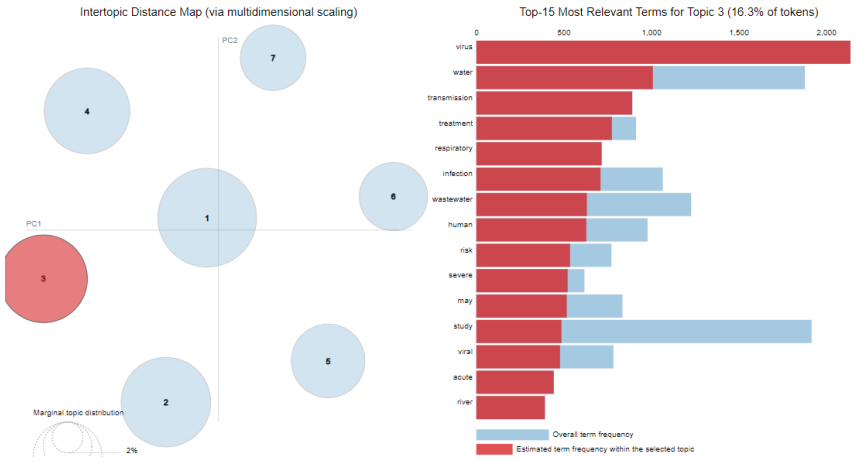

Figure 6: *Graphic with the top 15 most relevant terms for topic 3 represented in red.*

Of all the topic clusters, topic 3 seems to be the most likely to be of interest for this study due to having the terms “water”, “wastewater”, “river” in the most represented words. Other clusters only have one or none of these previous terms so it can be assumed that they aren’t about wastewater epidemiology.

Atempting to make a topic model using the sklearn package is a slightly different methodology but it results in similar results. This model was made by discarding words that were present in over 95% of the articles and counting words that had to show up in atleast two documents. The most notable difference in the model cluster is that two of the topics (clusters 3 and 6) overlap each other. With this sklearn model the most likely topic number was associated for each article abstract and the number of articles was counted for each topic. The results are shown in figure 7.

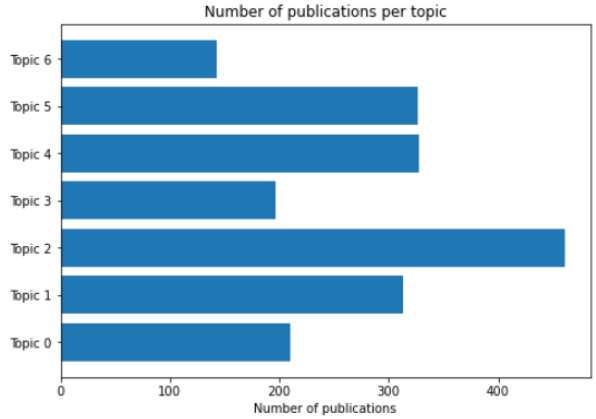

Figure 7: *Number of publications for each topic cluster.*

3.4 Sentiment Analysis

For this part of the work, a sentiment analysis was performed for each article in the csv file and for the author with the biggest h-index value.

Using the nltk package It was possible to give a sentiment analysis score for each and every article abstract, articles that score over zero were given a positive sentiment while articles with a negative score less than zero were on the other hand given a negative sentiment. The results showed that of the 1978 articles in the csv file, 1481 were given a positive sentiment and 497 were deemed to be negative in sentiment. The highest h-index calculated was for author Kitagima, M. and has value of 11. The H-index for the top 10 authors with the most publications can be seen in the plot represented below in figure 8.

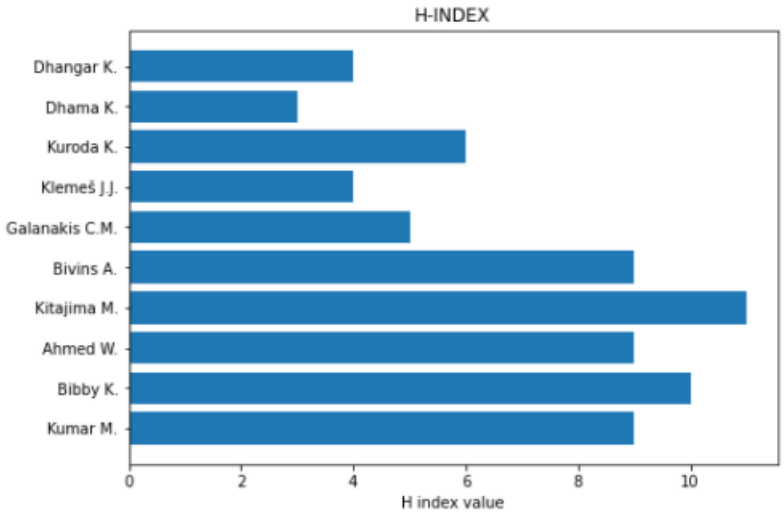

Figure 8*: H-index value calculated for the top 10 authors with the most citations for this search on scopus.org.*

It should be noted that although the nltk package is a very powerful tool to analyze texts, due to the nature of how academic reports are written It can be hard to figure out the overall sentiment of a text. Also, because an abstract text is a very summarized version of the contents of an article It may not give the entire opinion of how the study went through. Usually, the discussion or the conclusions part of an article is the one that can give out the best overall sentiment of an article.

Because of the previous reason, and to see if there was any correlation between the sentiment of the writing of the articles and a high index value a more extensive analysis was made for all the articles produced by Kitajima M. Results below show the 5 articles with the most positive sentiment and the 5 with the most negative.

```
Most positive articles:

Comparison of virus concentration methods for the RT-qPCR-based recovery of murine hepatitis virus, a surrogate for SARS-CoV-2 from untreated wastewater
---------------
compound score: 0.9715
---------------

Institutionalising wastewater surveillance systems to minimise the impact of COVID-19: Cases of Indonesia, Japan and Viet Nam
---------------
compound score: 0.9686
---------------

Early warning of COVID-19 via wastewater-based epidemiology: potential and bottlenecks
---------------
compound score: 0.9287
---------------

Duration of SARS-CoV-2 viral shedding in faeces as a parameter for wastewater-based epidemiology: Re-analysis of patient data using a shedding dynamics model
---------------
compound score: 0.8748
---------------

Surveillance of SARS-CoV-2 RNA in wastewater: Methods optimization and quality control are crucial for generating reliable public health information
---------------
compound score: 0.8658
---------------
```

Figure 9: *Most positive articles produced by Kitajima, M.*

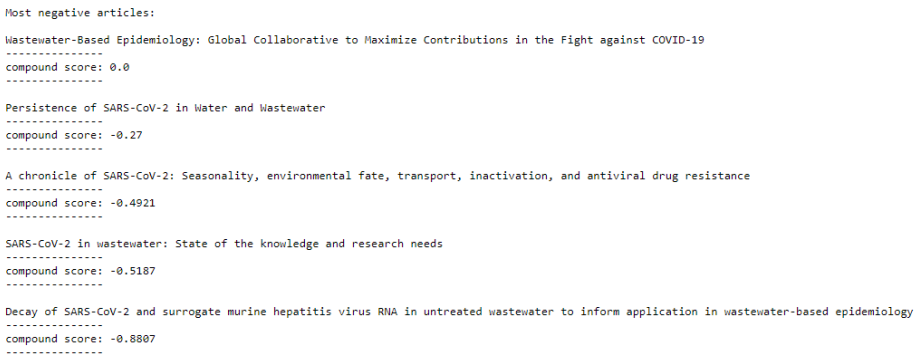

Figure 10: *Most negative articles produced by Kitajima, M.*

From the list above the 2 articles with the most positive score ("Comparison of virus concentration methods for the RT-qPCR-based recovery of murine hepatitis virus, a surrogate for SARS-CoV-2 from untreated wastewater") and most negative score ("Decay of SARS-CoV-2 and surrogate murine hepatitis virus RNA in untreated wastewater to inform application in wastewater-based epidemiology") were selected and the discussion part of the article was extracted and inserted in a new csv so the sentiment analysis can be performed again. The results showed that both articles still remained with their respected sentiment, the positive article got a score of 0.99 and the negative article got a score of -0.97 for the discussion part of both articles.

4 Conclusions

This project aimed to detect the impact the research of covid-19 wastewater-based epidemiology had by analyzing trends of all the research produced via a bibliometric approach and an added natural language processing to give more information about each article. It's possible to say the research has been substantially getting more numerous as time passes, this is considering that just for the search made for one single database the number of articles nearly tripled in just one year. However, the research being made about wastewater epidemiology related to covid detection in water is still relatively small in comparison to other subject areas so more effort needs to be put for this area of research. It should be remarked there are some authors that have made some significant efforts in helping this area of research like Ahmed and Kitagima as shown previously.

Using sentiment analysis It was possible to see that most of the articles made where positive in nature so one could assume the research done had good results that could help fight this disease in the future and even thought this could be true for some articles It needs to be said that the way articles are written in a objective , mostly neutral stance, makes this type of analysis not the most reliable for scientific work so one should look deeply into each article to see the results themselves and what they mean for future research.

5 References

Ahmed, W., Tscharke, B., Bertsch, P. M., Bibby, K., Bivins, A., Choi, P., Clarke, L., Dwyer, J., Edson, J., Nguyen, T. M. H., O'Brien, J. W., Simpson, S. L., Sherman, P., Thomas, K. V., Verhagen, R., Zaugg, J., & Mueller, J. F. (2021). SARS-CoV-2 RNA monitoring in wastewater as a potential early warning system for COVID-19 transmission in the community: A temporal case study. *Science of the Total Environment*, *761*. https://doi.org/10.1016/j.scitotenv.2020.144216

Bialek, S., Boundy, E., Bowen, V., Chow, N., Cohn, A., Dowling, N., Ellington, S., Gierke, R., Hall, A., MacNeil, J., Patel, P., Peacock, G., Pilishvili, T., Razzaghi, H., Reed, N., Ritchey, M., & Sauber-Schatz, E. (2020). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*(12), 343–346. https://doi.org/10.15585/mmwr.mm6912e2

Gerrity, D., Papp, K., Stoker, M., Sims, A., & Frehner, W. (2021). Early-pandemic wastewater surveillance of SARS-CoV-2 in Southern Nevada: Methodology, occurrence, and incidence/prevalence considerations. *Water Research X*, *10*, 100086. https://doi.org/10.1016/j.wroa.2020.100086

Hao, T., Chen, X., Li, G., & Yan, J. (2018). A bibliometric analysis of text mining in medical research. *Soft Computing*, *22*(23), 7875–7892. https://doi.org/10.1007/s00500-018-3511-4

Jones, A. W. (2015). Forensic Journals: Bibliometrics and Journal Impact Factors. In *Encyclopedia of Forensic and Legal Medicine: Second Edition* (Vol. 2). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-800034-2.00181-6

Ling, Y., Xu, S. B., Lin, Y. X., Tian, D., Zhu, Z. Q., Dai, F. H., Wu, F., Song, Z. G., Huang, W., Chen, J., Hu, B. J., Wang, S., Mao, E. Q., Zhu, L., Zhang, W. H., & Lu, H. Z. (2020). Persistence and clearance of viral RNA in 2019 novel coronavirus disease rehabilitation patients. *Chinese Medical Journal*, *133*(9), 1039–1043. https://doi.org/10.1097/CM9.0000000000000774

Nishiura, H., Kobayashi, T., Miyama, T., Suzuki, A., Jung, S. mok, Hayashi, K., Kinoshita, R., Yang, Y., Yuan, B., Akhmetzhanov, A. R., & Linton, N. M. (2020). Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *International Journal of Infectious Diseases*, *94*(February), 154–155. https://doi.org/10.1016/j.ijid.2020.03.020

Petersen, E., Koopmans, M., Go, U., Hamer, D. H., Petrosillo, N., Castelli, F., Storgaard, M., Al Khalili, S., & Simonsen, L. (2020). Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet Infectious Diseases*, *20*(9), e238–e244. https://doi.org/10.1016/S1473-3099(20)30484-9

Shen, M., Zhou, Y., Ye, J., Abdullah AL-maskri, A. A., Kang, Y., Zeng, S., & Cai, S. (2020). Recent advances and perspectives of nucleic acid detection for coronavirus. *Journal of Pharmaceutical Analysis*, *10*(2), 97–101. https://doi.org/10.1016/j.jpha.2020.02.010

Tahamtan, A., & Ardebili, A. (2020). Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Review of Molecular Diagnostics*, *20*(5), 453–454. https://doi.org/10.1080/14737159.2020.1757437

West, C. P., Montori, V. M., & Sampathkumar, P. (2020). COVID-19 Testing: The Threat of False-Negative Results. *Mayo Clinic Proceedings*, *95*(6), 1127–1129. https://doi.org/10.1016/j.mayocp.2020.04.004

WebReferences

1-WHO, 2020. Novel coronavirus (2019-nCoV) situation report - 1. WHO Bull. 1–7.
2- https://github.com/MiguelAndreDias/Projecto_Bioinform-tica/tree/main