Disaster Medicine and Public Health Preparedness

www.cambridge.org/dmp

Predicting the Utility of Scientific Articles for Emerging Pandemics Using Their Titles and Natural Language Processing

Kinga Dobolyi PhD, Sidra Hussain BS and Grady McPeak MS

Department of Computer Science, George Washington University, Washington, DC, USA

Original Research

Cite this article: Dobolyi K, Hussain S, McPeak G. Predicting the utility of scientific articles for emerging pandemics using their titles and natural language processing. *Disaster Med Public Health Prep*. **18**(e103), 1–6. doi: https://doi.org/10.1017/dmp.2024.109.

Keywords:
natural language processing; pandemic; policy; public health; scientific articles; utility

Abbreviations:
ACE-2, Angiotensin-converting Enzyme 2; BERT, Bidirectional Encoder Representations from Transformers; COVID-19, Coronavirus disease; DHS, Department of Homeland Security; MERS, Middle Eastern respiratory syndrome; MeSH, Medical Subject Headings; MQL, Master Question List; NIH, National Institutes of Health; NLP, Natural language processing; PPE, Personal protective equipment; SARS, Severe acute respiratory syndrome.

Corresponding author:
Kinga Dobolyi, PhD; Email: kinga@gwu.edu

Abstract

Objective: Not all scientific publications are equally useful to policy-makers tasked with mitigating the spread and impact of diseases, especially at the start of novel epidemics and pandemics. The urgent need for actionable, evidence-based information is paramount, but the nature of preprint and peer-reviewed articles published during these times is often at odds with such goals. For example, a lack of novel results and a focus on opinions rather than evidence were common in coronavirus disease (COVID-19) publications at the start of the pandemic in 2019. In this work, we seek to automatically judge the utility of these scientific articles, from a public health policy making persepctive, using only their titles.

Methods: Deep learning natural language processing (NLP) models were trained on scientific COVID-19 publication titles from the CORD-19 dataset and evaluated against expert-curated COVID-19 evidence to measure their real-world feasibility at screening these scientific publications in an automated manner.

Results: This work demonstrates that it is possible to judge the utility of COVID-19 scientific articles, from a public health policy-making perspective, based on their title alone, using deep natural language processing (NLP) models.

Conclusions: NLP models can be successfully trained on scienticic articles and used by public health experts to triage and filter the hundreds of new daily publications on novel diseases such as COVID-19 at the start of pandemics.

Not all peer-reviewed or preprint scientific publications are equally useful to policy-makers, and this is especially true in the case of emerging epidemics and pandemics, when there is an urgent need for information and research. For example, the outbreak of coronavirus disease (COVID-19) in late 2019 and its spread throughout early 2020 generated a deluge of publishing that grew rapidly in the first 6 months of the pandemic.[1] Opinion pieces,[2] often without novel research results, and other vanity articles[3] overwhelmed scientific publications in this domain and timeline. Even when such papers contain novel results, they may not translate directly into actionable policies for public health.

This work argues that it is possible to judge the *utility* of a scientific publication for public health policy-making, based solely on its title and/or abstract in many, if not most, cases. For example, a paper that is an obvious opinion based on a case study may be less useful than information collected from thousands of emergency room patients demonstrating the benefits of a drug. While focusing on titles and abstracts to triage publications based on their policy-making utility undoubtedly leaves room to miss important results, this work argues that introducing a way to rapidly select relevant articles is necessary, since policy-makers simply cannot read the hundreds of papers published daily during these times.[4] A tool that helps classify articles as likely speculative or opinion would provide additional means for policy-makers to make decisions.

In this work, a deep-learning model was trained to triage *useful* versus *not useful* papers to study during such emerging crises. To do so, expert human annotations were used to label several thousand COVID-19 articles as *useful* or not, in terms of public health policy-making. These labels were then used to train a deep natural language processing (NLP) model to predict scientific article utility based on paper titles and/or abstracts. Most closely related to this work are those papers that try to predict the impact of scientific articles using machine learning.[5] For example, researchers have discovered that shorter titles have higher citation counts,[6] and making a title amenable to search queries increases its impact.[7] However, scientific impact is measured by citation and social media networks along with altmetrics, rather than attempting to directly measure something akin to utility. This study is the first that we are aware of that attempts to model and predict this utility for scientific articles.

Methods

In this study, NLP was used to build predictive models that can label a COVID-19 scientific article as *useful* versus *not useful*, based on the title or abstract only (see Figure 1). These models

© The Author(s), 2024. Published by Cambridge University Press on behalf of Society for Disaster Medicine and Public Health, Inc.

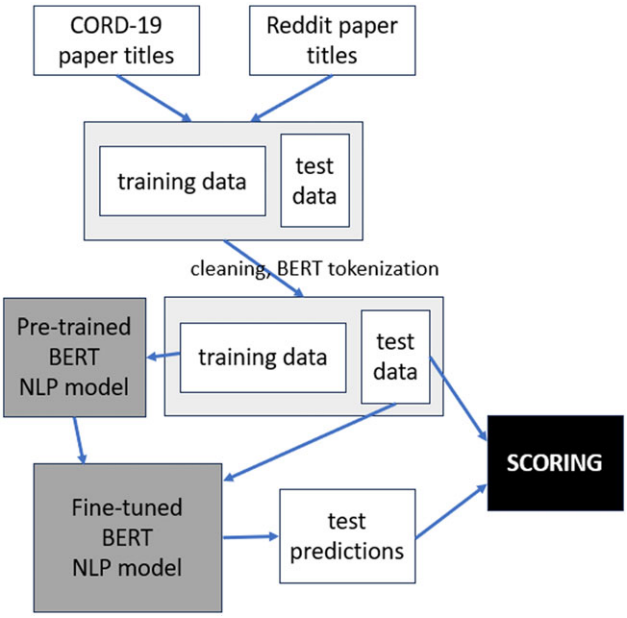


Figure 1. NLP model training in this study.

were trained from both social media engagement metrics (sourced from Reddit[8]) and human annotations on article usefulness. The ground truth of what characteristics make an article *useful* or *not useful* was determined by referring to the kinds of scientific articles cited in any version of the Department of Homeland Security's (DHS) Master Question List (MQL) for COVID-19,[9] an expert-curated document that was updated weekly throughout the pandemic with scientific (and commercial) sources of information curated to answer basic questions about the novel disease.

CORD-19 Training Data for the Model

Scientific articles and preprints were obtained from the CORD-19 dataset[10] as the collection of training data samples for the model. Focusing on the first 6 months of the declared pandemic (February 2020 through July 2020), this study randomly sampled about 1000 articles per month from this dataset for the training data, which were later annotated as *useful* or *not useful* (described below).

Ground Truth Utility Dataset From DHS

The ground truths for utility, used for testing across all the models, were sourced from citations in the DHS MQL for COVID-19.[9] The MQL was updated weekly throughout the pandemic, collecting literature to answer basic questions about the disease across topics of infectious dose, transmissibility, incubation period, clinical presentation, treatments, personal protective equipment (PPE), and others. Dozens of citations might answer these questions, and these human-curated answers were updated regularly throughout the pandemic.

This study used the DHS MQL obtained on December 21, 2020, providing about 300 cited papers. Only DHS citations that also appeared in CORD-19 were used because the DHS often cites non-scientific sources in their MQL (such as news articles). The DHS timeline was also extended to December 2020 instead of ending it in July to allow for a larger test set to evaluate the approach, a necessary move considering a dataset consisting of 300 or fewer entries is considered very small in the world of machine learning.

Reddit Training Data for the Model

To generate an alternative, larger testing dataset for the experiments besides just DHS, the COVID-19 subreddit (*reddit.com/r/covid19*[8]), which is a social media forum focusing exclusively on scientific articles, preprints, and data only, was scraped. From February 2020 through July 2020, 1913 posts were collected and then matched against the CORD-19 via their paper titles. There were 44, 107, 396, 495, 380, and 380 such papers across each month of February 2020–July 2020, respectively.

This same Reddit dataset was also added into training data for certain versions of the model, as long as it never tested and trained on the same data. This was done in order to achieve a more balanced training dataset in terms of *useful* versus *not useful* articles, given that the vast majority of CORD-19 papers from this date range were expected to not be useful. When categorizing these data, we identified forum posts with a high number of "upvotes" as being viewed by the forum community as *useful*, given the stated purpose of the forum is to "facilitate scientific discussion of this potential global public health threat."

Human Annotations of Training Data

Reddit is a public forum without membership moderation; it is not possible to know the details of its members who choose to post. Therefore, an alternative training dataset of 5298 papers was annotated by a single bioinformatician who had followed the COVID-19 literature and was familiar with the DHS MQL since the beginning of the pandemic. The annotator was given access to both the CORD-19 training data and the Reddit dataset and asked to label them on a 0-2 utility scale (defined below). They were not informed which papers came from CORD-19 versus Reddit. Together, there were ~6700 unique labeled papers in this joined dataset that could be used for training the model. Approximately 15% of their annotations were labeled as *useful* (a non-zero score) when told their goals were to flag papers that could help organizations that are tasked with coordinating a response to the pandemic to minimize the amount of transmission, hospitalizations and deaths, and post-viral disability due to the virus (our definition of utility).

Training the NLP Model to Predict Utility

The data above were used to fine-tune a deep learning NLP model to label unseen papers as either *useful* or *not useful* from a public health policy perspective. The training took advantage of a pretrained version of BERT,[11] using the *bert-base-uncased* weights,[11] and was fine-tuned over 2 epochs with a learning rate of 2.00E-05 and the maximum length of the tokens set to 30 (for titles, as these tended to be short in general) with a batch size of 16. A *WeightedRandomSampler* was used during the data loading of the BERT model training to help balance the training dataset, which was overwhelmingly made up of *not useful* articles (85%). The chronological appearance of any paper was not a feature fed into the model.

Experimental Setup

This study evaluated the feasibility of using a deep learning model to predict scientific article *utility* across 3 separate ground truths: (1) papers labeled by the human annotator; (2) papers that appeared on the scientific COVID-19 subreddit; and (3) papers judged as useful by experts at the DHS in their COVID-19 Master Question List (MQL) answers. To investigate whether the approach could generalize beyond just COVID-19, an additional

https://doi.org/10.1017/dmp.2024.109 Published online by Cambridge University Press

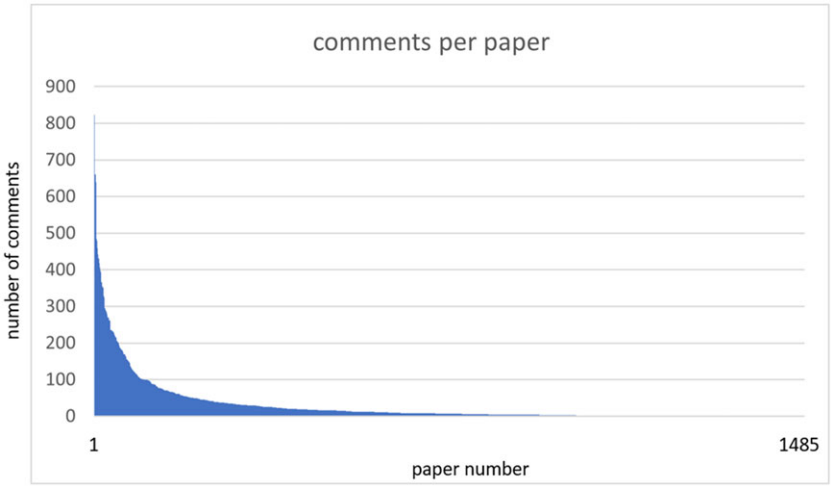


Figure 2. The number of comments on the COVID-19 subreddit that appeared for the matching papers in the CORD-19 dataset during February–July 2020.

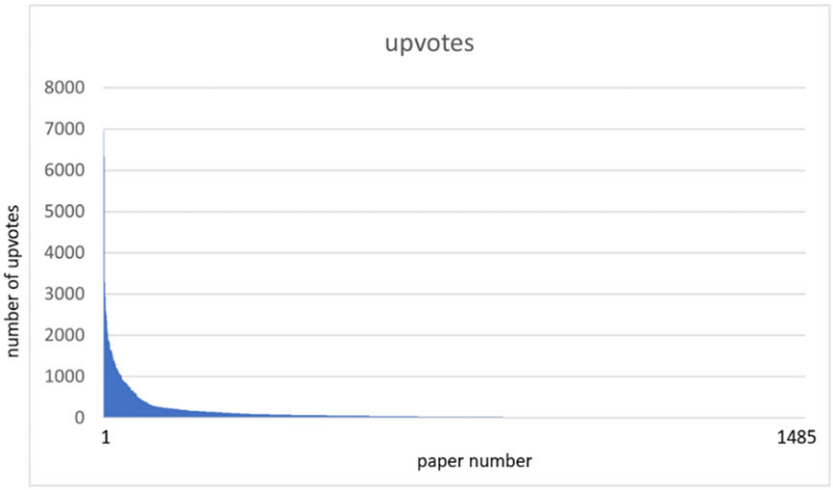


Figure 3. The number of upvotes on the COVID-19 subreddit that appeared for the matching papers in the CORD-19 dataset during February–July 2020.

cleaning of all the training datasets above was performed to remove keywords that pertained to any specific disease or medical terminology. MeSH[12] keywords from the NIH Database replaced any word in the title/abstract that matched any MeSH keyword with the string *WORD*. For example, an original paper title, *Incidental CT Findings Suspicious for COVID-19-Associated Pneumonia on Nuclear Medicine Examinations: Recognition and Management Plan,* was scrubbed to be: *incidental ct findings suspicious for WORD-associated WORD on nuclear WORD exWORD: reWORD and management plan*. Models trained on such a relatively topic-agnostic set of titles/abstracts may be better able to generalize to other diseases and emerging pandemics or may even generalize outside the biomedical domain. All papers with duplicate titles from the training datasets were removed. Further, for each run of the experiments below, no paper title in the test dataset was in the training dataset.

For experiments testing the model on papers that appeared on Reddit, this study also examined only such papers that had at least 100 comments in their discussion section, or at least 500 upvotes overall, as a more stringent definition of utility where these papers were a magnet for comments and votes. Figures 2 and 3 show these respective distributions, illustrating the choice of these cutoffs.

Results

Validating That a Predictive Model Can Be Built

The first set of experiments sought to establish the feasibility of using the title and/or abstract of COVID-19 scientific articles to predict their *utility* from a public health policy-making perspective. This study therefore trained and tested 3 different iterations of the BERT pretrained model on such tasks, using the CORD-19 and Reddit papers that were labeled as *useful* or *not useful* by the human annotator; such a model had about an 80% weighted accuracy, precision, and recall, in correctly labeling these papers. Table 1 presents the results. These 3 experiments were performed with 10-fold cross validation.

When the same approach was applied to paper abstracts only, the model weighted accuracy fell to 71%, with recall and precision dipping similarly. It is possible that limitations in terms of feeding only 128 tokens (thus, truncating the abstracts) into BERT may have contributed to these observations. However, it is also likely that, although BERT arguably makes some attempt to "understand" the natural language text it is presented with, asking it to comprehend a full paragraph of abstract text is infeasible.

https://doi.org/10.1017/dmp.2024.109 Published online by Cambridge University Press

Table 1. Experimental results using 10-fold cross validation for the CORD-19+Reddit datasets (95% confidence intervals are reported)

Number	Training dataset (N)	Test dataset (N)	Purpose	Weighted accuracy	Weighted recall	Weighted precision
Experiment 1 [titles]	Random CORD-19 + Reddit (6000) with 15% considered *useful* papers	10% of training dataset not used in training (668)	Validate that a predictive model can be built	0.804 ± 0.03	0.806 ± 0.03	0.829 ± 0.024
Experiment 6 [titles]	Same as Ex. 1 above, with MeSH scrubbing	Same as Ex. 1 above, with MeSH scrubbing	See how the model above might generalize outside the COVID-19 domain	0.7709 ± 0.0314	0.7724 ± 0.029	0.8004 ± 0.0265
Experiment 12 [abstracts]	Random CORD-19 + Reddit (4190) with 18% considered *useful* papers	10% of training dataset not used in training (467)	Validate that a predictive model can be built	0.7129 ± 0.0341	0.7129 ± 0.0341	0.7506 ± 0.0398

Table 2. Experimental results for the models on the DHS ground truth dataset (95% confidence intervals are reported)

Number	Training dataset (N)	Test dataset (N)	Purpose	Recall
Experiment 2	Random CORD-19 + Reddit (6629) with 15% considered *useful* papers	DHS (290); all considered *useful* papers	Measure potential real-world utility of the model compared to experts	0.653 ± 0.025
Experiment 7	Same as Ex. 2 above, with MeSH scrubbing	Same as Ex. 2 above, with MeSH scrubbing	See how the model above might generalize outside the COVID-19 domain	0.6112 ± 0.01552
Experiment 11	Random CORD-19 + Reddit (6572) with 15% considered *useful* papers	DHS (254) with case studies, small population, and modeling papers removed manually; all considered *useful* papers	Examine the potential bias of the human annotator vs DHS experts	0.6671 ± 0.0211

This study also investigated how the approach might generalize outside of COVID-19, in particular. All biomedical keywords that matched anything in the MeSH database of terms were removed and replaced with the string *WORD*. In doing so, this study wanted not only to remove COVID-19 synonyms (such as SARS CoV 2), but also to not have the model learn anything important about specific biologic pathways (such as ACE-2) or treatments (such as hydroxychloroquine). The weighted metrics dropped about 3% when training and testing on such a scrubbed dataset. One could imagine a real-world scenario for a novel pandemic that takes advantage of the models trained on unscrubbed data and then fine-tunes them as time goes by with online training methods for papers in the new disease domain.

Measuring Model Efficacy Against the DHS Dataset

This study also tested the model from Experiment 1 above on papers cited by DHS in their COVID-19 MQL; these are de facto expert judgments on what articles were *useful* between the start of the pandemic and December 2020. The results are shown in Table 2.

In this setup, all the DHS papers were, by this work's definition, *useful*, so only recall was measured. The performance of the model dropped by ~15% on this test set, indicating that it is missing *useful* papers. However, on average, roughly half of the DHS papers were not considered *useful* by the annotator. Although it is impossible to eliminate any chance of unintentional bias from the human annotation, not all the questions of the MQL are attempting to answer things that are immediately *useful* for public health policy. For example, questions around host range, which made up 5% of the raw DHS dataset of articles before matching these to CORD-19, are very unlikely to be relevant when they reference macaques, pangolins, and bats.

Similarly, the DHS citations that the annotator judged as *not useful* tended to be case studies, have small study populations, or discuss various modeling strategies. When these kinds of papers were removed from the DHS test set, the recall was 67%, illustrating the differences in what the human judged as important from everything cited by DHS. In other instances, DHS papers had titles such as *Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation*, which are unlikely to directly translate into actionable public health decision-making. As in the previous section, allowing COVID-19 specific keywords in the training and testing datasets improved the model performance by ~4% in terms of recall.

Measuring Model Efficacy Against the Reddit Dataset

As a complement to the DHS dataset, this study also sought to measure the model's performance on predicting what papers would be judged useful by people on the scientific *Covid19* subreddit.[8] While the forum is moderated, anyone can post a paper title and link. Therefore, the model was expected to perform more poorly on this test set, as it, by contrast, was trained by annotations from someone highly familiar with the COVID-19 literature at the time. In addition to that known limitation, this study had to remove any paper that appeared in the test set from the training data. Therefore, it is more likely than chance that the papers removed from the training set, in order to not train and test on the same data, happened to be *useful*; this is a disadvantage in terms of providing these flavors of the model robust training data.

The model performed worse at predicting what articles would show up on Reddit, compared to DHS citations or judged as *useful* by the annotator, as shown in Table 3. Recall dropped significantly from 0.653 to 0.3711; however, this experiment assumed all Reddit

https://doi.org/10.1017/dmp.2024.109 Published online by Cambridge University Press

Table 3. Experimental results for models trained on CORD-19 only and tested on the Reddit datasets (95% confidence intervals are reported)

Number	Training dataset (N)	Test dataset (N)	Purpose	Recall
Experiment 3	Random CORD-19 (5346) with 13% considered *useful* papers	All Reddit papers (1527)	Measure potential real-world utility for non-experts	0.3711 ± 0.022
Experiment 8	Same as Ex. 3 above, with MeSH scrubbing	Same as Ex. 3 above, with MeSH scrubbing	See how the model above might generalize outside the COVID-19 domain	0.4123 ± 0.0146
Experiment 4	Random CORD-19 (6563) with 14% considered *useful* papers	Reddit papers that had at least 100 user comments (114)	Measure potential real-world utility for the most popular papers on Reddit based on engagement	0.4078 ± 0.0422
Experiment 9	Same as Ex. 4 above, with MeSH scrubbing	Same as Ex. 4 above, with MeSH scrubbing	See how the model above might generalize outside the COVID-19 domain	0.4898 ± 0.0345
Experiment 5	Random CORD-19 (6606) with 14% considered *useful* papers	Reddit papers that had at least 500 user upvotes (76)	Measure potential real-world utility for the most popular papers on Reddit based on popularity	0.4529 ± 0.0337
Experiment 10	Same as Ex. 5 above, with MeSH scrubbing	Same as Ex. 5 above, with MeSH scrubbing	See how the model above might generalize outside the COVID-19 domain	0.5496 ± 0.0483

papers to be *useful*. This observable decrease in performance on a significantly lower quality test dataset demonstrates that the model was not simply overtrained on our human annotations, but that its performance decays the further the test set is removed from a stringent set of requirements about utility.

In reality, anyone can anonymously post to Reddit, so the quality of posts and/or upvotes is unknown. To investigate this further, the Reddit test sets were filtered into 2 groups: papers that had over 100 comments and (possibly the same) papers that had over 500 upvotes. As expected, the model's performance increased as the testing dataset was limited to papers that are more likely to be considered broadly useful.

Comparison to Baseline Shallow Learning Models

Although the BERT models performed well enough on the expert-judged dataset to be used in a real-world scenario to triage COVID-19 papers from a public health perspective, such deep learning models often lack an explanation of why the model made its predictions. Research into what makes a scientific paper title popular includes its sentiment,[13,14] length, use of colons, acronyms, question marks, humor or cliches, and if results versus methods are conveyed.[15] In other work, question-titles were less popular, while those that lead with results were more popular[15]; colons were preferred, as were longer titles.[15] The average title sentiment was lower for *useful* papers than *not useful* papers in our dataset.

While this work's metric of *utility* is orthogonal to popularity and preference, this study explored such features in a baseline shallow learning model. A simple Random Forest Classifier that used the features of the title length, number of verbs, number of nouns, the sentiment of the title, and whether it contained a colon/semicolon or a question to investigate how these basic features might influence predicting utility. Such a baseline shallow learning model was only 54% accurate (weighted)—barely better than chance. Unlike the BERT-based models, the engineered features in this baseline were not enough to make decisions about what is an important paper.

Discussion

During the emergence of a novel, high-impact pandemic such as COVID-19, being able to quickly and effectively generate evidence-based public health policies in the face of incomplete and contradictory research is important. Evidence is important, especially in the first few weeks and months of such an outbreak, and many, if not most, of the initial preprints and publications from the first 6 months of the pandemic[2] lacked original experimental results.

Given that only 20% of COVID-19 papers at the start of the pandemic later appeared in peer-reviewed journals,[16] it is not surprising that in this work it was found that, on average, around the same percentage of random papers sampled from the CORD-19 database were to be *useful* from a public health policy-making perspective. While these 2 sets do not fully intersect, it is clear that policy-makers tasked with setting public health recommendations at the start of the pandemic were faced with a deluge of unhelpful articles. This work demonstrated that, given the conditions of the experiment, it is possible to train a machine learning model to predict the utility of such scientific articles with up to 80% accuracy, based on their titles alone. Such a model could be used by policy-makers and scientists to triage the deluge of low-utility publications, especially at the start of pandemics. It is our hope, however, that this model is not in turn used by academics to attempt to optimize their papers for better "usefulness," search query results, and in turn more citations. At the same time, it must be acknowledged that it may be possible to manipulate article titles to effectively poison the data and undermine a reliable or useful outcome.

Although the deep learning NLP models were able to generate predictions that could be used in the real-world, it is difficult to understand why they made their decisions. Therefore, this work compared the successful deep learning models against shallow learning models, and performed ablation experiments to gain insight as to why the models were labeling papers as *useful* or not.

This study also explored how a reduction in the quality of what papers were considered *useful*—especially by relying on social media popularity for this metric—indicates that there is less of an intersection in terms of what people find popular on Reddit versus what policy-makers are flagging for closer inspection. Although the model was better at flagging highly popular papers on Reddit than just any paper that was posted, it performed better when biomedical keywords were allowed to remain in the title; this is the opposite of what was observed with the 2 datasets curated by experts (either the human annotator or DHS scientists). In both cases, the difference was small, indicating that it is possible to build a model that is paying attention to the structure of the paper titles more than any specific diseases or medical keywords. One would therefore expect the models to generalize to previous infectious disease outbreaks (such as Ebola or monkeypox) or future pandemics. Such an exploration will be examined in future work.

https://doi.org/10.1017/dmp.2024.109 Published online by Cambridge University Press

Limitations

While the results of this study suggest it is possible to build a machine learning model that successfully triages early stage COVID-19 scientific articles as *useful* or not from a public health policy perspective, there are several limitations to the work. First, as a human expert was responsible for judging CORD-19 papers as *useful* or not, it is possible their opinions may not generalize entirely. Furthermore, a single scientific article may change in utility over time, depending on where in an outbreak a population is in terms of time. Small case studies may be more valuable at the start of pandemics as opposed to later on; it is difficult to enforce this type of consideration when making utility judgments in hindsight. Similar limitations may exist in the DHS ground truth dataset in terms of deliberate or natural curation of evidence as the pandemic went on. Other groups outside DHS may also have different utility definitions. Finally, paper themes such as long-term sequelae may require the model be updated with new training data in the future, as these themes were less prominent in our training datasets from earlier in the pandemic.

Conclusions

This study demonstrates that it is feasible to predict the utility of COVID-19 scientific articles as information for public health policy-makers based on their title alone. Using a deep learning natural language processing model, a system was trained that could triage papers for further reading for the corpus of articles and preprints that were published during the first 6 months of the pandemic. Because model performance was minimally affected by removing all biomedical keywords from the paper titles, the approach could be theoretically used on other diseases as well as future pandemics.

Acknowledgments. The authors would like to thank George Sieniawski for preliminary discussions on the topic of scientific article utility and his suggestion of using the DHS MQL as the main comparator of this study. We would also like to thank Son Hoai Nguyen for his contributions to the code used to collect Reddit data, as well as Luca Caruso and Krish Sadhwani for their sentiment analysis code.

Author contributions. Conceptualization: Kinga Dobolyi; Methodology: Kinga Dobolyi, Sidra Hussain, Grady McPeak; Data curation: Sidra Hussain, Kinga Dobolyi; Formal analysis and investigation: Kinga Dobolyi; Writing – original draft preparation: Kinga Dobolyi; Writing – review and editing: Kinga Dobolyi, Sidra Hussain, Grady McPeak.

Funding statement. The authors have no relevant financial or non-financial interests to disclose. The authors did not receive support from any organization for the submitted work.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethical standards. For the human subject involved in annotating the scientific articles, this work was deemed exempt from IRB review under DHHS regulatory Category 2 by George Washington University.

References

1. **Odone A, Galea S, Stuckler D, Signorelli C.** The first 10000 COVID-19 papers in perspective: are we publishing what we should be publishing? *Eur J Public Health.* 2020;30(5):849-850. https://doi.org/10.1093/eurpub/ckaa170
2. **Raynaud M, Zhang H, Louis K, *et al.*** COVID-19-related medical research: a meta-research & critical appraisal. *BMC Med Res Methodol.* 2021;21(1):2313-2349. https://doi.org/10.1186/s12874-020-01190-w
3. **Jalali R, Hosseinian-Far A, Mohammadi M.** Contradictions in the promotion of publishing academic & scientific journal articles, & the inability to cope with the new coronavirus (COVID-19). *Antimicrob Resist Infect Control.* 2021;10(1). Published online 12 January 2021. https://doi.org/10.1186/s13756-021-00884-0
4. **Mohammed M, Sha'aban A, Jatau AI, *et al.*** Assessment of COVID-19 information overload among the general public. *J Racial Ethn Health Disparities.* 2022;9(1):184-192. https://doi.org/10.1007/s40615-020-00942-0
5. **Bai X, Liu H, Zhang F, *et al.*** An overview on evaluating and predicting scholarly article impact. *Information.* 2017;8(17). Published online 25 June 2017. https://doi.org/10.3390/info8030073
6. **Rossi MJ, Brand JC.** Journal article titles impact their citation rates. *Arthroscopy.* 2020;36(7):2025-2029. https://doi.org/10.1016/j.arthro.2020.02.018
7. **Beranová L, Joachimiak MP, Kliegr T, *et al.*** Why was this cited? Explainable machine learning applied to COVID-19 research literature. *Scientometrics.* 2022;127:2313-2349. https://doi.org/10.1007/s11192-022-04314-9
8. COVID-19. COVID19 subreddit. Published 2020. Accessed February 1, 2020–July 31, 2020. https://www.reddit.com/r/COVID19/
9. **Master Question List for COVID-19**. US Department of Homeland Security. Published 2020. Accessed December 21, 2020. https://www.dhs.gov/publication/st-master-question-list-COVID-19
10. **Wang LL, Lo K, Chandrasekhar Y, *et al.*** CORD-19: The COVID-19 open research dataset. Preprint. *ArXiv*. Published online April 22, 2020.
11. **Devlin J, Chang M, Lee K, Toutanova K.** BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2019;1(Long & Short Papers):4171-4186. https://doi.org/10.18653/v1/N19-1423
12. **Download MeSH Data**. National Library of Medicine. Published 2022. Accessed December 1, 2022. https://www.nlm.nih.gov/databases/download/mesh.html
13. **Fabiano N, Hallgrimson Z, Wong S, *et al.*** Selective tweeting of COVID-19 articles: does title or abstract positivity influence dissemination? Preprint. *medRxiv.* 2021. Published online 24 June 2021. https://doi.org/10.1101/2021.06.22.21259354
14. **Lockwood G.** Academic clickbait: articles with positively-framed titles, interesting phrasing, and no wordplay get more attention online. *The Winnower.* 2016;3. Published online 29 June 2016.
15. **Hallock RM, Bennett TN.** I'll read that!: what title elements attract readers to an article? *Teach Psychol.* 2021;48(1):26-31. https://doi.org/10.1177/0098628320959948
16. **Älgå A, Eriksson O, Nordberg M.** The development of preprints during the COVID-19 pandemic. *J Intern Med.* 2021;290(2):480-483. https://doi.org/10.1111/joim.13240