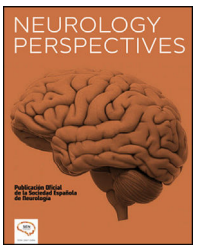
Neurology Perspectives 3 (2023) 100115

NEUROLOGY PERSPECTIVES

www.journals.elsevier.com/neurology-perspectives



SCIENTIFIC LETTER

Bias, coronavirus, nationality, gender and neurology article citation count prediction with machine learning

Sesgo, coronavirus, nacionalidad, género y neurología Predicción de recuento de citas de artículos con aprendizaje automático

S. Bacchi[a,b,c,*], S.C. Teoh[c], L. Lam[a,c], D. Schultz[b], Robert J. Casson[a,c], W. Chan[a,c]

[a] *University of Adelaide, South Australia, Australia*
[b] *Flinders Medical Centre, South Australia, Australia*
[c] *Royal Adelaide Hospital, South Australia, Australia*

Dear Editors,

The timely identification of impactful research, as may be indicated by citation count, may facilitate scientific advancement. It is possible that machine learning, including natural language processing, may be able to assist with this task. However, machine learning applications also have the potential to perpetuate biases, and this requires close examination.

One way in which machine learning may be applied to facilitate the research process is through the automatic analysis of abstracts. For example, previous analyses have suggested that natural language processing with sentiment analysis can be successfully applied to detect aspects of the impact of stroke trials.[1] This type of analysis has promise, but also requires interrogation prior to widespread use. There are multiple potential sources of bias in natural language processing analyses.[2] In particular, biases may occur due to the data upon which analyses are based, the labelling of these data, the analysis of these data, or the application of the tools in practice.

The aim of this study was to examine the performance of machine learning, namely natural language processing, in the prediction of citation count for neurology articles, relative to other articles published in the same year.

The study included all publications identified in the PubMed database published from inception to 2021 identified with the MeSH term 'neurology'. The title, abstracts, author lists, MeSH terms, journal international standardised serial number (ISSN), and citation count (March 2022) were extracted and subsequently used as input data fields. Articles were allocated percentile ranks, compared with other articles published in the same year, for the total number of citations received. Following pre-processing (including capitalisation removal and word stemming), models were developed on the training dataset to predict which articles would rank in the top quartile for citation count, as compared with other articles published in the same year. Logistic regression (LR) models were developed for each of the input data fields individually, and then combined. Regression coefficients were ranked to identify the 50-word stems most strongly associated with a top quartile citation count for each text field. Similarly, the 50-word stems most strongly associated with not having a top quartile citation count were examined. However, it was pre-specified that author name and journal ISSN analysis associated with a lower citation count would not be presented. A bidirectional encoder representations from transformers (BERT) model was developed for the best performing combination of input data. Performance was evaluated on the hold-out test dataset. The primary outcome was the area under the receiver operator curve (AUC) for the BERT model. Analysis was conducted using open-source Python libraries including Sci-Kit Learn and Tensorflow.[3,4]

There were 468,550 articles included in the study. Several patterns were apparent in the analysis of the

https://doi.org/10.1016/j.neurop.2023.100115

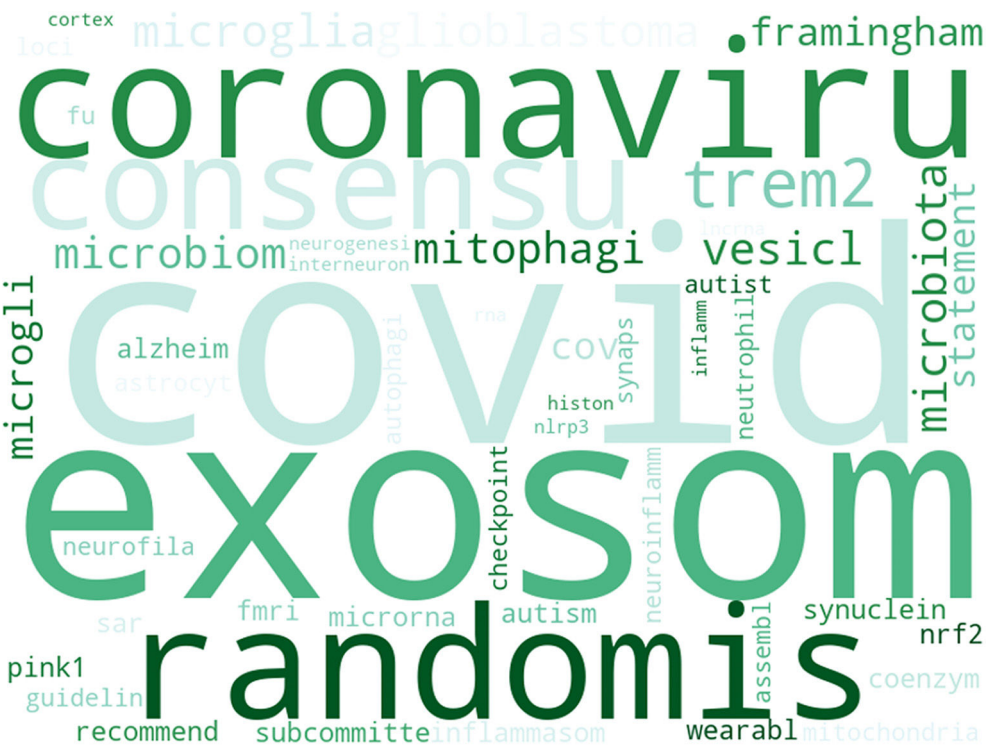


Fig. 1 Word stems most strongly associated with having a top quartile citation count based on the analysis of titles. In this visualisation the size of the word represents the magnitude of the regression coefficient.

regression coefficients associated with top quartile citation count (see Supplementary Information 1). In particular, coronavirus terms, terms related to multicentre randomised trials, and the nation 'america' were frequently associated with top quartile citation counts. For example, Fig. 1 illustrates a word cloud of the title word stems most strongly associated with a top quartile citation count. The prominence of the coronavirus related terms in Fig. 1 demonstrates that, while the terms may have been infrequent in articles near the beginning of the pandemic, those that were present were highly likely to receive top quartile citation counts. However, there was also a trend for certain country names and female pronouns to be associated with a lower likelihood of a high citation count. For example, when analysing titles, notable terms that were associated with a lower likelihood of high citation count were 'woman', 'polish', 'brazillian', 'poland', 'korean', 'iranian', and 'japanes'. In titles, the 10 words most strongly associated with not having a top quartile citation count were 'repli', 'author', 'reader', 'teach', 'comment', 'letter', 'commentari', 'editori', 'protocol', and 'case'. Notably many of these words relate to article type (e.g., 'editori' and 'commentari'), rather than article content.

The best performing logistic regression analysis used all of the available inputs, including title, abstracts, author lists, MeSH terms, and ISSN. This model returned an AUC of 0.81. When the BERT algorithm was applied with all input data it achieved an AUC of 0.86 for this task.

This study has shown that article citation counts can be successfully predicted with data available at the time of publication and natural language processing. However, such analyses have a signal suggesting there may be risks with respect to perpetuating geographic and gender biases. There are multiple means by which these algorithms can become biased, including data annotation, model development, and design of the applications of the models.[2] This potential for bias also is present in other medical applications of machine learning. For example, the relative underrepresentation of females in clinical trials has been shown to have the potential to bias machine learning models developed on such datasets.[5] Strategies to help mitigate bias include the collection of representative datasets, use of oversampling in model development with unbalanced datasets, and subpopulation analyses. Future research may seek to develop natural language processing algorithms to assist with other parts of the scientific writing process, including grant applications. However, such machine learning systems should exercise caution with respect to potential biases.

Funding

Nil.

Patient consent (informed consent)

Not applicable.

Ethical considerations

No ethics approvals required.

Conflict of interest

The authors have no conflicts of interest to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neurop.2023.100115.

References

1. Fischer I, Steiger HJ. Toward automatic evaluation of medical abstracts: The current value of sentiment analysis and machine learning for classification of the importance of PubMed abstracts of randomized trials for stroke. J Stroke Cerebrovasc Dis. 2020;29, 105042.
2. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Language and Linguistics. Compass. 2021:15.
3. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
4. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning.Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16); 2016.
5. Agmon S, Gillis P, Horvitz E, Radinsky K. Gender-sensitive word embeddings for healthcare. J Am Med Inform Assoc. 2022;29: 415–23.