Received: 11 March 2020 | Revised: 17 September 2020 | Accepted: 7 October 2020

DOI: 10.1111/geb.13219

MACROECOLOGICAL METHODS

Global Ecology and Biogeography A Journal of Macroecology WILEY

Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets

Richard Cornford[1,2,3] | Stefanie Deinet[1] | Adriana De Palma[2] | Samantha L. L. Hill[4] | Louise McRae[1] | Benjamin Pettit[5] | Valentina Marconi[1,3] | Andy Purvis[2,3] | Robin Freeman[1]

[1]Institute of Zoology, Zoological Society of London, London, United Kingdom

[2]Department of Life Sciences, Natural History Museum, London, United Kingdom

[3]Department of Life Sciences, Imperial College London, Ascot, United Kingdom

[4]UNEP World Conservation Monitoring Centre, Cambridge, United Kingdom

[5]Cleo AI Ltd., London, United Kingdom

Correspondence
Richard Cornford, Institute of Zoology, Zoological Society of London, London, NW1 4RY, UK.
Email: richard.cornford16@imperial.ac.uk

Funding information
Natural Environment Research Council, Grant/Award Number: NE/M014533/1 and NE/R012229/1

Editor: Pedro Peres-Neto

Abstract

Aim: Understanding broad-scale ecological patterns and processes is necessary if we are to mitigate the consequences of anthropogenically driven biodiversity degradation. However, such analyses require large datasets and current data collation methods can be slow, involving extensive human input. Given rapid and ever-increasing rates of scientific publication, manually identifying data sources among hundreds of thousands of articles is a significant challenge, which can create a bottleneck in the generation of ecological databases.

Innovation: Here, we demonstrate the use of general, text-classification approaches to identify relevant biodiversity articles. We apply this to two freely available example databases, the Living Planet Database and the database of the PREDICTS (Projecting Responses of Ecological Diversity in Changing Terrestrial Systems) project, both of which underpin important biodiversity indicators. We assess machine-learning classifiers based on logistic regression (LR) and convolutional neural networks, and identify aspects of the text-processing workflow that influence classification performance.

Main conclusions: Our best classifiers can distinguish relevant from non-relevant articles with over 90% accuracy. Using readily available abstracts and titles or abstracts alone produces significantly better results than using titles alone. LR and neural network models performed similarly. Crucially, we show that deploying such models on real-world search results can significantly increase the rate at which potentially relevant papers are recovered compared to a current manual protocol. Furthermore, our results indicate that, given a modest initial sample of 100 relevant papers, high-performing classifiers could be generated quickly through iteratively updating the training texts based on targeted literature searches. These findings clearly demonstrate the usefulness of text-mining methods for constructing and enhancing ecological datasets, and wider application of these techniques has the potential to benefit large-scale analyses more broadly. We provide source code and examples that can be used to create new classifiers for other datasets.

KEYWORDS
automated classification, biodiversity indicators, Biodiversity Intactness Index, ecological data, Living Planet Index, machine learning, text mining

1 | INTRODUCTION

Substantial anthropogenic change is degrading the natural world, creating an urgent need to understand the drivers and consequences of biodiversity loss to inform mitigation strategies (IPBES, 2019; WWF, 2020). For example, monitoring progress towards international conservation policy objectives, such as the Aichi Biodiversity Targets (CBD, 2010), requires the reliable, accurate and rapid tracking of changes in the state of nature (Collen & Nicholson, 2014; Walpole et al., 2009). Currently, collating data for such analyses is a time-consuming and largely manual process (Collen et al., 2009; Hudson et al., 2014), typically involving literature searches, manual screening of titles and abstracts for relevance, assessment of data quality, liaising with study authors to obtain data when necessary and entering usable data into the database. Estimates from other fields, such as medical systematic reviews, suggest that an experienced reviewer may take between 30 s and several minutes to assess an abstract (O'Mara-Eves et al., 2015). The annual rate of publication of scientific papers is growing at 8–9% per year (Landhuis, 2016) and over 15,500 ecology-related papers were indexed in Web of Science in 2019. Manually creating and updating ecological databases will therefore become ever more laborious (Ananiadou et al., 2009; Cohen et al., 2012). If current data-collection techniques cannot keep pace, large portions of relevant, available data might not be incorporated, potentially leading to suboptimal and potentially biased outputs that could not only hinder scientific progress (Nunez-Mir et al., 2016) but may misinform policy makers.

Combining text-mining and machine-learning approaches has the potential to substantially increase the rate of data discovery and database growth. These techniques have so far had relatively limited use in the biological sciences (Nunez-Mir et al., 2016), but evaluations in the context of producing medical systematic reviews show that they can classify accurately and save time (O'Mara-Eves et al., 2015), and can even correct human error (Bannach-Brown et al., 2019). Within ecology, Roll et al. (2018) recently used automated content analysis and artificial neural networks to accurately determine whether texts associated with the term 'reintroduction' were linked to conservation biology or another topic, and recommended further use of text mining and machine learning in conservation to better inform policy and management practices (Roll et al., 2018).

In this paper, we demonstrate how text classifiers trained through supervised machine-learning can identify papers containing ecological data, applying the approach to two high-profile biodiversity indicator databases as examples. The Living Planet Database (LPD: http://livingplanetindex.org/data_portal) contains population time-series data on over 4,000 vertebrate species collected from over 25,000 populations and is used to produce the Living Planet Index (LPI: Collen et al., 2009; Loh et al., 2005; McRae et al., 2017), one of the most widely used indicators of biodiversity (Mace & Baillie, 2007). The database of the PREDICTS (Projecting Responses of Ecological Diversity in Changing Terrestrial Systems) project (Hudson et al., 2017) collates ecological assemblage data from terrestrial sites worldwide that face different pressures relating to land-use change. More than 50,000 taxa and over 32,000 sites are included and from this the global status of a range of indicators, including the Biodiversity Intactness Index (BII: Scholes & Biggs, 2005), can be calculated (Newbold et al., 2016; Purvis et al., 2018). Both indicators have been used widely in high-profile reports, such as the IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services) Global Assessment (IPBES, 2019) and the Living Planet Report (WWF, 2020). By tuning our workflow on LPD data and then applying this to PREDICTS, we illustrate the generality of our approach, which can be applied to any (ecological) database created using data extracted from literature sources.

Given the variety of text classifiers available (Khan et al., 2010; Kotsiantis et al., 2007), we compare the performance of easy-to-create logistic regression (LR) to 'black-box' neural networks, which can capture complex, nonlinear, non-additive relationships (LeCun et al., 2015). Furthermore, we identify aspects of text processing (e.g., 'stop word' removal) that influence the performance of classification models—a facet of methodology that can significantly affect performance but is often overlooked (Ananiadou et al., 2009; Uysal & Gunal, 2014). Specifically, we address the following questions:

1. Can automated text-classifiers accurately identify papers that are relevant to ecological datasets?
2. Which aspects of text processing influence the performance of automated classifiers?
3. Are there trade-offs, in relation to performance and scrutability, when comparing state-of-the-art 'black-box' neural networks to simpler, more tractable LR classifiers?
4. Can the application of these models increase the rate at which relevant literature is identified?

2 | METHODS

2.1 | Data

The use of full-text articles can be preferable in terms of accuracy (Westergaard et al., 2018), but restricts the number of documents on which a classifier can be trained and subsequently applied. We therefore focus on the initial article-screening stage, and a method that uses only the titles and abstracts of scientific texts (see Supporting Information Appendix S1 for further details).

Using our two example databases (LPD and PREDICTS), we define relevant texts as articles that have an English abstract, and have contributed data to, or have been identified as likely to contain data for that database. We identified 633 such records linked with the LPD and 536 with the PREDICTS database. Using these databases allowed us to test and explore our methods, but they could be applied to any such database.

We downloaded the top 125,000 'ecology' articles from the National Center for Biotechnology Information, using the Entrez Programming Utilities (Sayers, 2010); these served as irrelevant texts. For each database, we took a random sample of irrelevant records equal in size to the number of relevant records, with any papers known to contribute to the focal database being excluded from sampling (i.e., papers contributing data to the LPD could not be irrelevant for the LPD but could be for PREDICTS and vice versa). Combining relevant and irrelevant records yielded 1,266 titles and abstracts for the LPD and 1,072 for the PREDICTS database that were used to train and test the classifiers.

2.2 | Text classifiers: construction, training, testing and analysis

We compared two binary classification techniques: logistic regression (LR) representing a strong but easy-to-create baseline while a convolutional neural network (CNN) offers a leading-edge alternative (Zhang & Wallace, 2015). For each method, the specific text-processing stages were varied to assess how these factors impacted the performance of the classifiers. Figure 1 summarizes the computational workflow, a detailed description of which can be found in Supporting Information Appendix S1.

To assess classifier performance, we used 10-fold cross validation and average area under the receiver operating characteristic curve (AUC) scores (LeDell et al., 2015). Generalized linear models (GLMs) were used to determine the influence of different workflow choices on classifier performance and we retained the best models for subsequent testing and application (see Supporting Information Appendix S1.3 for details).

Although improving data discovery for a single database has value, the broader potential of a text classifier depends on how well and how readily it can be transferred to other biodiversity databases. We optimized our text-processing workflow using the LPD texts before applying the best procedures to the PREDICTS data, providing an example of how such techniques can be readily transferred to various databases.

2.3 | Comparing classifiers to search engines

To compare the data discovery rate of our workflow with a search engine, we conducted targeted literature searches for each of the two example databases (Supporting Information Appendix S1.5 and Table S1.3). Articles from each search were ordered separately according to their relevance, as predicted by the search engine (Scopus for LPD, Web of Science for PREDICTS) or our best classifiers (see Results). For each ranked list, RC spent 15 min manually classifying papers as relevant or not based on the content of their titles and abstracts. Binomial mixed-effects models were then used to compare the effect of ranking type (search engine or model) on the proportion of relevant articles found for each database. Search topic (the queries used to conduct each literature search) was specified as a random intercept and equivalent models were also fitted to the

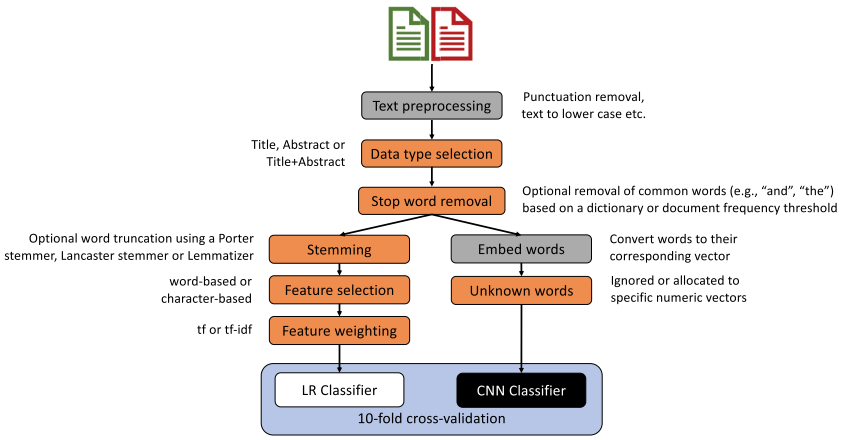

FIGURE 1 Graphical depiction of text-processing workflow and the classifier training. Orange boxes indicate stages of the workflow that were systematically varied; grey boxes represent processing that was constant across all indicated models. For details of the text-processing stages, see Supporting Information Appendix S1 and Table S1.1. Stemming reduces words to their root, for example, 'ecological' would be shortened to 'ecolog'. tf = term frequency (the number of times a term occurs in the text being considered); tf-idf = term frequency-inverse document frequency whereby the term frequency is multiplied by the term's inverse document frequency (idf = $log\left(\frac{D}{d}\right)$, where D is the total number of documents in the training corpus and d the number of documents in which the term of interest occurs); LR = logistic regression; CNN = convolutional neural network

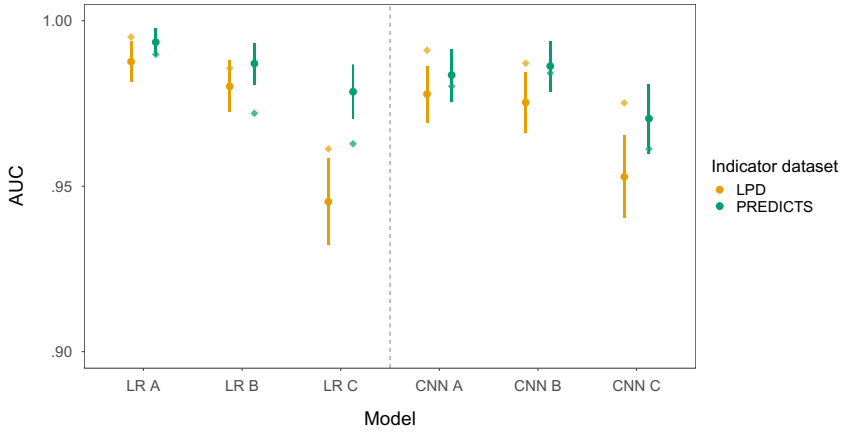

FIGURE 2 Cross-validation and test AUC scores for selected models. All models display strong performance on both the LPD and PREDICTS texts. The LR A and CNN A models make use of abstracts and a number of text-processing stages. Foregoing the text-processing stages (B models) causes model performance to drop only slightly, or improve in the case of the PREDICTS-trained CNN B. However, not using abstracts (C models) leads to a larger performance decrease, especially for the LPD texts. Note truncated *y* axis starts at .90. Circles and error bars show the mean and 95% confidence intervals, respectively, of the AUC scores from 10-fold cross-validation. Diamonds represent the test AUC scores. AUC = area under receiver operating characteristic curve; LR = logistic regression; CNN = convolutional neural network; LPD = Living Planet Database; PREDICTS = Projecting Responses of Ecological Diversity In Changing Terrestrial Systems

TABLE 1 Summary of configurations for the models presented in Figure 2

	LR A	LR B	LR C	CNN A	CNN B	CNN C
Data type:	Title + Abstract	Title + Abstract	Title	Title + Abstract	Title + Abstract	Title
Stop words:	df > .85	None	df > .85	NLTK	None	NLTK
Stemmer:	Lancaster	None	Lancaster			
Feature:	Word-based	Word-based	Word-based			
Weighting:	tf-idf	tf	tf-idf			
Unknown words:				Random	Removed	Random

Note: A = models identified as best using AUC (area under the receiver operating characteristic curve); B = models equivalent to A but using simpler text-processing; C = models equivalent to A but not using abstract text; LR and CNN = logistic regression and convolutional neural network models, respectively; NLTK = the Natural Language Toolkit stop word list; df = document frequency; tf = term frequency; tf-idf = term frequency-inverse document frequency.

manual classifications generated after the first 10 and 5 min. Ten per cent of the manually classified papers were sampled at random and double-checked by experienced members of the LPD and PREDICTS teams to determine the level of agreement between the manual categorizations. Mixed-effects models were re-fitted using the expert classifications to assess how any re-classifications affected the coefficient estimates for ranking type.

2.4 | Potential for iterative improvement of classifiers

Performance is expected to improve with the size of the training set (Liu et al., 2019), and the speed of improvement is an important determinant of the general usefulness of an approach. To explore whether iteratively expanding the classifier training data has the potential to improve the classifiers, we used cross-validation and AUC scores to separately assess how the size of the training dataset and addition of new texts from literature searches influence the predictive performance of the models used in 2.3 (see Supporting Information Appendices S1.6 and S1.7 for details).

2.5 | Insight into classifier decision making

Within the LR models, each term in the training texts—for example, an individual word or word stem—is associated with a learned weight (see Supporting Information Appendix S1.1.1 for details). To identify the terms having the most influence on predicted relevance, term weights were extracted from the best-performing LR models. The 50 most positively and 50 most negatively weighted terms were inspected to see if they could cause biases in the classifications.

3 | RESULTS

Overall, AUC scores indicate that both the LR and neural network models performed very well, with little difference between the approaches. Among the text-processing choices tested (Figure 1), the type of text data is the most important factor influencing classification performance, for both models (Figure 2, Supporting Information Figure S5.4 and Table S4.5). For example, when considering the LPD-trained models, average AUC drops from .988 for the best logistic model (LR A: using titles and abstracts, see Table 1 for details) to .945, if abstracts are not considered (LR C: using only titles). The workflow developed using the LPD also performs as well or better on the PREDICTS texts.

Additional metrics calculated on the labelled test data, and thus likely representing the upper limits of model performance, indicate that if 100 relevant texts were present in a corpus, 95 would be labelled as such by the LR A model; and that for every 100 texts labelled as relevant, 98 actually would be, demonstrating high recall and precision, respectively (see Supporting Information Table S4.6 for metrics associated with the selected models).

Ranking search results using the LR A classifier led to a significantly higher proportion of potentially relevant papers being discovered after 15, 10 and (for the LPD searches) even 5 min than if the search engine rankings were used (Figure 3 and Supporting Information Table S4.7). For example, when manually screening LPD-related searches for 10 min, use of the classifier increased the average proportion of relevant papers found from .48 to .65. Experienced database users (LPD and PREDICTS team members) agreed with RC's manual classifications in 87% (47/54) and 95% (37/40) of cases, respectively. The positive effects of the classifier on discovery rate increased slightly when using the expert classifications of the sampled texts in combination with the rest of RC's classifications (Supporting Information Figure S5.5), suggesting that the benefits of using the LR A models are not driven by any initial classification errors.

Larger training datasets enhance predictive performance of LR A-style classifiers but with diminishing returns. Furthermore, even models trained with just 200 texts achieve average AUC ≥ .98 (Figure 4a). Expanding the training data to include texts identified during the literature screening also substantially improves the performance of the LR A-style classifiers on real-world search results. Up-weighting new negatives relative to the original negatives produces the best performance (Figure 4b).

Generally, the most positively weighted terms are associated with the respective indicator database; for example, 'pop' and 'abund' for the LPD and 'specy' (stemmed form of 'species') and 'landscap' for PREDICTS (Figure 5 and Supporting Information Table S4.8). The most negatively weighted terms for both datasets represent topics in ecology less connected to either the LPD or PREDICTS, such as 'evolv'. The greater similarity of negative terms across the dataset-specific models is illustrated by the fact that whilst 21 terms are shared between the 50 most negatively weighted features for the LPD and PREDICTS models, only 9 are when considering the 50 most positive terms. Interestingly, there are some terms that stand out as potential artefacts of biases in the training texts, for example, 'declin' for the LPD and 'forest' for PREDICTS.

4 | DISCUSSION

Collating ecological data is essential for understanding the natural world and how it is affected by anthropogenic activity. Macroecological datasets in particular are critical for exploring the extent to which impacts of such activity can be generalized across space and taxa. We have shown that by using text mining and automated classifiers we can speed up the identification of newly

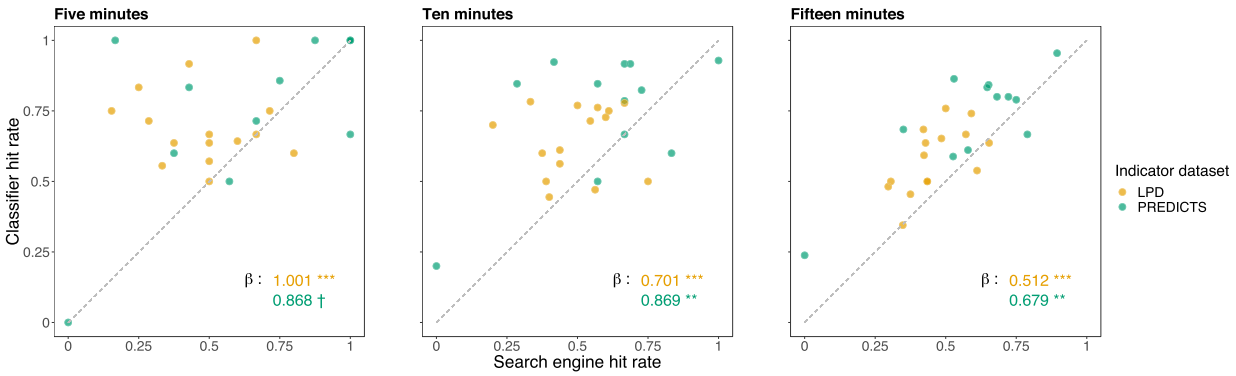

FIGURE 3 A comparison of the proportion of papers that are manually classified as relevant, when working through papers according to the search engine or our best logistic regression model. Using the automated classifiers is beneficial across all timespans (positive β values and most points above dashed grey line). β values represent the impact of using the classifiers compared to the search engine ranking. Significance is indicated ($^{\dagger}p < .1$, $^{**}p < .01$, $^{***}p < .001$) from two-tailed tests. Residual degrees of freedom are 21 for LPD related models and 27 for PREDICTS. The 1:1 dashed grey line represents a scenario where the different rankings lead to the same proportion of relevant papers being found. LPD = Living Planet Database; PREDICTS = Projecting Responses of Ecological Diversity In Changing Terrestrial Systems. See Supporting Information Table S1.3 for a breakdown of search terms and the number of papers returned

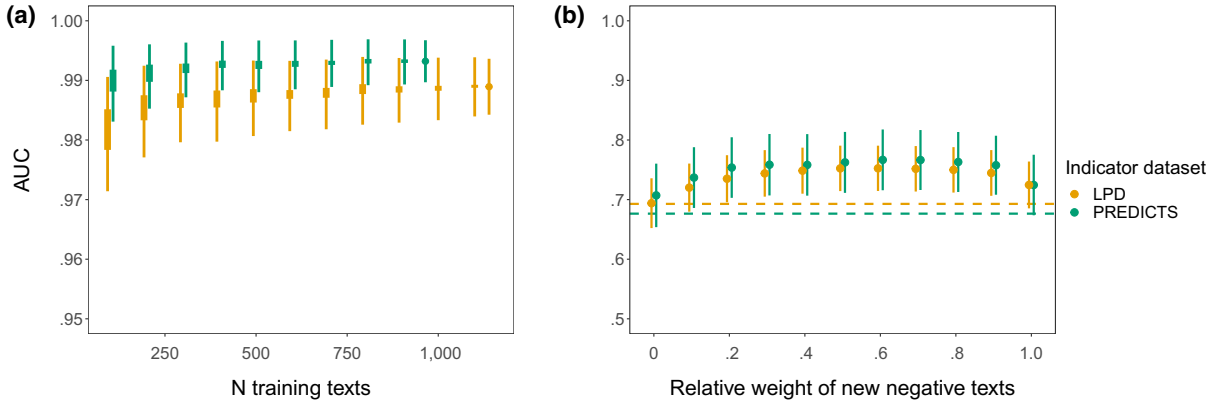

FIGURE 4 The impact of training set size and sample weighting on classifier performance. (a) When using LR A specified models, larger training sets improve predictive performance. However, a plateau is apparent indicating that beyond approx. 1,000 texts, little additional improvement is made. Crucially, even when using training sets of 200 texts (100 positives) average AUC scores exceed .98 for LPD classifiers and are around .99 for PREDICTS classifiers. (b) Incorporating new texts demonstrably boosts classifier performance when compared to the original (dashed lines). Up-weighting new negatives relative to original negatives confers benefits, with optimal relative weighting being around .6–.7. Note different axis limits in (a) and (b). In (a), the thick lines show the range of mean AUC values calculated over 10 replicates of 10-fold cross-validation, the associated thin lines show the range of the 95% confidence intervals. The circles and error bars show the mean and 95% confidence intervals, respectively, of the AUC scores from 10-fold cross-validation of the complete datasets, as reported in Figure 2. In (b), circles and error bars show the mean and 95% confidence intervals, respectively, of the AUC scores from cross-validation incorporating new texts from the literature searches. AUC = area under receiver operating characteristic curve; LPD = Living Planet Database; LR A = best logistic regression model; PREDICTS = Projecting Responses of Ecological Diversity In Changing Terrestrial Systems

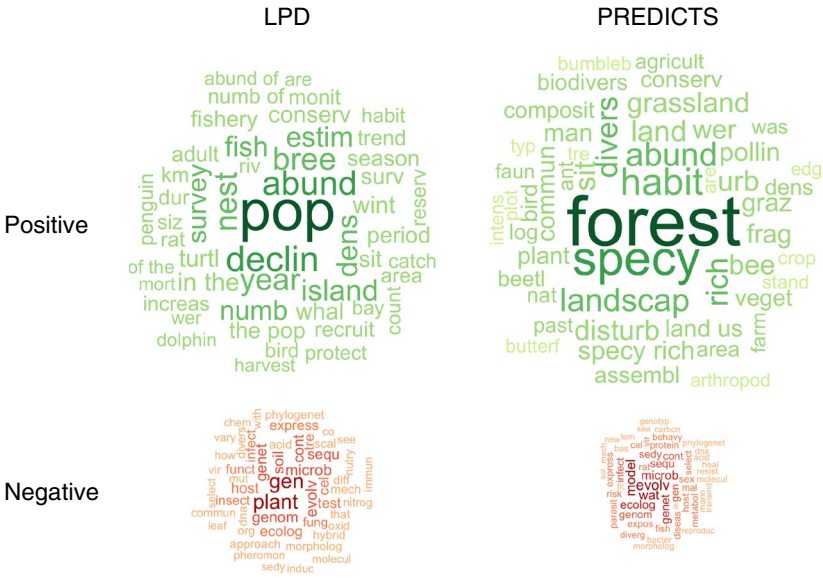

FIGURE 5 Word clouds showing the 50 most positively and negatively weighted features from the best performing logistic regression models trained on LPD and PREDICTS texts. Larger, darker font is associated with higher absolute weighting and thus greater influence on a text's predicted relevance. Note: 'specy' is the result of applying a Lancaster stemmer to 'species'. LPD = Living Planet Database; PREDICTS = Projecting Responses of Ecological Diversity In Changing Terrestrial Systems

published data relevant to two example databases: the Living Planet Database and the PREDICTS database. The real-world performance of the trained classifiers can also be further improved by incorporating new relevant and irrelevant texts identified during the manual screening of targeted literature searches. Combined with the fact that even models trained with just 100 relevant texts have high predictive power, our findings suggest that the demonstrated techniques can assist in the creation and growth of other ecological datasets, such as BIOFRAG (Pfeifer et al., 2014) and BioTIME (Dornelas et al., 2018). We therefore encourage others to use and improve upon our work to collate urgently needed biodiversity data.

Biodiversity indicators represent an important output of ecological datasets whose usefulness is maximized if they are derived from up-to-date information (Collen et al., 2008; Walpole et al., 2009). The growing availability of dashboards and portals serving indicators and similar derived biodiversity data products—for example, The Biodiversity Indicators Dashboard (Han et al., 2014) and Map of Life (Powers & Jetz, 2019)—further attests to the need for a dynamic view of the state of nature. Our

technique can potentially reduce the lag time between publication, discovery and incorporation of new data into such established indicators.

4.1 | Overall performance of classifiers

Our best classifiers have accuracy, precision and recall that compare favourably with studies using automated text classification to identify papers relevant to medical reviews, which typically report similar (Adeva et al., 2014; Ananiadou et al., 2009; Bannach-Brown et al., 2019) or lower values (Cohen et al., 2012). Although one might naively expect the more complex neural network to outperform the simple logistic classifier (Joulin et al., 2017), we did not find that to be the case here. Researchers may therefore choose a classification technique based on other considerations, for example, bias assessment/mitigation where LR classifiers offer higher levels of transparency concerning model 'decision making' than do 'black-box' neural networks (see 4.3).

Incorporating abstracts, rather than just using titles, substantially improves classifier performance. Adeva et al. (2014) found the same qualitative pattern, and Westergaard et al. (2018) demonstrated that text mining biomedical literature is significantly better if using full-text articles rather than abstracts. These findings all suggest that text mining benefits from the greater information content of longer texts. While a classifier trained on full texts may display increased performance, the articles available for subsequent screening would be smaller due to current access limitations such as pay-walls and copyright issues, though the increasing prevalence of open-access publishing means that these limitations may be transient.

By optimizing our workflow using the LPD texts and then transferring the identified procedure to the PREDICTS data, we demonstrate the general applicability of our methods. Given that PREDICTS-trained models perform at least as well as those trained on the LPD data (Figures 2 and 3), applying our approach to databases like BioTIME (Dornelas et al., 2018) could quickly help identify additional relevant data sources.

4.2 | Limitations

Classifiers that we did not consider may perform differently and/or be more sensitive to text-processing procedures. We have also not addressed how the architecture of deep-learning networks could influence model performance. While a thorough exploration of CNN hyperparameters would be expected to improve performance (Zhang & Wallace, 2015), the principal aim of this paper has been to develop the use of text-mining techniques within ecological data collation workflows and demonstrate their potential benefits. Having assessed both a strong baseline (LR using bag-of-features) and a leading-edge option (deep learning with word vector representation) (Joulin et al., 2017), our work shows the usefulness of such techniques within ecology.

4.3 | Future development

Concerns have been raised recently that machine-learning models may contain bias, primarily due to being trained on imperfect data (Bolukbasi et al., 2016; Tramer et al., 2017). Given the imbalances that exist within ecological datasets (Gonzalez et al., 2016; McRae et al., 2017), text classifiers like ours could propagate bias, as evidenced by the strong influence of certain terms in the logistic models, for example, 'forest' and 'fish'. The accumulation of biases within biodiversity datasets is detrimental to their scientific goals (Gonzalez et al., 2016). Consequently, there is a clear need to assess classifiers carefully for bias prior to their widespread application and to check the representativeness of any subsequently acquired data to mitigate this risk. Further research into this area, especially with regard to biodiversity data coverage, could provide substantial insight into these issues and how best to combat them effectively.

One potential solution could involve technical developments of the text-mining process to ignore specified 'bias terms' and/or preferentially return information associated with entities (e.g., taxa and locations) that are currently under-represented in the focal dataset. Text-mining techniques could therefore not only increase the rate at which data are incorporated into biodiversity assessments but might also contribute to making ecological databases more representative of reality, to better inform conservation and policy decisions.

Although the methods discussed here can help researchers collate available biodiversity data, in the longer term, it is also critical that published/collected ecological data are made more accessible for use in syntheses (McMahon et al., 2011; Poisot et al., 2019). Approaches to facilitate this include searchable, centralized repositories (similar to genetic sequence databases, e.g., GenBank; Benson et al., 2012), standardized data formats (Poisot et al., 2019), and/or the use of a machine-readable mark-up language within articles (Bourne et al., 2008). Crucially, such changes require strong incentives to ensure that the original data collectors receive appropriate recognition for their scientific contributions (Bourne, 2005; Ewers et al., 2019). Although a substantial challenge, these developments would enable the more complete and rapid synthesis of ecological data, which is essential for mitigating biodiversity loss and its associated challenges.

5 | CONCLUSION

We have shown that combining text mining and simple machine-learning classifiers is highly effective in identifying papers relevant to ecology datasets. We demonstrate this using two globally recognized biodiversity indicators, but our method is applicable to any dataset comprised of data from literature sources. Interestingly, even relatively simplistic models based on LR perform very well, on a par with more complex neural networks. The wider adoption of these techniques could therefore rapidly increase the rates of data discovery and collation across a wide range of ecological datasets. Removing the discovery bottleneck would substantially help

researchers to keep datasets up-to-date and representative of the natural world, both of which are critical for accurately monitoring conservation progress and informing policy. To facilitate further application and development we provide code for building and using such classifiers.

ACKNOWLEDGMENTS

This work was supported by the Natural Environment Research Council (grants NE/R012229/1 and NE/M014533/1). This paper is a contribution to the Grand Challenges in Ecosystems and the Environment initiative.

DATA AVAILABILITY STATEMENT

All relevant code and data can be found at https://zenodo.org/badge/latestdoi/302323925 and https://github.com/rcornf/geb_text_class_2020

ORCID

Richard Cornford https://orcid.org/0000-0002-9963-3603
Adriana De Palma https://orcid.org/0000-0002-5345-4917
Louise McRae https://orcid.org/0000-0003-1076-0874
Andy Purvis https://orcid.org/0000-0002-8609-6204

REFERENCES

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, *41*(4), 1498–1508.

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*, *27*(4), 509–523.

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, *8*(1), 23.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems* (ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett), Vol. *29*, pp. 4349–4357. Curran Associates, Inc, Barcelona.

Bourne, P. E. (2005). Will a biological database be different from a biological journal? *PLoS Computational Biology*, *1*(3), e34.

Bourne, P. E., Fink, J. L., & Gerstein, M. (2008). Open access: Taking full advantage of the content. *PLoS Computational Biology*, *4*(3), e1000037.

CBD. (2010). COP 10 Decision X/2: The strategic plan for biodiversity 2011–2020, 10th Meeting of the Conference of the Parties to the Convention on Biological Diversity, Nagoya. 18–29 October 2010. https://www.cbd.int/decision

Cohen, A. M., Ambert, K., & McDonagh, M. (2012). Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making*, *12*(1), 33.

Collen, B., Loh, J., Whitmee, S., McRae, L., Amin, R., & Baillie, J. E. (2009). Monitoring change in vertebrate abundance: The Living Planet Index. *Conservation Biology*, *23*(2), 317–327.

Collen, B., & Nicholson, E. (2014). Taking the measure of change. *Science*, *346*(6206), 166–167.

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75–88. https://doi.org/10.1177/194008290800100202

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmaker, J., Benedetti-Cecchi, L., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Ewers, R. M., Barlow, J., Banks-Leite, C., & Rahbek, C. (2019). Separate authorship categories to recognize data collectors and code developers. *Nature Ecology & Evolution*, *3*(12), 1610. https://doi.org/10.1038/s41559-019-1033-9

Gonzalez, A., Cardinale, B. J., Allington, G. R. H., Byrnes, J., Arthur Endsley, K., Brown, D. G., Hooper, D. U., Isbell, F., O'Connor, M. I., & Loreau, M. (2016). Estimating local biodiversity change: A critique of papers claiming no net loss of local diversity. *Ecology*, *97*(8), 1949–1960. https://doi.org/10.1890/15-1759.1

Han, X., Smyth, R. L., Young, B. E., Brooks, T. M., Sánchez de Lozada, A., Bubb, P., Butchart, S. H. M., Larsen, F. W., Hamilton, H., Hansen, M. C., & Turner, W. R. (2014). A biodiversity indicators dashboard: Addressing challenges to monitoring progress towards the Aichi biodiversity targets using disaggregated global data. *PLoS One*, *9*(11), e112046. https://doi.org/10.1371/journal.pone.0112046

Hudson, L. N., Newbold, T., Contu, S., Hill, S. L. L., Lysenko, I., De Palma, A., Phillips, H. R. P., Alhusseini, T. I., Bedford, F. E., Bennett, D. J., Booth, H., Burton, V. J., Chng, C. W. T., Choimes, A., Correia, D. L. P., Day, J., Echeverría-Londoño, S., Emerson, S. R., Gao, D. I., … Purvis, A. (2017). The database of the PREDICTS (Projecting Responses of Ecological Diversity in Changing Terrestrial Systems) project. *Ecology and Evolution*, *7*(1), 145–188. https://doi.org/10.1002/ece3.2579

Hudson, L. N., Newbold, T., Contu, S., Hill, S. L. L., Lysenko, I., De Palma, A., Phillips, H. R. P., Senior, R. A., Bennett, D. J., Booth, H., Choimes, A., Correia, D. L. P., Day, J., Echeverría-Londoño, S., Garon, M., Harrison, M. L. K., Ingram, D. J., Jung, M., Kemp, V., … Purvis, A. (2014). The PREDICTS database: A global database of how local terrestrial biodiversity responds to human impacts. *Ecology and Evolution*, *4*(24), 4701–4735. https://doi.org/10.1002/ece3.1303

IPBES. (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (ed. by E. S. Brondizio, J. Settele, S. Díaz and H. T. Ngo). IPBES secretariat, Bonn.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (ed. by M. Lapata, P. Blunsom and A. Koller), pp. 427–431. Association for Computational Linguistics, Valencia.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, *1*(1), 4–20.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*, 3–24.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, *535*(7612), 457–458.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436. https://doi.org/10.1038/nature14539

LeDell, E., Petersen, M., & van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic Journal of Statistics*, *9*(1), 1583. https://doi.org/10.1214/15-EJS1035

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loh, J., Green, R. E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., & Randers, J. (2005). The Living Planet Index: Using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 289–295.

Mace, G. M., & Baillie, J. E. (2007). The 2010 biodiversity indicators: Challenges for science and policy. *Conservation Biology*, *21*(6), 1406–1413. https://doi.org/10.1111/j.1523-1739.2007.00830.x

McMahon, S. M., Harrison, S. P., Armbruster, W. S., Bartlein, P. J., Beale, C. M., Edwards, M. E., Kattge, J., Midgley, G., Morin, X., & Prentice, I. C. (2011). Improving assessment and modelling of climate change impacts on global terrestrial biodiversity. *Trends in Ecology and Evolution*, *26*(5), 249–259. https://doi.org/10.1016/j.tree.2011.02.012

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS ONE*, *12*(1), e0169156. https://doi.org/10.1371/journal.pone.0169156

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L., Hoskins, A. J., Lysenko, I., Phillips, H. R., & Burton, V. J. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science*, *353*(6296), 288–291.

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., & Fei, S. (2016). Automated content analysis: Addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, *7*(11), 1262–1272. https://doi.org/10.1111/2041-210X.12602

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, *4*(1), 5.

Pfeifer, M., Lefebvre, V., Gardner, T. A., Arroyo-Rodriguez, V., Baeten, L., Banks-Leite, C., Barlow, J., Betts, M. G., Brunet, J., Cerezo, A., Cisneros, L. M., Collard, S., D'Cruze, N., da Silva Motta, C., Duguay, S., Eggermont, H., Eigenbrod, F., Hadley, A. S., Hanson, T. R., … Ewers, R. M. (2014). BIOFRAG—A new database for analyzing BIOdiversity responses to forest FRAGmentation. *Ecology and Evolution*, *4*(9), 1524–1537. https://doi.org/10.1002/ece3.1036

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., & Peres-Neto, P. (2019). Ecological data should not be so hard to find and reuse. *Trends in Ecology and Evolution*, *34*(6), 494–496. https://doi.org/10.1016/j.tree.2019.04.005

Powers, R. P., & Jetz, W. (2019). Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios. *Nature Climate Change*, *9*(4), 323–329. https://doi.org/10.1038/s41558-019-0406-z

Purvis, A., Newbold, T., Palma, A. D., Contu, S., Hill, S. L., Sanchez-Ortiz, K., Phillips, H. R., Hudson, L. N., Lysenko, I., Börger, L., & Scharlemann, J. P. (2018). Modelling and projecting the response of local terrestrial biodiversity worldwide to land use and related pressures: The PREDICTS project. In D. A. Bohan, A. J. Dumbrell, G. Woodward, & M. Jackson (Eds.), *Next generation biomonitoring: Part 1* (Vol. *58*, pp. 201–241). Academic Press. https://doi.org/10.1016/bs.aecr.2017.12.003

Roll, U., Correia, R. A., & Berger-Tal, O. (2018). Using machine learning to disentangle homonyms in large text corpora. *Conservation Biology*, *32*(3), 716–724. https://doi.org/10.1111/cobi.13044

Sayers, E. (2010). A General Introduction to the E-utilities. *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US), Bethesda, MD. https://www.ncbi.nlm.nih.gov/books/NBK25497/

Scholes, R. J., & Biggs, R. (2005). A biodiversity intactness index. *Nature*, *434*(7029), 45–49. https://doi.org/10.1038/nature03289

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., & Lin, H. (2017). FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 401–416).

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112.

Walpole, M., Almond, R. E., Besançon, C., Butchart, S. H., Campbell-Lendrum, D., Carr, G. M., Collen, B., Collette, L., Davidson, N. C., Dulloo, E., & Zimsky, M. (2009). Tracking progress toward the 2010 biodiversity target and beyond. *Science*, *325*(5947), 1503–1504.

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Computational Biology*, *14*(2), e1005962.

WWF (2020). *Living Planet Report 2020: Bending the curve of biodiversity loss* (ed. by R. E. A. Almond, M. Grooten and T. Peterson). WWF, Gland.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 253–263. Asian Federation of Natural Language Processing, Taipei.

BIOSKETCH

Richard Cornford is a PhD student interested in the use and development of biodiversity indicators for informing conservation policy. His work focuses on methods for improving the representation of indicator databases and better understanding how current biases in ecological data may influence our ability to predict biodiversity trends.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Cornford R, Deinet S, De Palma A, et al. Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Global Ecol Biogeogr*. 2020;00:1–9. https://doi.org/10.1111/geb.13219