INTRODUCTION

Addressing rapid environmental changes has quickly increased the body of literature on global ecological changes. Systematic literature reviews are important for synthesizing accumulated evidence and informing decision-making and public relations research gaps or publishing bias. Traditional literature synthesis for extracting information from high volumes of text data are expensive, inefficient, and not easily replicable. There is a growing need for improved methods that can rapidly and efficiently synthesize the ever-growing body of literature.

2 METHODS

1. Screen articles for eligibility using inclusion/exclusion criteria.
2. Review and classify articles manually and extract data.
3. Train machine learning models using manually classified data.

We conducted a systematic literature review using the Web of Science search engine to identify studies that documented one or more direct threats to inland fisheries at a basin scale. We then performed initial eligibility screening to identify articles describing one or more documented, direct threats. Manually reviewed abstracts (n=4,336) served as the training and testing data for automated classification of the remaining abstracts (n=4,092) using machine learning by NLP. We used four types of text classification algorithms: binary logistic regression, naïve Bayes classification, linear single vector machine, and k-nearest neighbor and evaluated model performance using recall, precision, and F1 score.

3 RESULTS

The results of this study synthesize documented anthropogenic effects on fishes, highlight research gaps from the absence of studies linking certain drivers to direct impacts, and demonstrate opportunities for improving efficiency in literature reviews through integrated machine learning and natural language processing approaches. Results suggest machine learning may be most useful for eliminating extraneous literature during preliminary review steps and point to the need for improved refinement of machine learning processes and noise reduction in data feeds (e.g., earth sciences) where syntax may be less standardized or structured.

LITERATURE SELECTION

We screened 9,361 abstracts from 45 major river basins in the USA and found 1,008 distinct peer-reviewed journals from 1990 to 2020.

MACHINE LEARNING CLASSIFICATION

Of the abstracts screened by NLP (n=14,092) using the best performing model (linear regression), 9.9% (n=404) were classified as direct threats. These were classified with ~65% precision (i.e., positive predictive value). Manual review revealed 67% (n=251) were indeed correctly classified by NLP classification and were suitable for inclusion in our data review. The remaining abstracts were classified as direct threats included for review and 82.3% (n=3,488) were excluded.

4 DISCUSSION

This study advances the understanding of trends with initial literature synthesis using NLP for inland fisheries. Motivated by the need to better understand direct-impact associations of fish for the development of indicators and proxy measures of stressors to inland fisheries, this study compiled existing driver-response-impacts links in the literature and applied a paired method for literature classification.

CONCLUSIONS

1. Documented links suggest all direct drivers of anthropogenic change exist in all major fisheries; some relationships are strongly documented.
2. Documented links reflect publishing and research bias by basin locations and types of studies; these biases create noticeable gaps in the literature.
3. Despite relatively standardized abstract structure and length, there is still inherent human objectivity in the literature and variable quality of evidence and interpretability.
4. NLP may be most useful for improving efficiency in preliminary steps of fisheries literature synthesis (i.e., classifying extraneous or redundant articles not pertinent to the study).
5. NLP and automated machine learning performance for fisheries and ecology literature may benefit from integrated noise reduction techniques.

BIBLIOGRAPHY


KEY REFERENCES


Acknowledgments

We thank Chelsie Romulo for conceptualization of machine learning application and use, Siddharth Ambalik for technical assistance with manual including label and construct review feedback. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. NSF DGE 1144462. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.