Classecol: classifiers to understand public opinions of nature

Received: 1 December 2020 | Accepted: 18 February 2021

DOI: 10.1111/2041-210X.13596

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

classecol: Classifiers to understand public opinions of nature

Thomas F. Johnson | **Hebe Kent** | **Bethan M. Hill** | **Georgia Dunn** | **Leonie Dommett** | **Natasha Penwill** | **Tom Francis** | **Manuela González-Suárez**

Ecology and Evolutionary Biology, School of Biological Sciences, University of Reading, Reading, UK

Correspondence
Thomas F. Johnson
Email: Thomas.frederick.johnson@outlook.com

Funding information
Natural Environment Research Council, Grant/Award Number: J71566E; The Royal Society, Grant/Award Number: IE160539

Handling Editor: Laura Graham

Abstract

1. Human perceptions of nature, once the domain of the social sciences, are now an important part of environmental research. However, the data and tools to tackle this research are lacking or are difficult to apply.
2. Here, we present a collection of text classifier models to identify text relevant to the broad topics of hunting and nature, describing whether opinions are pro- or against-hunting, or show interest, concern or dislike of nature. The methods also include a biographical classification—describing whether the author of the text is a person, nature expert, nature organisation or 'Other'. The classifiers were developed using an extensive social media dataset, and are designed to support qualitative analysis of big data (especially from Twitter).
3. The classifiers accurately identified biographies, text related to hunting and nature and the stance towards hunting and nature (weighted *F*-scores: 0.79–0.99; 1 indicates perfect accuracy).
4. These classifiers, alongside an array of other text processing and analysis functions, are presented in the form of an R package classecol. classecol also acts as a proof of concept that nature-related text classifiers can be developed with high accuracy.

KEYWORDS

conservation, cultural ecosystem services, culturomics, human perceptions, human–nature relationship, sentiment analysis, text classifier

1 | INTRODUCTION

Ecology has become more transdisciplinary to better understand our environment. For example, ecosystem services reflect health, economic and cultural values (Kareiva et al., 2011), and journals and societies want to study human relationships with nature (Gaston et al., 2019; Society for Conservation Biology working groups, 2020). This transdisciplinary shift has brought the human dimension of nature into focus, but the study of human–nature relationships largely falls outside the traditional expertise of an ecologist or conservationist, who may be unfamiliar with the available methods and data.

Social media could help us understand human–nature relationships. Historically, surveys (or other qualitative approaches) have assessed perceptions, often providing a detailed understanding of a person's thoughts. Social media does not offer such detail, but is cost-effective, less time-intensive and offers enormous amounts of information (Fox et al., 2020). In 2020, social media has become widely used in most countries, with approximately half of the world's population (and increasing) being active users (Clement, 2020). Social media captures many data types (e.g. text, photos, videos, sound and interaction networks with other people) with spatial representation and temporal time series that could allow holistic analyses (Toivonen et al., 2019).

In recent years, the use, and diversity of uses, of social media analysis across the environmental sciences has rapidly increased (Ghermandi & Sinclair, 2019). Social media has been used to develop species distribution models (August et al., 2020), measure aesthetic and recreational ecosystem services (Graham & Eigenbrod, 2019; Van Zanten et al., 2016), track illegal wildlife trade (Di Minin et al., 2018) and determine the role of wildlife in nature-based tourism (Hausmann et al., 2017). The abundance and availability of data on these platforms—many now 15 years old, open the door for more research. Analyses of social media could revolutionise our understanding of the human—nature relationship and how it impacts the environment, but this requires new and improved tools (Toivonen et al., 2019).

There are many approaches to 'mine' opinions and gain insights from text data (Aggarwal & Zhai, 2013). For example, sentiment analysis aims to understand the emotion of a text, often classifying the text's language use as negative, neutral or positive (Liu, 2020). This can be done with machine learning approaches, but a more readily accessible approach for interested ecologists and conservationists would be lexicon-based sentiment analysis. Lexicon-based approaches assign scores to words to calculate an average score for a text passage, for example, if more negative words are used, the text will be labelled as negative. Overall scores are effective for describing sentiment, but meaning may be unclear (Aldayel & Magdy, 2019; Mohammad et al., 2017). For example, lexicon-based sentiment analyses would return negative scores for these two messages 'It is sad that Pangolin are vanishing' and 'Pangolins are bad' (both use negative language), failing to recognise that only the second message indicates a dislike for pangolins. Furthermore, in some lexicons, species names can have negative scores (e.g. 'shark') which can bias results if we are interested in human–nature relationships (Lennox et al., 2020).

Stance analysis is an alternative approach (Aggarwal & Zhai, 2013; Liu, 2020; Srivastava & Sahami, 2009), more targeted towards assessing opinions about topics or specific questions. Stance analysis could help recognize the dislike of pangolins in the example above, but this method is often time-consuming to develop as it requires large training datasets alongside complex machine learning models. Furthermore, the generality of the stance analysis models can be low. For example, if a stance analysis model was built to detect fondness of pangolins, it may be of limited use for other species. So whilst stance analysis gets far closer (relative to lexicon-based sentiment analysis) to understanding a user's opinion, for it to be useful, it would also need to be derived from a broad array of training data themes, and answer general and pertinent questions.

2 | CLASSECOL DESCRIPTION

With the massive growth in social media analysis, and especially in studies using text data to look at people's perceptions of and relationships with nature (Ghermandi & Sinclair, 2019), there is a great need for text analysis tools (Toivonen et al., 2019). To meet this demand, we present classecol a text cleaning, processing and classification tool to support analysis of public opinions of nature in a big data setting. classecol avoids the interpretation issues of sentiment analysis and the specificity issues of stance analysis. classecol can identify relevant texts, describe their stance and determine the type of user that produced the text. This provides a proof of concept to guide and encourage further text analysis development for ecology, and we hope other groups developing classifiers would consider uploading them to our package—becoming formal contributors (see package VIGNETTE). classecol's 10 text classifiers have been trained and tested on Twitter data, and fall within three topics:

1. Hunting—Are texts discussing the hunting of wildlife? If so, what's the user's opinion, for example, pro or against hunting?
2. Nature—Are texts relevant to nature? If so, what's the user's opinion, for example, expressing interest, concern or dislike of nature?
3. Biographical (bio hereafter)—Is the author of the text a person? If so, is that individual a member of the general public or an individual discussing nature in a professional or academic capacity?

3 | DEVELOPING CLASSIFIERS

Prior to developing the ten classifiers in the classecol collection, we developed base classifiers for each of the three topics following eight steps: (a) Defined a protocol to describe the criteria text must meet to fall in a category (e.g. What text characteristics distinguish pro- and against-hunting?). (b) Ensured the human classifiers could accurately and consistently use the protocol. (c) Seven individuals classified 1,100 texts for each topic (tweets for hunting and nature, and user provided descriptions for bio) creating a training dataset of 7,700 texts per topic. (d) Built six text classification models for each topic including multinomial logistic regression, support vector machines, naïve Bayes, random forest, *K* nearest neighbour and a four-layer neural network. A logistic regression was then used to merge the outputs from these models generating an ensemble text classifier. (e) Tested the performance of the ensemble model and identified cases of misclassification to refine the protocol and classification criteria. (f) Corrected misclassified training texts using the refined protocol. (g) Finalised the classification protocol. (h) Tested different text cleaning options (e.g. from raw text to very clean text—see Table S1) to identify that which maximised ensemble model precision and recall (both defined below). These eight steps are further detailed in Supporting Information: Developing classifiers.

In the final protocol, there are three categories for the hunting topic and four for the nature and bio (one added during the reclassification steps) topics:

Hunting

1. Irrelevant—text does not discuss the hunting of animals.
2. Pro-hunting—text indicates support for hunting.
3. Against-hunting—text indicates opposition to hunting.

Nature

1. Irrelevant—text does not discuss nature or nature related activities.
2. Pro-nature (positive phrasing)—text endorses nature with positive language, for example, interest.
3. Pro-nature (negative phrasing)—text endorses nature with negative language, for example, concern.
4. Against-nature—text indicates opposition or frustration towards nature, for example, fear.

Bio

1. Expert—user has professional status, or qualifications to indicate expertise, in nature or a nature related field.
2. Person—user is an individual without nature expertise.
3. Nature org (added)—user is an organisation, company or group working in a nature-related activity.
4. Other—user is none of the above.

4 | CLASSIFIER ACCURACY

We report the *F*-score (Zhang & Zhang, 2009) accuracy of each category in each classifier, and an overall accuracy per classifier (average *F*-score weighted by the proportional abundance of each category). Accuracy was measured on an independent data sample, that is, not used to develop the classifiers. $F = 1$ indicates perfect classification.

The hunting classifiers had high overall (0.87–0.97) and category accuracies (Figure 1), except for Irrelevant, where lower accuracy (0.64–0.72) was driven by low recall (0.54–0.61). Nearly half of the Irrelevant texts were assigned to the wrong category. In the nature classifiers, overall accuracies ranged from 0.82 to 0.92, with moderate to high accuracy across all categories except Pro-nature (negative phrasing) and Against-nature in the 'full' model. Against-nature had low model recall (0.67) and precision (0.4), probably because this category only represented 1.1% of all classifications. This low coverage could make the model unreliable, which may explain why Pro-nature (negative phrasing) also had low accuracy in the 'full' model, despite good accuracy in other models. Given this finding, we removed Against-nature from the stance and trimmed models and would recommend using the trimmed over the full model. Finally, in the bio models, overall accuracies ranged from 0.79 to 0.87, with moderate to high accuracy in all categories. All topics are characterised in Figures S6–S8.

5 | USING CLASSECOL

Prior to data collection and analyses, any research project involving public opinion should consider the legal and ethical requirements—see Data rights and ethics in the Supporting Information.

The classecol functions fall into two groups: (a) general text cleaning and analysis and (b) text classification. The first group includes five functions of value for anyone interested in natural language processing. The clean function provides comprehensive text cleaning options, including the conversion of common emoticons, abbreviations, slang and environment-related hashtags into readable text. valence detects the presence of terms that can alter, reverse or amplify meaning. contract performs word stemming and lemmatisation to reduce term complexity (e.g. consulting becomes consult). lang_eng detects the presence of non-English terms. Finally, senti_matrix pulls together 11 popular sentiment analysis approaches into one function, to produce a matrix of average sentiment scores for

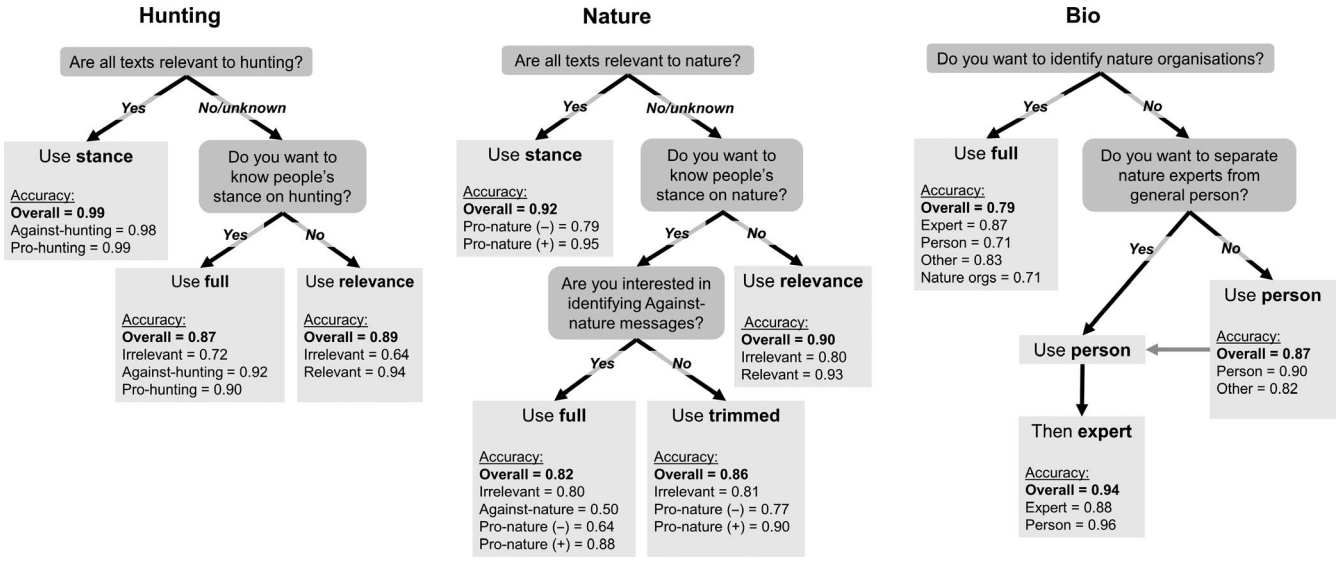

FIGURE 1 Flowchart to assist in selecting a suitable classecol classifier for each of the hunting, nature and bio topics. Flowchart questions are depicted in dark grey boxes with rounded edges, and classifier options are in the lighter shade of grey. The bold text in the classifier boxes describes the classifiers name and overall accuracy. Accuracy (measured as the classification *F*-score, a value of 1 is perfect classification accuracy) is also broken down into each classifier category

each sentence. All of these functions can be used in conjunction, for example, to assess sentiment analysis of some text, you may use lang_eng to remove non-English texts, then clean and contract the text, before running the senti_matrix function.

Our second group of functions are the most important component of classecol. These text classifiers are processed through a Python backend, thus require downloading and installing Python (we recommend version 3.6). This can be done automatically in R through the addeR::py_download function (Johnson, 2021a). The load_classecol function then automatically downloads the text classification models and Python module dependencies. load_classecol also links R to the Python backend and needs to be run every time a new R environment is loaded; the text classification models and Python modules will only need to be downloaded once. The hun_class, nat_class and bio_class functions perform the text classifications in the hunting, nature and bio topics respectively. Prior to using the classifiers, we recommend running clean(level = "simple") for hun_class and clean(level = "full")- for nat_class, but no cleaning is required before using bio_class. nat_class also requires a matrix of valence and language indicators, as well as sentiment scores for each text record (see package VIGNETTE on https://github.com/GitTFJ/classecol).

The hun_class, nat_class and bio_class functions each contain multiple text classifiers which could be valuable in different scenarios (Figure 1). For hun_class, the relevance model identifies whether text is relevant or irrelevant to hunting, stance classifies relevant texts as pro- or against-hunting and full runs both relevance and stance. Similarly, for nat_class, relevance identifies whether text is relevant or irrelevant to nature, stance identifies whether relevant pro-nature texts are using positive or negative phrasing and the trimmed model combines both. nat_class also has a full model which includes the low-accuracy Against-nature category, which should be used with caution. Finally, for bio_class, the person model identifies whether a user is a person or not, expert classifies persons as nature experts or general public and full combines both and adds the additional 'Nature organisation' category.

Classifiers can be used hierarchically (e.g. use relevance followed by stance) rather than using the combined classifiers. This increased computational processing time but had little impact on accuracies, except in the bio model, where accuracy is improved by using the person classifier, followed by the expert classifier. Classifiers can also be stacked. For example, to explore the general public's stance towards hunting in the USA, we could remove non-English texts with lang_eng, identify members of the public with bio_class(type = "full") and then determine hunting stance with hun_class(type = "full"). When running any of the text classifiers, we recommend manually classifying a sample of your data, so classification accuracy can be determined.

classecol's suite of text processing, analysis and classifier functions can assist academics and policy-makers interested in exploring the human dimensions of nature in big data. This research theme, and in-turn classecol's value, extends far beyond the fields of ecology and conservation, with social scientists, human geographers and environmental scientists all working with human–nature relationship data. classecol provides evidence that moderate to high accuracies can be achieved from text classifiers and we hope this will inspire future classifier development (methods and code are openly available). Admittedly, there are time costs to consider as supervised classifiers like classecol require lengthy training datasets, which are laborious to compile, and as mentioned earlier, can lack generality. Whilst we have designed classecol across a broad array of training data themes,

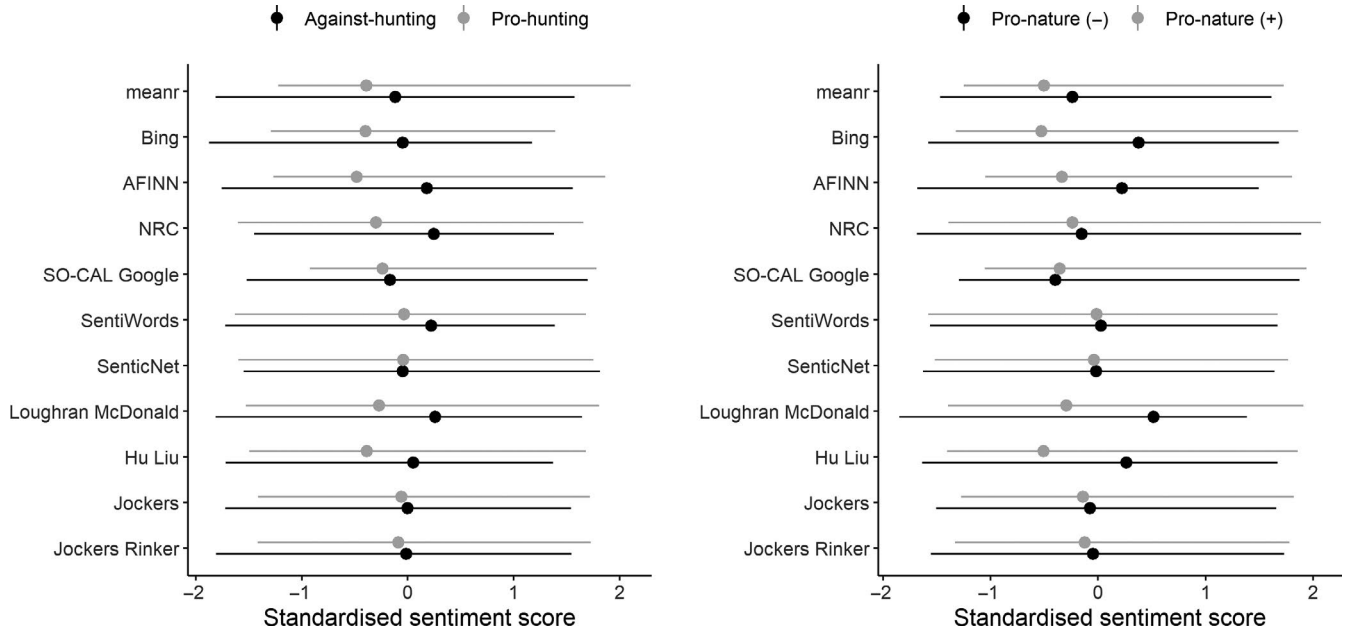

FIGURE 2 Assessment of 11 sentiment approaches ability to distinguish between the hunting and nature stances. The points represent the median sentiment score and error bars are the 95% quantiles [2.5%, 97.5%], displayed for each approach in each stance. If the sentiment analysis approaches were able to distinguish between the stances, we would expect to see little to no overlap in the black and grey points

its generality (or accuracy) across different data types is unknown. classecol should be used cautiously on non-Twitter data, and a sample of data must always be manually classified (by a human), so accuracy can be tested.

Despite hundreds of studies in the environmental sciences using social media analysis, there is a scarcity of method comparison and testing which means the accuracy and representativeness of these text analysis tools remains largely unknown, and could be error-prone. For example, when we measure sentiment analysis scores for texts in our human-classified hunting and nature stance data, we may expect sentiment analysis to detect the opposing hunting stances, or the opposing language use in pro-nature tweets, that is, Against-hunting tweets would primarily have negative scores, and Pro-hunting tweets would have positive scores. However, the sentiment scores between the categories largely overlap in both the hunting and nature topics (Figure 2). Sentiment approaches were unable to distinguish the classifications and detect our stances (lexicon-based sentiment analysis can only describe the text's polarity, not infer meaning). To ensure social media data are used robustly in the environmental sciences, its pivotal that methods are tested and frameworks for analysis are developed.

Big data culturomics within the ecological and conservation sciences are already reliant on transdisciplinary work involving social science. Transdisciplinary research is key to harnessing the data's massive potential, but requires careful method development and testing. This scrutiny extends onto classecol for which next steps include further testing of the text classifiers especially on non-Twitter data. The full potential of classecol, to our knowledge the first publicly available text classifier of opinions on nature, is yet to be explored, but we hope this tool will be the first of many in a growing community.

ACKNOWLEDGEMENTS

The authors thank R.C. and two reviewers for valuable feedback, and M.G. for the package name. T.F.J. thanks NERC (Natural Environment Research Council) Centre for Doctoral Training studentship (J71566E), and M.G.-S. thanks The Royal Society (IE160539) for funding.

AUTHORS' CONTRIBUTIONS

T.F.J. developed the classification protocol, which was reviewed by M.G.-S.; T.F.J., L.D., G.D., T.F., B.M.H., H.K. and N.P. labelled the training datasets; T.F.J. developed and refined the classification models, and prepared the first manuscript draft. All authors critically reviewed the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13596.

DATA AVAILABILITY STATEMENT

Twitter terms and conditions prevent sharing of the training data. Code to develop classification models at https://github.com/GitTFJ/classecol_dev and the classecol R package is located at https://github.com/GitTFJ/classecol. A static package version 0.4.0 is archived on Zenodo (Johnson, 2021b).

ORCID

Thomas F. Johnson https://orcid.org/0000-0002-6363-1825
Manuela González-Suárez https://orcid.org/0000-0001-5069-8900

REFERENCES

Aggarwal, C. C., & Zhai, C. X. (2013). *Mining text data* (1st ed.). Springer. https://doi.org/10.1007/978-1-4614-3223-4

Aldayel, A., & Magdy, W. (2019). Assessing sentiment of the expressed stance on social media. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-030-34971-4_19

August, T. A., Pescott, O. L., Joly, A., & Bonnet, P. (2020). AII naturalists might hold the key to unlocking biodiversity data in social media imagery. *Patterns*. https://doi.org/10.1016/j.patter.2020.100116

Clement, J. (2020). *Number of social network users worldwide from 2017 to 2025*. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Di Minin, E., Fink, C., Tenkanen, H., & Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. *Nature Ecology and Evolution*, *2*(3), 406–407. https://doi.org/10.1038/s41559-018-0466-x

Fox, N., August, T., Mancini, F., Parks, K. E., Eigenbrod, F., Bullock, J. M., Sutter, L., & Graham, L. J. (2020). 'photosearcher' package in R: An accessible and reproducible method for harvesting large datasets from Flickr. *SoftwareX*. https://doi.org/10.1016/j.softx.2020.100624

Gaston, K. J., Aimé, E., Chan, K. M. A., Fish, R., Hails, R. S., & Maller, C. (2019). People and Nature – A journal of relational thinking. *People and Nature*. https://doi.org/10.1002/pan3.7

Ghermandi, A., & Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, *55*, 36–47. https://doi.org/10.1016/j.gloenvcha.2019.02.003

Graham, L. J., & Eigenbrod, F. (2019). Scale dependency in drivers of outdoor recreation in England. *People and Nature*, *1*(3), 406–416. https://doi.org/10.1002/pan3.10042

Hausmann, A., Toivonen, T., Heikinheimo, V., Tenkanen, H., Slotow, R., & DI Minin, E. (2017). Social media reveal that charismatic species are not the main attractor of ecotourists to sub-Saharan protected areas. *Scientific Reports*. https://doi.org/10.1038/s41598-017-00858-6

Johnson, T. F. (2021a). *addeR*. Retrieved from https://github.com/GitTFJ/addeR

Johnson, T. F. (2021b). GitTFJ/classecol (version 0.4.0). *Zenodo*. Retrieved from https://zenodo.org/record/4569449#.YDyz49zLdEY

Kareiva, P., Tallis, H., Rickets, T. H., Daily, G. C., & Polasky, S. (2011). *Natural capital: Theory and Practice of Mapping Ecosystem Services* (1st ed.). Oxford University Press.

Lennox, R. J., Veríssimo, D., Twardek, W. M., Davis, C. R., & Jarić, I. (2020). Sentiment analysis as a measure of conservation culture in scientific literature. *Conservation Biology*, *34*(2), 462–471. https://doi.org/10.1111/cobi.13404

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions* (2nd ed.). Cambridge University Press.

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in Tweets. *ACM Transactions on Internet Technology*, *17*(3), 1–23. https://doi.org/10.1145/3003433

Society for Conservation Biology working groups. (2020). *Conservation culturomics*. Retrieved from https://conbio.org/groups/working-groups/conservation-culturomics

Srivastava, A. N., & Sahami, M. (2009). *Text mining classification, clustering, and applications* (1st ed.). CRC Press.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, *233*, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Van Zanten, B. T., Van Berkel, D. B., Meentemeyer, R. K., Smith, J. W., Tieskens, K. F., & Verburg, P. H. (2016). Continental-scale quantification of landscape values using social media data. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(46), 12974–12979. https://doi.org/10.1073/pnas.1614158113

Zhang, E., & Zhang, Y. (2009). F-measure. *Encyclopedia of Database Systems*. https://doi.org/10.1007/978-0-387-39940-9_483

SUPPORTING INFORMATION

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

How to cite this article: Johnson TF, Kent H, Hill BM, et al. classecol: Classifiers to understand public opinions of nature. *Methods Ecol Evol*. 2021;00:1–6. https://doi.org/10.1111/2041-210X.13596