Sentiment Analysis of Conservation Studies Captures Successes of Species Reintroductions

Kyle S. Van Houtan  
*Monterey Bay Aquarium; Duke University*

Tyler Gagne  
*Monterey Bay Aquarium*

Clinton N. Jenkins  
*IPÊ - Instituto de Pesquisas Ecológicas, clinton.jenkins@fiu.edu*

Lucas Joppa  
*Microsoft*

Follow this and additional works at: [https://digitalcommons.fiu.edu/earth_environment_fac](https://digitalcommons.fiu.edu/earth_environment_fac)  
Part of the [Earth Sciences Commons, and the Environmental Sciences Commons](https://digitalcommons.fiu.edu/)

**Recommended Citation**  
Van Houtan, Kyle S.; Gagne, Tyler; Jenkins, Clinton N.; and Joppa, Lucas, "Sentiment Analysis of Conservation Studies Captures Successes of Species Reintroductions" (2020). *Department of Earth and Environment*. 93.  
https://digitalcommons.fiu.edu/earth_environment_fac/93

This work is brought to you for free and open access by the College of Arts, Sciences & Education at FIU Digital Commons. It has been accepted for inclusion in Department of Earth and Environment by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.
Sentiment Analysis of Conservation Studies Captures Successes of Species Reintroductions

Graphical Abstract

Highlights
- The volume of science is rapidly growing, challenging learning within the community
- Machine learning can automate the process of extracting data from human language
- We applied five natural language processing models to over 4,300 scientific abstracts
- The approach captured broad trends and known lessons, offering future promise

Authors
Kyle S. Van Houtan, Tyler Gagne, Clinton N. Jenkins, Lucas Joppa

Correspondence
kvanhoutan@mbayaq.org

In Brief
Science and technology are increasingly integrated into our everyday lives. A key aspect of science is that the community learns through verified, published findings. Online archives and publications have vastly increased the volume of published science, affording greater access to research results while also presenting new challenges. This study uses established methods in artificial intelligence to assess whether reading scientific papers can be automated. The results are promising, although technical disciplines with specific vocabulary will require special considerations.
Article

Sentiment Analysis of Conservation Studies Captures Successes of Species Reintroductions

Kyle S. Van Houtan,1,2,5,* Tyler Gagne,1 Clinton N. Jenkins,3 and Lucas Joppa4
1Monterey Bay Aquarium, 886 Cannery Row, Monterey, CA 93940, USA
2Nicholas School of the Environment, Duke University, Box 90328, Durham, NC 27708, USA
3IPÊ - Instituto de Pesquisas Ecológicas, Rod. Dom Pedro I, km 47, Caixa Postal 47, Nazaré Paulista, São Paulo 12960-000, Brazil
4Microsoft, One Microsoft Way, Redmond, WA 98075, USA
5Lead Contact
*Correspondence: kvanhoutan@mbayaq.org
https://doi.org/10.1016/j.patter.2020.100005

SUMMARY

Learning from the rapidly growing body of scientific articles is constrained by human bandwidth. Existing methods in machine learning have been developed to extract knowledge from human language and may automate this process. Here, we apply sentiment analysis, a type of natural language processing, to facilitate a literature review in reintroduction biology. We analyzed 1,030,558 words from 4,313 scientific abstracts published over four decades using four previously trained lexicon-based models and one recursive neural tensor network model. We find frequently used terms share both a general and a domain-specific value, with either positive (success, protect, growth) or negative (threaten, loss, risk) sentiment. Sentiment trends suggest that reintroduction studies have become less variable and increasingly successful over time and seem to capture known successes and challenges for conservation biology. This approach offers promise for rapidly extracting explicit and latent information from a large corpus of scientific texts.

INTRODUCTION

The sheer volume of scientific literature challenges the goal of capturing knowledge from the published body of peer-reviewed science. A recent review1 of species reintroductions, for example, manually extracted information from 361 published articles. While this was admittedly a small fraction of the total literature on the topic, it still required months of effort from highly trained experts just to obtain the raw data, which they then had to analyze. However, because population reintroductions are an effective means to accomplish an elusive task—to recover species and restore ecosystems2–4—understanding what determines their success or failure is considered broadly important (S.L. Becker, T.E. Nicholson, K.A. Mayer, M.J. Murray, and K.S.V.H., unpublished data).5 Therefore, such narrated lessons from established evidence are critical for conservation practices and management decisions.1,6 How can we lower the barrier to learning them?

Natural language processing (NLP) is a branch of artificial intelligence, or machine learning, which analyzes strings of human language to extract usable information. One goal of NLP is to automate the processing of large volumes of text with minimal human supervision7,8 yet crucially in a manner that approximates the performance of a human reader. As applied here, sentiment analysis (SA) parses different affective states of sentiment to capture either single or combined emotions, attitudes, or traits.9,10 While an array of methods exists, the basic principle of
SA is to use a trained set of text that has previously been attributed with a sentiment score to define the sentiment for a separate body of unlabeled text. Although SA has been developed to describe more complex and sophisticated human emotions (e.g., empathy, greed, trust, and fear), such results are, perhaps expectedly, variable. By comparison, a more simplistic score of negative to positive sentiment, in the form of a weighted polarity, is far more robust in capturing basic attitudes across various types of texts and fields of study. Aside from labeling raw texts, the lexicon-based and machine learning-based SA models have additional value as they are resilient to various structures of text strings (e.g., letter case, punctuation, and stop words) and require little to no text preprocessing. Although NLP is a dynamic and rapidly advancing research area with extensive scientific and commercial applications, and there are more sophisticated approaches to NLP than SA, the SA approach we deploy here can produce straightforward and robust results with broad and intuitive interpretative value.

In this study, we explore the use of supervised SA to facilitate a new meta-analysis of the species reintroduction literature with a goal of understanding effectiveness and identifying what determines success. We query public databases to build a robust corpus, numbering in the thousands of scientific abstracts and use existing or “off-the-shelf” lexicons and NLP models to identify the terms that drive sentiment and the domain concepts associated with success. This basic yet novel application shows the potential to enable a more rapid understanding of the growing volume of scientific literature. Such research on research, or meta-analyses, can produce results important for research journal practices as well as the topical domain itself.

RESULTS AND DISCUSSION

Since we are applying language models created for general use to extract information specifically from scientific abstracts, it is important to evaluate how the terms contained in our abstracts correlate with sentiment scores. Figure 1A shows the contribution of frequent terms from our reintroduction abstracts to sentiment scores using the AFINN lexicon. The most common terms associated with positive sentiment are success, protect, growth, support, greater, recommend, strong, allow, great, help, benefit, reach, best, better, fe, clear, highlight, conflict, poor, kill, trap, avoid, prevent, lack, problem, restrict, threat, risk, loss, threaten. As the term success drives sentiment more than any other term (either positively or negatively), and as this list seems to capture terms that genuinely reflect how authors communicate successes and failures of population metrics as well as reintroduction programs, the abstract-level sentiment score serves as a rational proxy to capture the lessons learned from the studies.

Figure S5 and S6 provide both sentence- and abstract-level sentiment scores from a range of abstracts highlighting a range of polarity values across the collected corpus.

Figure 1B summarizes the annual multi-model ensemble of sentiment of reintroduction abstracts over three decades. Though there is some variation between the five models (see Results and Discussion in Figure S4), the models converge on a general trend of decreasing variation and increasingly positive sentiment through time. To this point, uniquely over the study period, the ensemble, including the confidence interval, is positive from 2007 to 2016, encapsulating the final 10 years of the study. Taken with Figure 1A, this suggests that reintroduction science is achieving positive results and becoming more successful over the last 30 years. The positive framing and success associated with specific settings, however, is of interest. Figure 2A shows the frequency and sentiment trends of ten issues
in reintroductions, ranked by their final modeled sentiment score. Abstracts containing terms that reference methods known to be effective, such as adaptive management and population viability analysis,2,3,22 have increasing sentiment throughout the study period. More recently deployed yet successful methods, such as Bayesian statistics and SNPs (single-nucleotide polymorphisms), also unsurprisingly demonstrate positive sentiment.23,24 Conversely, reintroduction studies dealing with fragmented populations25,26 and invasive species2,27,28 have negative sentiment, indicating that barriers to success are likely impactful.

Figures 2B and 2C repeat this approach but for ecosystem and taxa contexts. Although many studies address rat depredation and island environments, neither set shows a clear signal of success, likely owing to the many conservation challenges associated with both.2,23 Likewise, the management of small populations (such as with condors and tortoises) are challenging29–32 while other heavily studied species with improving conservation status (such as giant pandas and river otters)30,33,34 have increasing sentiment. Although there is variability between years, reintroduction studies in scrub, grassland, and savanna habitat, in the ocean and in coral reefs, and in forests reflect a positive sentiment that is either stable or increasing through the study. By contrast, studies involving zoo collections have persistently negative sentiment. This result may simply reflect that zoo-based captive breeding programs are often supporting species that are critically endangered in the wild7,35,36 and face extreme challenges. Lastly, commonly applied techniques such as telemetry27,28 and translocation may not depart from the overall trend simply because they are so widespread, and thus their trend dominates the overall corpus.

Conclusions

While text-mining algorithms are broadly applied in a range of powerful applications,7,8 they have yet to find a regular use in conservation science. Here, we developed a novel use of NLP by using SA to review and extract knowledge from a large body of scientific literature on population reintroductions. Our approach provided several lessons and recommendations. To begin, although the underlying models are trained on words and sentences from a general domain,20,39,40 our application here captures the specific terms (Figure 1A) as well as the broad success and failures associated with specific management settings (Figure 2) for conservation biology more generally and reintroductions more specifically.

Next, this application is a modest proof of concept toward automating meta-analyses of the scientific literature. Future analyses will achieve greater success by using lexicons and other NLP models that are trained for the specific domain, that compare word- versus sentence-level methods of scoring, and when possible are benchmarked and validated against specific empirical metrics of interest (see Supplemental Information). Such investments may be particularly important for recursive neural tensor network (RNN) models that incorporate syntax on top of word significance, as syntax (more so than individual words) may transfer less easily from a general to an applied domain (see Figure S4). Therefore, while the significance of individual words (see Figure 1A) may retain significance across applications and word-based lexicons may find transdisciplinary applications apt, RNN models may more require domain-based treatments.

These improvements need to be weighed against the logistical cost of model training and validation, but the proposed value of automation and generation of latent information and novel
syntheses is well known from other scientific disciplines. For reintroduction biology, for example, such approaches may pinpoint whether polarity from SA models correlates with quantitative data on reintroduction successes reported within the studies themselves, or more reflects the overall conservation status of the population in question. However, popular metrics within information theory, such as TF*IDF scores, though useful in locating clusters of uncommon terms and key phrases, may have limitations for the purposes of training domain-specific lexicons for SA (see Figure S3).

Lastly, aside from innovating the analytical methods, greater access and data sharing from publishing groups and popular indexing services will provide significant advances. Here, we manually accessed and extracted abstracts from a single repository of published literature. Although this presented a substantial body of information, future analyses will be improved by the development of public application programming interfaces and automated and open access to entire journal articles, not simply the condensed abstracts. Such advances will facilitate the intended greater good of science by making more scientific articles, and more of each article, freely available to the public, which may advance future analyses through facilitating increased meta-analyses.

EXPERIMENTAL PROCEDURES

We developed our corpus by searching the Web of Science indexing service, a representative database of the peer-reviewed scientific literature. We used the Boolean operators AND, OR to retrieve studies containing all of the terms species, conservation, population, and either reintroduction or translocation in their abstracts. As translocation also indicates chromosomal transfer, we used the NOT operator to remove studies with the terms protein, yeast, Arabidopsis, Drosophila, Saccharomyces, and Escherichia most associated with the non-target term use and thus generating false positives. We removed studies lacking abstracts, and excluded studies without English abstracts. This resulted in a corpus of 4,313 studies (Figure S1) published from 1987 to 2016 with searchable abstracts containing 1,030,558 words.

From this body of text, we estimated abstract-level sentiment with an ensemble model. We built sentiment scores from four lexicon-based models (AFINN, Bing, NRC, and Syuzhet) and one trained RNN model, the Stanford coreNLP. The lexicons classify sentiment in text strings from the accumulation of sentiment scores of individual words, which each lexicon labels via some form of crowdsourcing. The RNN model is also derived from labeled training text (in this case over 10,000 scored film reviews) and models sentiment not of each individual word but rather by considering the syntax at the sentence level. We used the R packages Syuzhet, coreNLP, and NLP to generate a polar score (negative to positive) of mean sentiment for each abstract, we derived a multi-model ensemble average. We employ this approach, as NLP and climate studies show that such ensembles are simply the condensed abstracts. Such advances will facilitate the intended greater good of science by making more scientific articles, and more of each article, freely available to the public, which may advance future analyses through facilitating increased meta-analyses.

E thics Statement

Our study did not require human or animal subjects.

DATA AND CODE AVAILABILITY

All the raw data used in this project are available at the Web of Science in a third-party project repository (https://osf.io/l4dc7/), at Github (https://bit.ly/2S6dahD), or included in Supplemental Information.

Supplemental Information

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2020.100005.

ACKNOWLEDGMENTS

J. Alt, S. Pimm, M. Cunha, J. Moxley, and two anonymous reviewers made improvements to earlier versions of the manuscript. This study was made possible from generous contributions to the Monterey Bay Aquarium.

AUTHOR CONTRIBUTIONS

All authors conceived of the study and reviewed the manuscript. K.V. gathered the data, T.G. developed and ran the models, and both generated the figures. K.V. wrote the manuscript with contributions from T.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 21, 2019
Revised: December 16, 2019
Accepted: January 14, 2020
Published: March 19, 2020

REFERENCES

1. Taylor, G., Canessa, S., Clarke, R.H., Ingwersen, D., Armstrong, D.P., Seddon, P.J., and Ewen, J.G. (2017). Is reintroduction biology an effective applied science? Trends Ecol. Evol. 32, 873–880.
2. Groom, M.J., Meffe, G.K., and Carroll, R. (2006). Principles of Conservation Biology (Sinauer Associates).
3. Sarrazin, F., and Barbault, R. (1996). Reintroduction: challenges and lessons for basic ecology. Trends Ecol. Evol. 11, 474–478.
4. Mayer, K.A., Tinker, M.T., Nicholson, T.E., and Murray, M.J. (2019). Surrogate rearing a keystone species to enhance population and ecosystem restoration. Oryx 53, https://doi.org/10.1017/S0030605319000346.
5. Van Houtan, K.S., Halley, J.M., van Aarde, R.J., and Pimm, S.L. (2009). Achieving success with small, translocated mammal populations. Conserv Lett. 2, 254–262.
6. Sutherland, W.J., Pullin, A.S., Dolman, P.M., and Knight, T.M. (2004). The need for evidence-based conservation. Trends Ecol. Evol. 19, 305–308.
7. Aggarwal, C.C., and Zhai, C. (2012). Mining Text Data (Springer Science & Business Media).
8. Feldman, R., and Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data (Cambridge University Press).
