Text classification to streamline online wildlife trade analyses

Oliver C. Stringham[1,2*], Stephanie Moncayo[1], Katherine G.W. Hill[1], Adam Toomes[1], Lewis Mitchell[2], Joshua V. Ross[2], Phillip Cassey[1]

1. Invasion Science & Wildlife Ecology Lab, University of Adelaide, SA 5005, Australia
2. School of Mathematical Sciences, University of Adelaide, SA 5005, Australia

*Corresponding Author: Oliver C. Stringham, The University of Adelaide, North Terrace Campus, Adelaide, SA, 5000, Australia. oliverstringham@gmail.com

Running headline: Text classification and online wildlife trade

Keywords: data cleaning, internet, natural language processing, text mining, wildlife trade

Abstract

1. Automated monitoring of websites that trade wildlife is increasingly necessary to inform conservation and biosecurity efforts. However, e-commerce and wildlife trading websites can contain a vast number of advertisements, an unknown proportion of which may be irrelevant to researchers and practitioners. Given that many of these advertisements have an unstructured text format, automated identification of relevant listings has not traditionally been possible, nor attempted. Other scientific disciplines have solved similar problems using machine learning and natural language processing models, such as text classifiers.
2. Here, we test the ability of a suite of text classifiers to extract relevant advertisements from an Australian classifieds website where people can post advertisements of their pet birds (n = 16.5k advertisements). Furthermore, in an attempt to answer the question 'how much data is required to have an adequately performing model?', we conducted a sensitivity analysis by simulating decreases in sample sizes to measure the subsequent change in model performance.
3. We found that text classifiers can predict, with a high degree of accuracy, which listings are relevant (ROC AUC $\geq$ 0.98, F1 score $\geq$ 0.77). From our sensitivity analysis, we found that text classifiers required a minimum sample size of 33% (c. 5.5k listings) to accurately identify relevant listings (for our dataset), providing a reference point for future applications of this sort.
4. Our results suggest that text classification is a viable tool that can be applied to the online trade of wildlife to reduce time dedicated to data cleaning. However, the success of text classifiers will vary depending on the advertisements and websites, and will therefore be context dependent. Further work to integrate other machine learning tools, such as image classification, may provide better predictive abilities in the context of streamlining data processing for wildlife trade related online data.

Introduction

The global wildlife trade is a major concern for biodiversity conservation and biosecurity enforcement (Smith et al. 2009). Information on the composition and volume of species, and where they are traded, is highly valuable for informing conservation research and practice (Scheffers et al. 2019). The Internet is an emerging source of data on the wildlife trade (Siriwat and Nijman 2020; Jarić et al. 2020). Researchers, NGOs, and government agencies monitor websites that trade wildlife to quantify various aspects of the trade (e.g., Sung & Fong 2018). Data gathered from the Internet are typically not immediately ready for analysis (i.e., they are 'messy') and must be cleaned or processed to identify the desired attributes for subsequent analysis (Dobson et al. 2020). This is especially true for classifieds, forums, and social media sites where human users type their advertisements into an open (or 'free form') text box. Consequently, relevant attributes cannot be extracted automatically (i.e., through web scraping or computer-based data manipulation) due to non-uniformity across users' advertisements (different species names, abbreviations, misspelling, etc.) (Stringham et al. 2020). Likewise, depending on the website, many online listings (i.e., posts) may contain items or taxa that are irrelevant for a given research context. For instance, in a pet reptile forum, one can find users trading tanks, food, or other accessories, which may not be relevant to researchers exploring the trade of live reptiles (e.g., Stringham and Lockwood 2018). The most common method to extract online wildlife trade data is to manually inspect each listing and record the desired attributes. Depending on how many listings are collected, the data cleaning process could represent an enormous amount of time and effort for researchers. Wildlife-related web data is notorious for its scale: for example, Xu et al. (2018) tracked around 140k tweets from a two-week period relevant to ivory and pangolin trade.

Automated methods of data cleaning such as machine-learning techniques and Natural Language Processing (NLP) tools have potential to streamline the processing of wildlife trade data derived from the Internet (Di Minin et al. 2019). A useful but unexplored application is to predict and extract relevant online listings based on their text, which could save time in manual data processing steps if many irrelevant listings exist in the dataset. In particular, text classification models relate the words associated with a particular label, such as 'relevant' or 'irrelevant', to predict the label of an unknown data point. A well-known application of text classifiers is filtering spam emails (Guzella and Caminhas 2009). In this context, a text classification model uses a training dataset of labelled emails (span or not spam) and trains a model to predict those labels based on their constituent words. The resulting model labels new incoming emails as spam or not. In the context of wildlife-trade data derived from the Internet, text classification models have the potential to identify relevant listings and remove irrelevant listings that do not sell wildlife (i.e., fish tanks, bird cages, food) by using the words in the listings. If shown to be effective, this could save researchers substantial time in the data cleaning process.

Here, we examine if text classification models can predict which Internet listings are relevant to wildlife trade research (for our own specific research purposes; e.g., Toomes et al. 2020). Further, to assist future implementation of such models, we sought to identify how much data is needed for a text-classification model to perform adequately well. We collected advertisement listings from a popular Australian classifieds website where people trade their pet birds and accessories (e.g., bird cages or bird toys). Bird trade is largely unregulated in Australia (but see Woolnough et al. 2020) and is highly diverse with a large number of both native and alien species; with potential conservation and biosecurity consequences (Vall-llosera & Cassey 2017). We observed three major categories of advertisements that were irrelevant to our research objectives: (i) 'junk' listings (not trading birds); (ii) wanted advertisements (requesting a bird); and (iii) the sale of domestic poultry – e.g., gamebirds, waterfowl, and pigeons (non-target wildlife taxa). We manually labelled around 16.5k listings and tested the efficacy of three commonly used text classification models at determining which listings were relevant versus irrelevant. Next, we systematically removed records from our dataset and recorded the change in model performance. Our results imply that text classification can be an incredibly useful time-saver when cleaning data on the wildlife trade, which is structurally (textually) similar to the data we explore here.

Materials and Methods

Data collection and data curation

We collected data from a popular Australian classifieds website daily over the course of five months (5 July 2019 to 5 December 2019). All information collected from the website was publicly available. We received ethics approval from The University of Adelaide (ethics number: H-2020-184) to collect this data and have anonymized the name of the website as good ethical practice (Hinsely et al. 2016). On the website, people can post advertisements (i.e., listings) of items/animals they are trading. From each listing, we collected: (i) the title; (ii) text description; (iii) date; and, (iv) images (if provided). The title and text description fields are open text boxes where the user can type whatever they desire up to a character limit. We collected a total of 66,704 unique listings. Given the large number of unique listings collected, and the substantial resources required to manually clean the data, we labelled a random subset of around 25% of the listings (n = 16,509). This took approximately 103 hours to label (at an average rate of 161 listings per hour). Four different authors were labelers (SM, KH, AT, OS), and we did not overlap labelling, although this is preferred practice (e.g. see Sheng et al. 2008).

For each listing, we manually labelled the taxa (e.g., species) being traded based on the title, the text, and the pictures provided in the listing. Some listings contained more than one species being traded. We identified the listing to the most specific taxonomic rank as possible (species or subspecies), but occasionally not enough information was provided and the listing was identified to genus, family, order, or class (i.e., bird). We resolved taxa names and obtained upper-level taxonomy using the Global Biodiversity Information Facility database (GBIF 2020). For each listing, we recorded if the user was requesting a bird species (i.e., a wanted advertisement), except in the case of domestic poultry species (see below). We labelled listings not trading a live bird as 'junk' (i.e., bird accessory such as cage or bird food).

Preparing text for text classification models

We considered all text written by the user (title and text description) for our analyses. To prepare or 'clean' the text for the NLP text classification models, we followed standard NLP text cleaning procedures (Silge and Robinson 2017) and removed special characters (emojis, dollar signs, numbers, etc.), removed all punctuation, converted text to lowercase, and removed all numbers. Next, we removed all stop words found in the following lexicons: *SMART*, *snowball*, and *onix*. We did not remove the stop words: "want", "wants", "wanting", or "wanted", so we could distinguish wanted advertisements. We stemmed each word using the Snowball stemmer. For the text classifier models, we tokenized the text to be unigrams (i.e., one word) and did not consider further n-grams. Text cleaning was performed in the statistical software R version (R Core Team 2020) using the following packages: stringr (Wickham 2019), dplyr (Wickham et al. 2020), tidytext (Silge & Robinson 2016), and corpus (Perry 2020).

To test the classification of irrelevant listings (see '*Text classification models*' below), we applied three separate labels for each listing. The first label was for 'junk' listings, where a live bird was not being traded (e.g., bird cage). The second label was for 'wanted' listings where a user was requesting a bird species and not selling one. The final label was for taxa that we considered non-target for our purposes (i.e., farm, poultry, or domesticated species). We called this label 'domestic poultry' and applied it to listings that were selling birds in the taxonomic orders of Anseriformes (waterfowl) and Galliformes (gamebirds) or trading domestic pigeons (*Columba livia domestica*). For text classification models, we removed listings categorized as more than one label (i.e., 'domestic poultry' and 'wanted'). Further, for the 'wanted' label, we removed listings if eggs were being advertised, as we did not simultaneously record if egg advertisements were also labeled as 'wanted'. This resulted in a sample size (number of listings used for text classification models) of 16,475 for 'domestic poultry', 16,446 for 'junk', and 13,751 for 'wanted'.

Text classification models

To classify irrelevant listings, we used three common supervised text classifiers: Logistic Regression, Multinomial Naive Bayes, and Random Forest. At a basic level, each classifier considers each word (i.e., gram) and their frequency as a covariate (i.e., 'feature') (Bird et al. 2009). However, each classifier varies in the algorithm used to classify observed listings as relevant or not (Bird et al. 2009). For each classifier, the order of the words in the listing was unaccounted, thus earning the name 'bag of words' classifier. We ran each model for each of the three labels mentioned above. We used 10-fold cross validation to train the model and evaluate predictions. We used the cross-validated macro-average of the following metrics to evaluate the performance of each model: receiver operating characteristic (ROC) curve and its area under the curve (ROC AUC), precision-recall curve and its area under the curve (PR ROC), precision, recall, negative predictive value (NPV), specificity, and F1 score (see Appendix S1 for more information evaluation metrics). We extracted the top features (e.g., covariates) for each model. Text classification models were performed in Python using the sci-kit learn library (Pedregosa et al. 2011), while plotting was conducted in R using ggplot2 (Wickham 2016).

Sensitivity analysis: degradation of model performance with diminishing sample size

To test the sensitivity of model performance to changes in sample size, we implemented the text classification model with iteratively smaller sample sizes. We systematically decreased the sample size of the training set by 500 records at a time, removing at most 15k records (c. 91% of entire dataset). We repeated this for each label and used 10-fold cross validation. To account for the variability in model performance due to cross-validation, we repeated the text classification model for 100 iterations, for each sample size explored. We recorded 10-fold cross validation statistics across each fold and model iteration (1,000 values in total for each sample size). For this sensitivity analysis, we only considered the logistic regression classifier and used the F1 value to evaluate model performance. We recorded the maximum training set sample size at which the F1 score was 99% of its maximum value (i.e., the F1 score without reducing sample size).

Results

We manually categorized 16,509 listings, of which 15.0% (n=2,473) were labeled as 'junk', 21.9% (n=3,615) were labeled as 'domestic poultry', 4.8% (n=787) were labelled as 'wanted' advertisements, and the remaining (c. 58%) were 'for sale' advertisements of relevant bird taxa.

The text classifiers performed extremely well for the 'domestic poultry' label (Figure 1; Appendix S2), with a cross-validated average ROC AUC of >0.99, Precision-Recall AUC of ≥0.97, and F1 score of >0.95 for all text classifiers (Figures 1-3). The text classifiers for the 'junk' label also performed very well, with marginally lower metric values compared to 'domestic poultry' (Figure 1). Further, all other metrics evaluated suggested that the text classification models performed very well for these two labels (Figures 1-3; Appendix S2; see Appendix S3 for confusion matrices). The text classification models for the 'wanted' advertisement label performed less well, however, the Logistic Regression and Random Forest classifiers for this label performed moderately well and each was much better than chance with a ROC AUC > 0.98, Precision-Recall AUC > 0.88, and F1 score > 0.77. Overall, the 'wanted' classifiers were not as good at predicting positive outcomes (e.g., if a listing is 'wanted'), yet did not struggle with predicting negative outcomes (Specificity = 0.99, and Negative Predictive Value = 0.99 for Logistic Regression classifier). In terms of relative performance between the classifiers, the Logistic Regression and Random Forest classifiers slightly outperformed the Naive Bayes Classifier; however, overall, their performances were comparable (Figure 1-3).

The top features for each label aligned with what should be expected and were similar across all text classifiers (Figure 4). For the 'junk' label, grams such as "condit" (i.e., condition), "cage", "birdcag" (i.e., birdcage) were the top features. For the 'domestic poultry' label, grams such as "pigeon", "rooster", and "chicken" were the top features. Finally, for the 'wanted' label, grams such as "want", "buy", "wtb" (an acronym for 'want to buy'), and "unwant" (i.e. "unwanted") were the top features.

As we reduced the sample size of the training set, we observed a non-linear decrease in model performance, where the F1 score initially declined gradually and then at an increasing rate at lower sample sizes (Figure 5). There were differences in this decline in performance among labels. The classifier for the 'domestic poultry' label realized 99% of the full model F1 score at c. 4.8k records (29% of dataset). For the 'junk' label, this was c. 9.3k records and c. 6.3k records for the 'wanted' label (57% and 45%, respectively). Stated another way, for the 'domestic poultry' label, the addition of c. 11k labelled records from our manual data labelling only increased the model F1 score by 0.01. For 'junk' and 'wanted', this value was c. 7.1k listings and c. 7.5k listings, respectively.

Discussion

Text classification can be a highly accurate method to extract relevant listings of wildlife found on the Internet. In particular, for listings trading non-target taxa and listings trading bird accessories (e.g., bird cages), text classification models were able to classify these listings with a very high degree of accuracy. Although the performance of the model varied between labels, our results suggest that this technique can be used to substantially lower the number of wildlife listings needed to be manually inspected, thus saving considerable time and resources. Further, we provide clarity around the question of 'how much data is needed to guarantee an adequately performing model?'. Of the more than 16k listings we manually labelled, our results suggest that, at most, only 9k listings were needed, although this number varied by label.

Text classification models are commonplace in other disciplines and industries, which work heavily with text data (e.g., Guzella & Caminhas 2009), yet have not been applied to data collected on the wildlife traded occurring on the Internet. Importantly, from our dataset, around 60% of the listings were relevant (for our purposes), representing a substantial amount of time and effort that would otherwise be spent on manually removing irrelevant data. For the website we explored, we showed that text classifiers predicted with great accuracy the advertisements that were not selling wildlife or selling non-target wildlife. In particular, text classification models performed the best for identifying listings trading non-target taxa (e.g., farm and domestic bird species). This is promising as a time saving tool because sometimes the most commonly traded taxa are the ones of least interest to researchers (i.e., pigeons and chickens; in our example). In contrast, the text classifiers had more difficulty distinguishing 'wanted' advertisements (where a user was requesting a bird) yet was still much better than chance. This suggests that the words people used in wanted advertisements have some overlap with those words used in non-wanted advertisements (e.g., names of species), and thus yields lower predictive abilities.

Importantly, we demonstrate that the model performance will likely not improve with more data because we observed a plateau of model performance after around 6k listings for wanted advertisements (45% of sample size). Therefore, even if we manually labelled many more listings, the model performance is unlikely to increase. This highlights an important point that model performance is a function of the underlying data itself (i.e., text) and not of the lack of data (once an adequate sample size is achieved).

How much data is required for an adequately performing text classification model? Our results show that this number will vary by what is being classified. For this study, we cleaned a substantial number of listings (c. 16k) yet found that model performance marginally increased after 5k to 9k records (31% to 56% of total effort). Thus, for other researchers who may not have the resources to invest this much effort, or are looking for a more efficient way to curate messy online data, our results provide guidance on how much data is needed before text classification can be used. We recommend establishing computer code to test the model performance and then repeatedly check the model performance at regular intervals (e.g., every 1k records cleaned). Ultimately, the labelled dataset will need to encapsulate the variation of words (i.e., vocabulary) used for a particular label for the text classifier to perform well. For instance, for the 'junk' label, the model performance plateaued at around 5k more records than it did for other labels. We hypothesize the words that Internet users write for the listings that fall under the 'junk' label has more variation (i.e., more words) and thus, we needed a larger sample size of labelled listings to account for that variation.

An important limitation of text classification (and other machine learning tools) is that they are highly context dependent (Lambda et al. 2019). Our specific classifiers were developed based on the text of birds being traded online in Australia and will likely be less useful for birds being traded in other countries and almost entirely useless if looking at other taxa (e.g., fish or plants) or in another language. The reason for this lack of generalization is because words used, and their frequency, will vary under different contexts. For instance, when looking at the trade of aquarium fish, a common irrelevant advertisement may be the sale of a fish tank, something that is not found when trading birds. We recommend that researchers consider each context separately when using these tools. Since manual data processing is likely always required to analyze the data, these tools can be tested throughout the cleaning stage to see if applicable.

Besides extracting relevant advertisements, text classifiers have the potential to identify the species being traded in online advertisements. Our results suggest that this will be possible for commonly traded taxa, with large amounts of data. For instance, in our study, advertisements for a group of species (waterfowl, gamebirds, and pigeons) comprised around 3.6k listings (22% of dataset) and were highly distinguishable using the text classifiers. The same kinds of models can be used to identify individual species of interest; however, text classifiers (like all machine learning techniques) require a large volume of data to perform well (Di Minin et al. 2019). In many cases, individual species of interest may not have enough advertisements to build adequate text classifiers. Thus, alternative methods such as matching species names (scientific, common, or trade names) to the text of advertisements using a fuzzy string-matching model (e.g., Levenshtein distance) may yield better results. In fact, if consistent patterns are used by users (e.g., the same species name is used by many users), string matching may yield just as good or better results than text classifiers. While our study relied exclusively on the text of the advertisement, there are other attributes of an Internet listing that can be considered for automated cleaning. For instance, a related study used metadata attributes of online listings (e.g., the number of views and the price) to classify illegal sales of elephant ivory (Hernandez-Castro & Roberts 2015). In cases with no or limited text provided (e.g., only a photo is posted), machine learning techniques such as image classification could assist in the classification of species or the product traded (Norouzzadeh et al. 2018). Integrating text classification with the aforementioned models may improve predictive ability, and we recommend this as a future area of research and development for the wildlife trade related online data.

Given that a substantial proportion of online listings may not be relevant to wildlife trade research (e.g., 40% irrelevant for our dataset), text classification methods can substantially decrease the amount of time spent processing raw data. Here, we demonstrate that text classification can be viable tool to identify irrelevant listings. When considering data on the scale of 'big data' of tens to hundreds of thousands of online advertisements (e.g. Olden et al 2020), text classifiers have the potential to save tens to hundreds of hours of curation effort. We recommend future application of text classifiers and testing other machine learning and natural language processing tools when cleaning messy data collected from the Internet on wildlife trade.

Data Availability

Data and code for text classification are available from the *figshare* repository at https://doi.org/10.6084/m9.figshare.14032742 and from GitHub at https://github.com/ocstringham/text_classification_wildlife_trade/.

Acknowledgements

This research was funded by the Centre for Invasive Species Solutions (Project PO1-I-002).

Author Contributions

OCS, LM, JVR, and PC conceived the ideas and designed the methodology. OCS collected the data. SM, KGWT, and AT cleaned the data. OCS analyzed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Dobson, A. D. M., Milner-Gulland, E. J., Aebischer, N. J., Beale, C. M., Brozovic, R., Coals, P., Critchlow, R., Dancer, A., Greve, M., Hinsley, A., Ibbett, H., Johnston, A., Kuiper, T., Le Comber, S., Mahood, S. P., Moore, J. F., Nilsen, E. B., Pocock, M. J. O., Quinn, A., … Keane, A. (2020). Making Messy Data Work for Conservation. *One Earth*, *2*(5), 455–465. https://doi.org/10.1016/j.oneear.2020.04.012

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. *Expert Systems with Applications*, *36*(7), 10206–10222. https://doi.org/10.1016/j.eswa.2009.02.037

Hernandez-Castro, J., & Roberts, D. L. (2015). Automatic detection of potentially illegal online sales of elephant ivory via data mining. *PeerJ Computer Science*, *1*, e10. https://doi.org/10.7717/peerj-cs.10

Hinsley, A., Lee, T. E., Harrison, J. R., & Roberts, D. L. (2016). Estimating the extent and structure of trade in horticultural orchids via social media. *Conservation Biology*, *30*(5), 1038–1047. https://doi.org/10.1111/cobi.12721

Jarić, I., Correia, R. A., Brook, B. W., Buettel, J. C., Courchamp, F., Di Minin, E., Firth, J. A., Gaston, K. J., Jepson, P., Kalinkat, G., Ladle, R., Soriano-Redondo, A., Souza, A. T., & Roll, U. (2020). iEcology: Harnessing Large Online Resources to Generate Ecological Insights. *Trends in Ecology & Evolution*, *35*(7), 630–639. https://doi.org/10.1016/j.tree.2020.03.003

Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental conservation. *Current Biology*, *29*(19), R977–R982. https://doi.org/10.1016/j.cub.2019.08.016

Minin, E. D., Fink, C., Hiippala, T., & Tenkanen, H. (2019). A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology*, *33*(1), 210–213. https://doi.org/10.1111/cobi.13104

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

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

Olden, J. D., Whattam, E., & Wood, S. A. (2020). Online auction marketplaces as a global pathway for aquatic invasive species. *Hydrobiologia*. https://doi.org/10.1007/s10750-020-04407-7

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Perry P.O. (2020). corpus: Text Corpus Analysis. R package version 0.10.1. https://CRAN.R-project.org/package=corpus

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scheffers, B. R., Oliveira, B. F., Lamb, I., & Edwards, D. P. (2019). Global wildlife trade across the tree of life. *Science*, *366*(6461), 71–76. https://doi.org/10.1126/science.aav5327

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 614–622. https://doi.org/10.1145/1401890.1401965

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach* (1st edition). O'Reilly Media.

Siriwat, P., & Nijman, V. (2020). Wildlife trade shifts from brick-and-mortar markets to virtual marketplaces: A case study of birds of prey trade in Thailand. *Journal of Asia-Pacific Biodiversity*. https://doi.org/10.1016/j.japb.2020.03.012

Smith, K. F., Behrens, M., Schloegel, L. M., Marano, N., Burgiel, S., & Daszak, P. (2009). Reducing the Risks of the Wildlife Trade. *Science*, *324*(5927), 594–595. https://doi.org/10.1126/science.1174460

Stringham, O. C., & Lockwood, J. L. (2018). Pet problems: Biological and economic factors that influence the release of alien reptiles and amphibians by pet owners. *Journal of Applied Ecology*, *55*(6), 2632–2640. https://doi.org/10.1111/1365-2664.13237

Stringham, O. C., Toomes, A., Kanishka, A. M., Mitchell, L., Heinrich, S., Ross, J. V., & Cassey, P. (2020). A guide to using the Internet to monitor and quantify the wildlife trade. *Conservation Biology*. https://doi.org/10.1111/cobi.13675

Sung, Y.-H., & Fong, J. J. (2018). Assessing consumer trends and illegal activity by monitoring the online wildlife trade. *Biological Conservation*, *227*, 219–225. https://doi.org/10.1016/j.biocon.2018.09.025

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

Toomes, A., Stringham, O. C., Mitchell, L., Ross, J. V., & Cassey, P. (2020). Australia's wish list of exotic pets: Biosecurity and conservation implications of desired alien and illegal pet species. *NeoBiota*, *60*, 43–59. https://doi.org/10.3897/neobiota.60.51431

Vall-llosera, M., & Cassey, P. (2017). Leaky doors: Private captivity as a prominent source of bird introductions in Australia. *PLOS ONE*, *12*(2), e0172851. https://doi.org/10.1371/journal.pone.0172851

Wickham H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

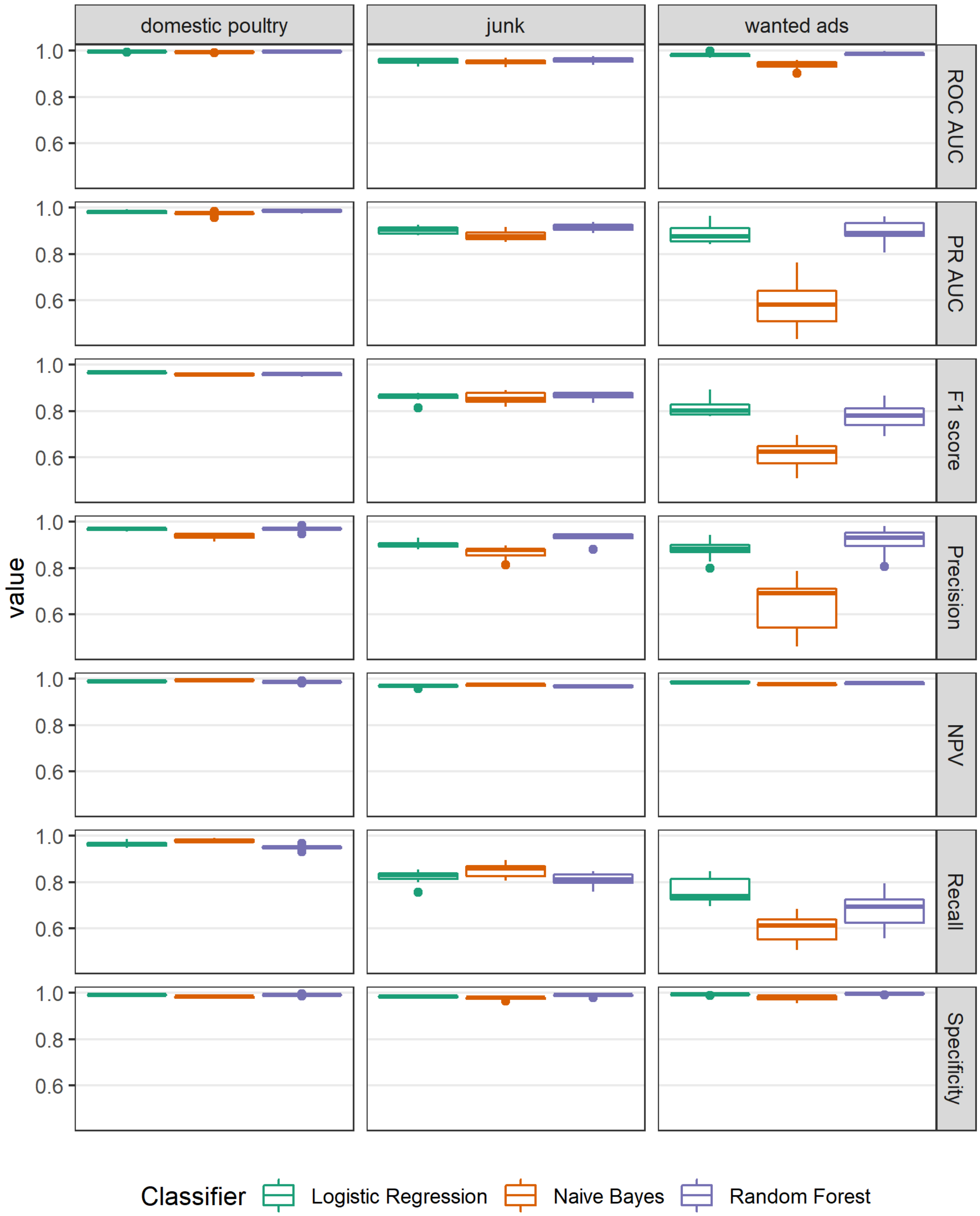
Wickham H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham H., François R., Henry L., Müller K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.0. https://CRAN.R-project.org/package=dplyr

Woolnough, A.P., de Milliano J.M., van Polen Petel T., and Cassey P. (2020). A policy approach to nonindigenous bird management in Victoria – managing potential threats to biodiversity, social amenity and economic values. Victorian Naturalist. 137 (6); 203-209

Xu, Q., Li, J., Cai, M., & Mackey, T. K. (2019). Use of Machine Learning to Detect Wildlife Product Promotion and Sales on Twitter. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00028

Figures



domestic poultry
junk
wanted ads
ROC AUC
PR AUC
F1 score
Precision
NPV
Recall
Specificity
value
1.0
0.8
0.6
Classifier
Logistic Regression
Naive Bayes
Random Forest

Figure 1.

Model evaluation metrics (rows) across 10 cross-validation folds using different text classifiers evaluated for three different labels (columns). See Appendix S1 for more information and calculation of the evaluation metrics and Appendix S2 for exact metric values.

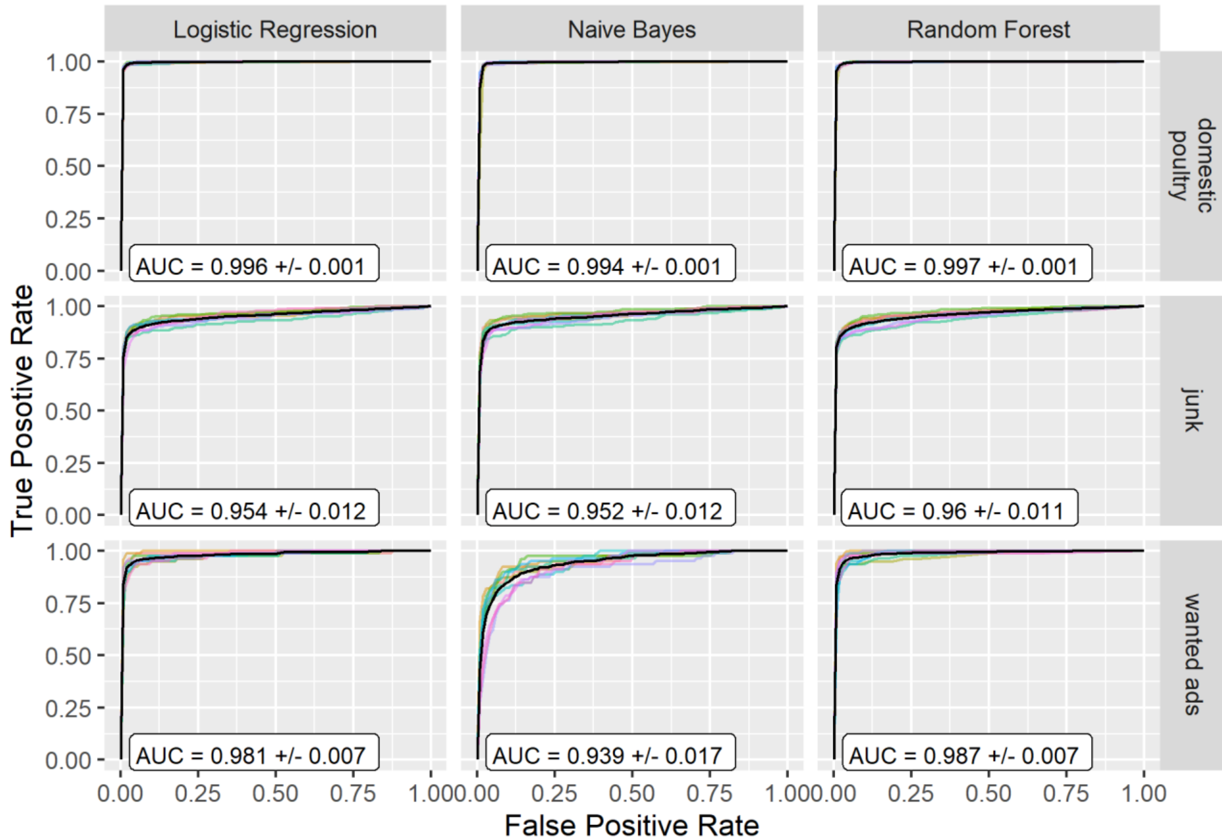


Figure 2.

Receiver operating characteristic curves and the area under the curve (ROC AUC). Three different text classifiers (columns) were tested across three different labels (rows). For each panel, each line represents one cross-validation fold and the solid black line represents the average across all cross-validation folds. Average AUC (area under curve) values are reported with standard deviation.

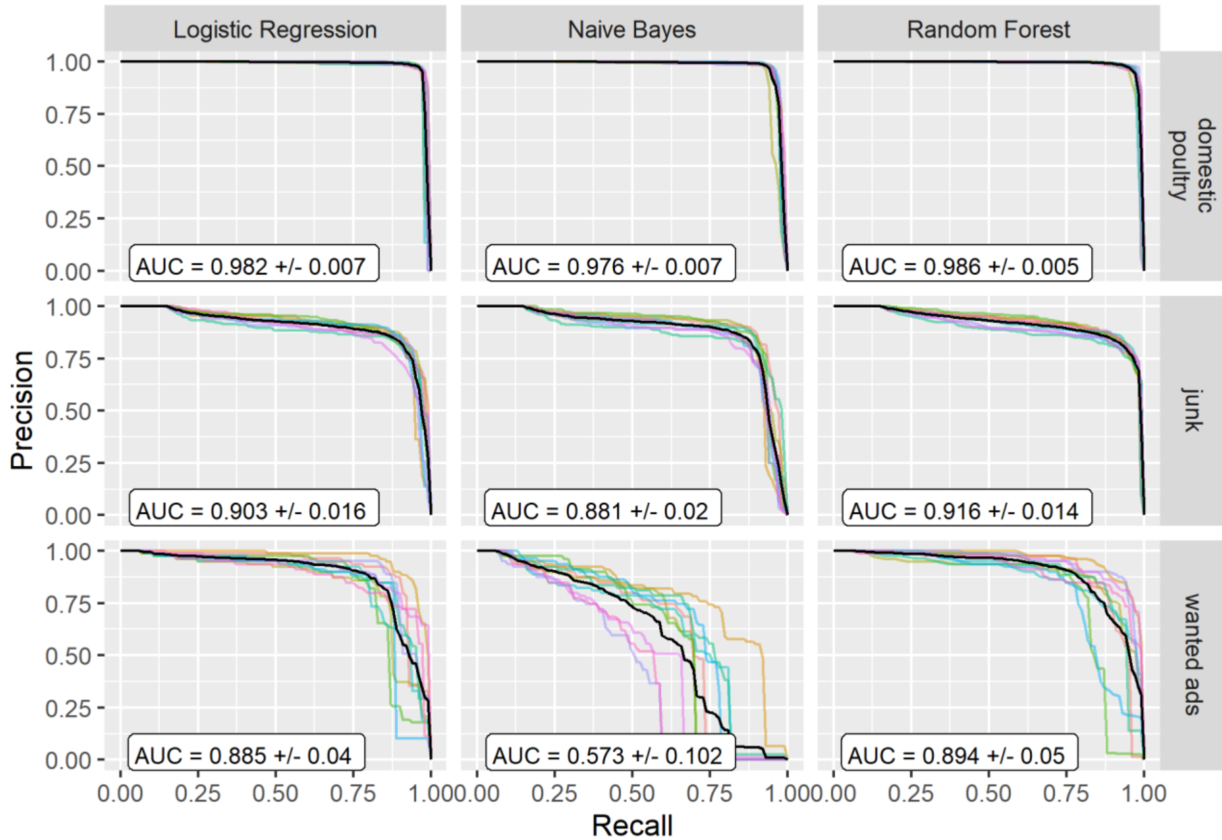


Figure 3.

Precision recall curves and the area under the curve (PR AUC). Three different text classifiers (columns) were tested across three different labels (rows). For each panel, each line represents one cross-validation fold and the solid black line represents the average across all cross-validation folds. Average AUC (area under curve) values are reported with standard deviation.

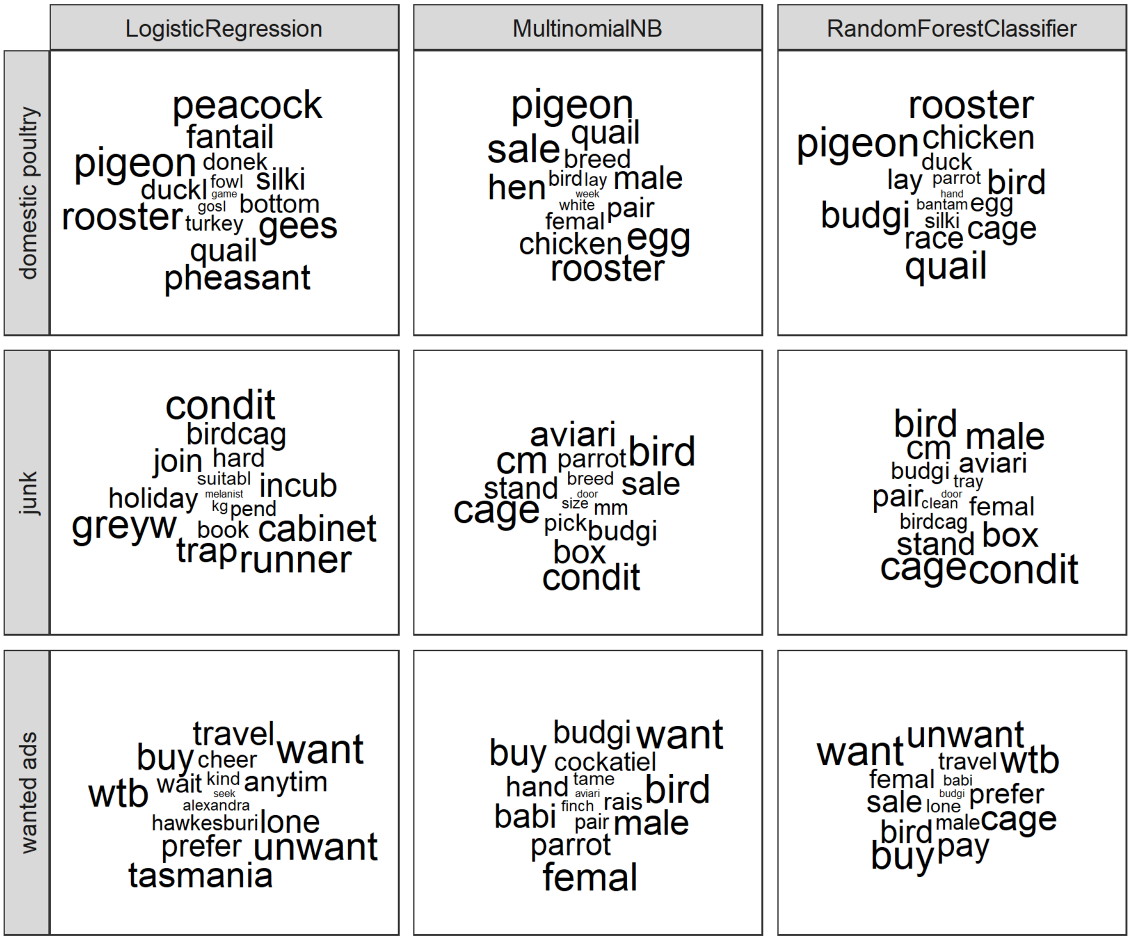


Figure 4.

Word clouds of top words (i.e., features or grams) for each label (rows) and classifier (columns). The size of the word corresponds to importance, where larger words indicate higher importance. Note that words are stemmed (e.g., condition is stemmed to condit).

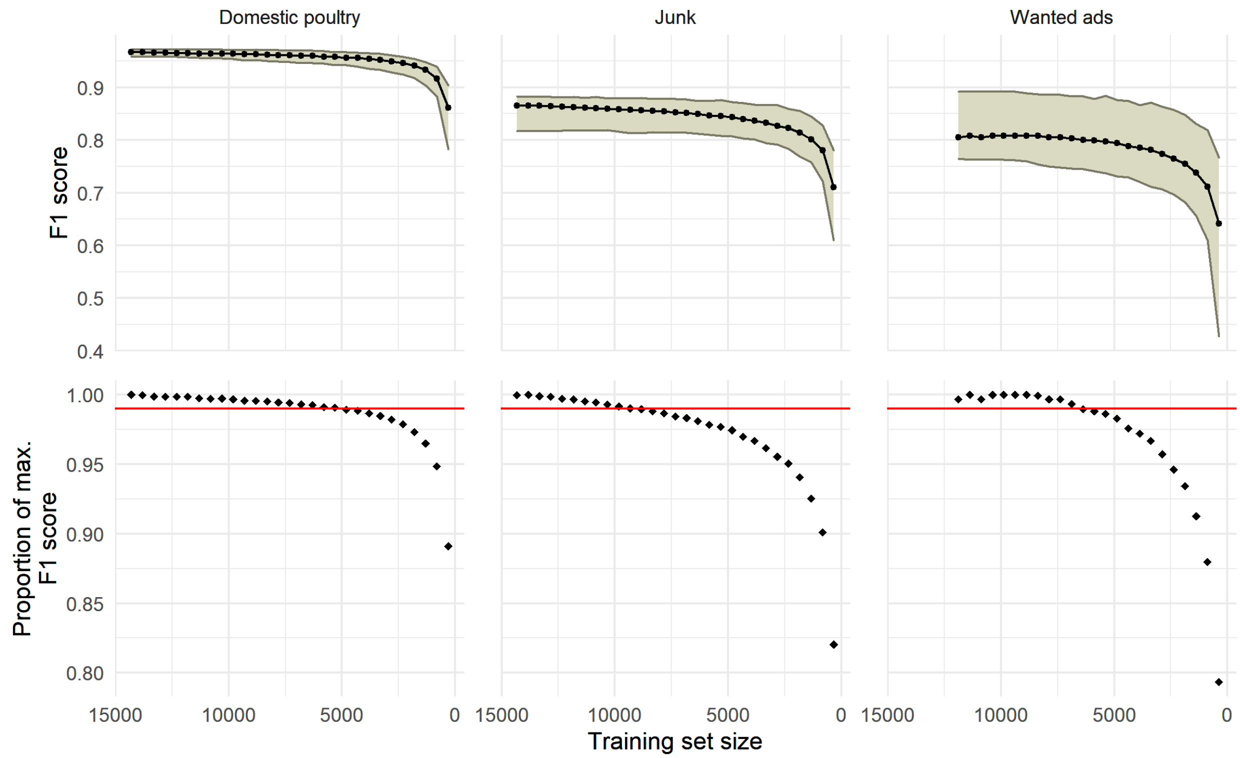


Figure 5.

The effects of reducing sample size on text-classifier model performance. Top row: the F1 score evaluated at decreasing sample size (training set) values. Ribbons represent the 95% quantile range from 100 iterations of 10-fold cross validation logistic regression text classification, repeated for each specified label (‘domestic poultry’, ‘junk’, and ‘wanted’). Bottom row: the proportion of the maximum F1 score, evaluated at each sample size, for each label. Only the median value was considered. The red horizontal line represents 0.99 of the maximum F1 score.

Supporting Information

- Appendix S1: Definitions of metrics used
- Appendix S2: Table of model metrics
- Appendix S3: Confusion matrices for all models

Appendix S1: Definitions of metrics used

Confusion matrix derived metrics

We evaluated several commonly used machine learning diagnostic metrics derived from confusion matrix values (Appendix S3): true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP) (Fielding and Bell 1997). *Precision* is the proportion of correctly predicted positives compared to all predicted positives. *Recall* is the proportion of correctly predicted positives compared to all observed positives. The *Negative predictive value* is the proportion of correctly predicted negatives compared to all predicted negatives. Finally, the *Specificity* is the proportion of correctly predicted negatives compared to all observed negatives. Mathematically, the metrics are defined (Fielding and Bell 1997) as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Negative\ predictive\ value = \frac{TN}{FN + TN}$$

$$Specificity = \frac{TN}{FP + TN}$$

Further, the *F1 score*, is defined as the harmonic mean of precision and recall, mathematically:

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Receiver operating characteristic (ROC) curve

The ROC curve shows the performance of a classification model at varying classification thresholds (Fewcett 2006). The curve plots two metrics: False positive rate (i.e., 1 - *Specificity*) and True positive rate (i.e., *Recall*). For each classification threshold (e.g., from 0.01 to 1.0 by units of 0.01), the false positive rate and true positive rate are plotted (e.g., main text Figure 2). The area under the curve for the ROC curve (ROC AUC) is a measure of the positive predictive ability of the classification model (e.g., the ability to predict true positives versus false positives), where an ROC AUC of 0.5 represents positive predictive ability equivalent to chance and an ROC AUC of 1 represents perfect positive predictive ability.

Precision-Recall (PR) Curve

Like the ROC curve, the precision-recall (PR) curve also displays the performance of a classification model at varying classification thresholds. However, for the PR curve, the tradeoff between *Precision* and *Recall* is examined (not the True versus False positive rate examined in ROC curves). The PR curve is useful when there are imbalanced class sizes (i.e., far fewer positives than negatives) because it does not consider true positives in its calculation (Sofaer et al. 2018).

References

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. https://doi.org/10.1017/S0376892997000088

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Appendix S2: Table of model metrics

The macro-averaged values (10-fold cross validated) of model performance metrics for each label-classifier combination.

Classifier	ROC AUC	PR AUC	F1 score	Precision	Recall	NPV	Specificity
domestic poultry							
Logistic Regression	0.996	0.982	0.966	0.969	0.964	0.990	0.991
Naive Bayes	0.994	0.975	0.958	0.938	0.979	0.994	0.982
Random Forest	0.997	0.986	0.959	0.969	0.950	0.986	0.991
junk							
Logistic Regression	0.954	0.903	0.860	0.902	0.822	0.969	0.984
Naive Bayes	0.952	0.879	0.857	0.866	0.849	0.973	0.977
Random Forest	0.960	0.914	0.866	0.931	0.810	0.967	0.989
wanted							
Logistic Regression	0.981	0.886	0.815	0.878	0.764	0.986	0.993
Naive Bayes	0.939	0.579	0.614	0.641	0.600	0.976	0.978
Random Forest	0.987	0.893	0.775	0.913	0.676	0.981	0.996

Appendix S3: Confusion "matrices"

The median number (10-fold cross validated) of true positives, false negatives, false positives, and true negatives for each label-classifier combination.

Label	Classifier	True positive	False negative	False positive	True negative
	Logistic Regression	348	13	11	1272
domestic poultry	Naive Bayes	354	8	22	1262
	Random Forest	344	18	11	1272
	Logistic Regression	205	42	22	1378
junk	Naive Bayes	212	34	29	1371
	Random Forest	200	46	14	1386
	Logistic Regression	58	20	8	1288
wanted	Naive Bayes	48	30	22	1274
	Random Forest	54	24	4	1292