Automatic Document Classification for Environmental Risk Assessment

Kyle Painter[1], Steven J. Dutton[2], Elizabeth Oesterling Owens[2], and Lyle D. Burgoon[2,*]

[1]Oak Ridge Institute for Science and Education, Research Triangle Park, North Carolina

[2] United States Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment, Research Triangle Park, North Carolina

[*]corresponding author e-mail: Burgoon.Lyle@epa.gov

Abstract

Motivation: In environmental risk assessment, information about potential health risks of chemicals released into the environment is compiled and distilled for use in informing public policy. The U.S. Environmental Protection Agency (EPA) produces Integrated Science Assessments (ISA) that provide a review of literature on air pollutants, including nitrogen oxides (NO_x). That review process currently requires much human labor to evaluate thousands of potentially-relevant documents published each year, a problem this study seeks to alleviate by using automated topic classification methods.

Results: For this study, abstracts and titles of scientific documents about NO_x were labeled by subject matter experts in four domains relevant to ISAs: toxicology, atmospheric science, epidemiology, and exposure science. In addition, documents not relevant to the four domains were included to simulate the background literature that we want to filter out of consideration. The labeled documents were used to train models using a Naive Bayes Multinomial classifier, via the Weka data mining platform. Separate tests were performed using multi-class or single-class models, and including background literature or not including it. For the multi-class models, recall (% of all documents in a class that are classified correctly) for scientific domains ranged between 74% and 94%, with precision (% of classified documents that are in the desired class) between 38% and 93%, with models created with background literature performing worse than models without the background documents. Single-class models had precision that ranged from 31% to 90%, and recall that ranged from 84% to 98%, with better precision for models not using background literature, but better overall recall for models using background literature. Single-class models generally performed better than multi-class models in recall, though multi-class models without the background screen tended to be best for precision.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.300v1 | CC-BY 4.0 Open Access | received: 24 Mar 2014, published: 24 Mar 2014

Introduction

In environmental risk assessment, information about potential health risks of chemicals released into the environment is compiled and distilled for use in policy recommendations. The National Center for Environmental Assessment (NCEA) within the United States Environmental Protection Agency (EPA) is tasked with developing assessments that are used to inform public policy. Scientific literature, identified from various electronic databases, provides the information that will be synthesized and evaluated in each assessment. For example, a recent literature search on nitrogen oxides identified over 79,000 potentially relevant documents from PubMed and Web of Science. Manual screening by scientific experts of the entire result set would require considerable time and effort, despite the fact that only a subset, perhaps a couple thousand, of the identified references will be included in the assessment. NCEA is currently examining ways to streamline various parts of the assessment development process, including the literature search and screening step, while keeping that process transparent to all relevant stakeholders. The goal of this study to determine the effectiveness of using automatic document classification to sort scientific literature so that scientists can spend their time considering the impact of relevant studies instead of looking at studies irrelevant to their task.

Specifically, this study focuses on the literature selection process for the Integrated Science Assessments (ISA). Sections 108 and 109 of the Clean Air Act (CAA) govern the establishment, review, and revision, as appropriate, of the National Ambient Air Quality Standards (NAAQS) to provide protection for the nation’s public health and the environment. ISAs are reports that provide a concise review, synthesis, and evaluation of the most policy-relevant science to serve as the scientific foundation for the review of the NAAQS. EPA has set NAAQS for six principal pollutants, which include: ozone, particulate matter, carbon monoxide, sulfur oxides, lead, and nitrogen oxides (NO_x). Called “criteria pollutants”, these originate from numerous sources and are generally considered harmful to public health and the environment. All ISA documents are vetted through a rigorous peer review process, including review by the Clean Air Scientific Advisory Committee and the public.

Since the ISAs play a critical role in informing public policy, the literature selection process needs to be both transparent and comprehensive. Given the large amount of labor required to meet these requirements, NCEA is currently developing methods to streamline and automate the process. Transparency has been aided by the creation of the Health and

Environmental Research Online (HERO) database (US EPA, 2008), a publicly available database that tracks citations used for NCEA publications. NCEA keeps recorded documentation of literature considered for inclusion, even if not cited in the ISAs, within the HERO database. Comprehensiveness is achieved by searching multiple indexing services, resulting in the identification of tens to hundreds of thousands of potential references per ISA. These documents must be examined by subject matter experts (SME) in one of several disciplines. This process follows a tiered evaluation strategy (US EPA, 2013). Documents from broad searches of multiple databases are first screened for topical relevance by looking at titles only. Documents are then routed to an SME of the relevant discipline to be considered for inclusion in the final assessment based on an evaluation of the scientific merits as determined first from a reading the abstract, and then eventually via the full text. Fig. 1 illustrates this process. This method requires a substantial number of man-hours to narrow down the list of documents to incorporate into the ISA. The initial screening and routing to SMEs presents an ideal scenario for computerized topic detection via classification algorithms.

The large influx of published material over the last quarter century has increased the difficulty of sorting through that material to find relevant documents, leading to a number of machine learning techniques to deal with the problem. Machine learning entails developing techniques and algorithms that allow computers to learn valuable patterns. One technique is document classification, the process of automatically extracting features from a document (usually text, but may also include bibliographic information and other metadata) and using an algorithm to create a model that predicts the class of new documents. Increasingly, the technique has been applied to sort through scientific, especially biomedical, literature (Cohen and Hersh, 2006). For example, Yu *et al*. (2008) identify gene association documents using a support vector machine-based classifier, and Wang *et al*. (2007) classify documents about epitopes using the Naïve Bayes algorithm. Similar to our work is the study by Hempel *et al*. (2012), which uses document classification methods to identify documents relevant to quality improvement, a type of health care literature review and evaluation similar to the environmental assessment process in that it requires searching literature across a number of scientific domains. They are primarily looking for literature about novel health procedures and outcomes, whereas ISAs typically are seeking literature in several pre-defined domains. However, we both find that PubMed's indexing of documents with Medical Subject Headings (MeSH) terms as keywords provides an inadequate solution to finding the relevant documents for such extensive reviews. We have found that the topical indexing of PubMed and other databases like Web of Science does not suit our needs because no database covers enough of the scientific literature, the indexing often lags behind publication date, and because taxonomy terms do not always align with categories most useful for our purposes.

This study aimed to classify a large scale broad pollutant search into scientific disciplines (i.e., epidemiology, toxicology, atmospheric science, and exposure science), which would simulate the initial screening step of an ISA literature review. Naïve Bayes (NB) was chosen as an easily-implemented baseline algorithm, although other benefits of using this algorithm were discovered. We believe this is the first time that document classification has been applied to sort references that will be used to develop environmental assessments that inform public policy.

Methods

Dataset Generation

A broad search was conducted to identify references related to the health effects of nitrogen oxides for use in the ISA for NO_x. This search was conducted on PubMed and Web of Science databases using a large set of search strings for nitrogen oxides (See Table S1). This search returned 79,511 distinct peer-reviewed documents published from 2008 to 2011. From this search, two datasets were generated: 1) a set of documents in each of several distinctly-defined domains and 2) a set of documents that are not relevant to those domains. The latter simulates the background literature to test specificity of the models. To develop the first dataset, subsets of documents corresponding to four scientific domains (atmospheric sciences (*As*), epidemiology (*Ep*), exposure sciences (*Es*), and toxicology (*Tx*)) were selected by SMEs in each particular discipline working independently of each other. Each subset contains at least 317 documents. To create the non-relevant set, the SMEs devised a set of discipline-specific exclusion terms from ranked frequency lists of journal names and non-overlapping single- and multi-word phrases generated from the reference title field. The documents that were excluded from all four discipline categories were placed in the category "Other" (*Oth*). This list of 8090 non-relevant references was not exhaustively checked by SMEs from all four domains, which would have been prohibitively time-consuming. Spot-checking gave us confidence that this set of references is a reasonable representation of the background of non-relevant documents from which we are trying to differentiate domain documents. Additionally, a large enough pool of non-relevant references should ensure that even if a few domain-specific references made it into that category, their influence on any final results would be minimized.

Some documents were tagged with multiple topics. These documents tend to be more substantial review articles and reports rather than single-study journal articles. In order to reduce the noise introduced by multiple-tagging, all tests used only documents that had been tagged with a single topic. Table 1 summarizes the number and topic of documents in the dataset.

Classification: Pre-processing and Algorithm

Classifier features were words extracted from document titles and abstracts. While every document had a title, a few did not have abstracts. The text for abstracts and titles were combined then tokenized using the Punkt tokenizer from the Natural Language Toolkit (in python) (Bird, Loper, and Klein 2009). Punctuation and word order were removed, leaving only word vectors that retained frequency counts for use in a standard bag of words representation of the documents. No stemming was used. Additionally, all html tags and a selection of common stop words were removed. Only the top ~3000 terms were used to create the models. Classification runs were performed using the NB algorithm, via Weka 3.6.8, an open-source machine learning software package (Hall, et al, 2009). In particular, these tests used Weka's NaïveBayesMultinomial implementation, which takes into account word frequency per document (McCallum and Nigam, 1998).

The standard Bayes equation (equation 1) finds a conditional probability of one event given a second event (P(A|B)) using knowledge of the reverse conditional probability (P(B|A)), and the independent probabilities of both events. For document classification (equation 2), the goal is to find the probability of a topic/class C given a document D. We can find P(C) from the proportion of classes in the original data set, or from what proportion we might expect to see in future data. Since NB seeks to find the best topic to match any given document, the P(D) term, which would otherwise be difficult to calculate, is simply dropped, as it would be the same for any comparison of classes. P(D|C) can be calculated as the product of the independent probabilities of each word appearing given the class; to avoid multiplying hundreds of small probabilities, this is typically simplified to taking a sum of the logs, which retains the relative rank that a document receives for each class (equation 3). After the models are created for each class, new documents are scored for how well their terms align with each model, and the highest scoring model is the predicted class.

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \qquad (1)$$

$$P(C_i \mid D_j) = \frac{P(D_j \mid C_i) * P(C_i)}{P(D_j)} \qquad (2)$$

$$\begin{aligned} P(D_j \mid C_i) &= \prod_n P(w_n \mid C_i) \\ P(D_j \mid C_i) &\propto \sum_n \log(P(w_n \mid C_i)) \end{aligned} \qquad (3)$$

NB is considered "naïve" because it assumes that each word in a document is independent from every other word in the document. In practice, we know that this is not how language works, but nonetheless, NB tends to have robust results. In addition, because a topic model is based only on the observed proportion of words in the test set, the model can be updated quickly and independently of the other topics, unlike many other computation-heavy classification algorithms. For the NO_x data set, document classification was done in the following ways:

- Multi-class classifier not including documents from the *Oth* class.
- Multi-class classifier including *Oth* documents.
- Single-class classifiers, in which each of the four scientific topics was tested independently.

All tests were performed using 10-fold cross validation. Results were evaluated using measurements of precision and recall, which are defined as follows:

- Precision = (True Positives) / (True Positives + False Positives)
- Recall = (True Positives) / (True Positives + False Negatives)

Results

Dataset

A dataset of titles and abstracts of scientific documents were generated and labeled by SMEs with domains for classification as described in the Methods. A few documents were considered relevant for multiple domains; those documents were eliminated from the dataset to avoid noise.

Document classification, multi-class without *Oth* documents

Our first experiment (Multi-1) tested how well the NB algorithm would predict topics of the NO_x references when *Oth* documents were not included. This test resulted in overall precision of 0.891 and recall of 0.892 (See Table 2; in this table and those that follow, columns in a confusion matrix indicate the number of documents that the model *predicted* for each topic (-P), while rows are the documents' gold-standard topic labels (-T for *true* topic)). Precision (prec) rates for individual topics ranged from 0.786 (*Es*) to 0.938 (*Ep*). Recall (rec) rates ranged from 0.767 (*Es*) to 0.945 (*Tx*).

Document classification, multi-class with *Oth* documents

The second test, Multi-2, included the references classified as *Oth* to test how well the NB algorithm would predict the topics of the NO_x references when added to a larger group of background documents (Table 3). Overall precision of the four target topics dropped to 0.702, whereas recall decreased to 0.853. Since we were not interested in producing a model for identifying *Oth* documents, those results are not included in the overall performance metrics for this test. Precision of individual topics was lower, ranging from 0.388 (*At*) to 0.865 (*Ep*), while recall ranged from 0.741 (*Es*) to 0.936 (*Ep*). Compared to the results without *Oth* documents, the precision was much lower for this test, but recall rates were only slightly lower.

Document classification, single-class with and without *Oth* documents

This round of tests predicts a single class for each test; there were separate tests for each of the four relevant domains, with the classifier choosing between the desired domain and the collection of the documents from the three other domains. Tests were performed both with no *Oth* documents (Single-1) and including all *Oth* documents (Single-2). The results of these tests are found on the left side of Table 4, with only precision and recall from the desired class reported. For Single-1, recall was higher in each topic compared to either multi-class test, and precision was higher than Multi-2 in all categories except *Es*, but was lower than Multi-1 in all categories. For Single-2, precision was lower than either multi-class test, but recall was higher for all categories except *Tx* of Multi-1. Compared to Single-2, precision of Single-1 was higher but recall was lower except for *Tx.*

Document classification, language models

To create a topic model, NB calculates the probability that any random term picked out of the bag of words is a given term. Table 5 shows the twenty most common terms in each of the five categories, along with their associated probabilities. However, the most common terms are not always the most determinative, as multiple categories can have similar highly ranked terms. Table 6 shows only those terms that are at least three times more likely to appear in the given category than any other category. Because NB differentiates based on the underlying language model of all domains being classified, terms that are similar among classes do little predictive work. The terms in Table 6, on the other hand, are much more likely to discriminate between these particular classes.

Discussion

NB models have been successful at producing high quality document classification results in studies involving biomedical texts. For example Frunza et al. (2011) and Barrajo et al. (2011) have recently used NB to classify scientific documents for systematic reviews and other cases like ours where the desired documents are vastly outnumbered by the non-desired documents. While other algorithms were considered, we chose NB because it is 1) simple to implement, 2) easy to explain intuitively to end users, 3) very fast, 4) completely transparent, and 5) well-regarded for its effectiveness on textual data. Given these factors and the results, we determined NB is a good choice for the identification of domain documents to consider for use in environmental risk assessment contexts.

A key question throughout this project was choosing the best measure of quality for this context. For scientists performing a broad comprehensive assessment of the environmental and health effects of chemicals, the most important criterion in filtering documents for further examination is to not miss any potentially relevant document. This is especially important in cases where those assessments ultimately have bearing on policy decisions. A false negative is highly problematic for this goal, as that document remains essentially invisible to the researcher. A false positive, on the other hand, adds only a marginal amount of work for the scientist who is manually examining classification results. Since no machine learning scheme (like any corollary human endeavor) is going to be perfect, it is desirable to tune a system to fail in the best direction. In this context, that means favoring low false negative rates/high recall. On that metric, the topic classifiers performed well, usually surpassing 80% recall. Precision was consistently high, but there were cases where precision fell as low as 31%. There is always a tradeoff in machine learning contexts between precision and recall. So while this method has lower precision rate than we ideally want, the context warrants prioritizing for higher recall.

These results suggest that using NB for document classification could significantly lower the time it takes to sort literature for environmental risk assessment. As described in the introduction, current methods for sifting through the literature are time-intensive and rely on various search engines using keywords to compile literature to be searched. The probability of terms in the models, as exemplified in Tables 5 and 6, demonstrate how document classification via NB can be superior to prior keyword-based methods at pinpointing worthwhile documents to read. Some of the results in Table 5 are curious. “Exposure” is highly ranked for exposure sciences, but it has a higher probability of indicating a toxicology document. “Air” is important in all four domains, but is ranked higher in epidemiology than atmospheric by a factor of three. Table 6 lists only those terms that are three times more likely in a domain compared to the other domains, and therefore it lets us see the terms that are most driving classification in this particular scenario. If we were comparing a different set of domains, these lists would be different. Even these lists are somewhat incomplete, for what drives classification is the entire widely defined feature space of, in this case, about 3000 terms. A list like table 6 is also useful to hand to end-users so that they understand how the system produces its results, which increases transparency. The benefit of automatic classification methods is that we do not have to guess which of those terms might be the best indicators for a given class. Rather, the best terms for any given classification context will bubble up when the algorithm is run. Keyword-based searching has its place (indeed, it is where the initial large set of documents comes from), but currently that method is not as effective at narrowing down documents based on ad hoc domain criteria.

Beyond this method’s potential increase for productivity, there are some extensions which could increase the method's effectiveness. The model described here is static, in that it uses only a set of pre-labeled data to create topic models. But the linguistic patterns of scientific domains change over time, often in subtle ways, so a model that moved with those patterns would be preferable. NB is a method that allows for quick updating. Probably the most well-known use of NB is spam detection, which can be updated in near real time and can be easily personalized as well (see, for example, Delaney, 2005). The class independence discussed above, along with the probabilistic foundation of the algorithm, make NB well-suited for quick updating.

There are two key results of this fact. First, the underlying model could be improved as an SME is analyzing results. As they verify (or not) results that the model classified, those judgments can be incorporated into the model quickly to incrementally improve results (at least up to a saturation point), even in the same session. Since their workflow will necessarily entail what amounts to an informal verification step anyway (they have to read the document to complete their comprehensive assessment), this is a benefit from their work that we can essentially get for free. Second, whenever there is a new domain to be classified, the model could be seeded with a small number of documents, and then progressively get better at predicting classes using this kind of updating. In cases where current assessments cover the same subject domains as one completed in the past, documents cited by the older assessment could be used to induce a model to classify results for the new assessment.

There are limitations to this method. Like all modeling activities, the models do not perfectly capture reality, and therefore there will be mistakes. As argued above, false positives are generally not much of a problem in the context we are considering. False negatives, however, can be. While a system may be engineered to decrease false negatives, they cannot be eliminated completely. But this will be true no matter what humans or algorithms are filtering the results. One way to deal with this limitation is to have a protocol that allows for documents to be considered when the algorithm passes on them. The EPA already regularly solicits public and peer review comments to help address this problem. The need for a protocol to find missed documents is one that cannot be avoided due to current limitations of classification technology, but the methods described in this paper will likely decrease the need to use of that kind of protocol.

Another limitation is the difficulty of producing a suitable background screen, here labeled as *Oth*. In the experiments above, we created a set of documents to serve as a model for the diverse range of what out-of-topic documents tend to look like. The results obtained using *Oth* documents was mixed. In the multi-class tests, the inclusion of *Oth* reduced precision and recall for all categories. However, for single-class results, recall was higher for three classes using *Oth* documents, though precision was somewhat lower for all classes. Since our project values recall over precision, the single-class model with *Oth* documents tends to be the best performing model. Given that many relevant documents like government reports and interdisciplinary research articles can naturally fit into multiple categories, single-class models allow those documents to be pushed to multiple SMEs if warranted. In addition, using *Oth* documents as a background screen more naturally simulates a real-world classification scenario. In summary, machine learning methods show great promise for improving the literature sorting process for environmental risk assessments. Specifically, we have shown that document classification using the Naïve Bayes algorithm can identify the domain of scientific documents with a high degree of accuracy, and therefore can increase efficiency of the assessment process. In particular, single-class Naïve Bayes classifiers using a background screen of additional out-of-topic documents produces high levels of recall that are desired for this kind of assessment.

Funding

This project was supported in part by an appointment to the Research Participation Program at the National Center for Environmental Assessment, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.

References

Bird S, Loper E, Klein E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc. Available at http://nltk.org/book/ (accessed 20 March 2014)

Borrajo F, Romero R, Iglesias EL, Redondo Marey CM. 2011. Improving imbalanced scientific text classification using sampling strategies and dictionaries. *Journal of Integrative Bioinformatics*, 8, 176. doi:10.2390/biecoll-jib-2011-176

Cohen A, Hersh W. 2006. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *Journal of Biomedical Discovery and Collaboration* 2006, **1-** 4. doi:10.1186/1747-5333-1-4

Delany SJ, Cunningham P, Tsymbal A, Coyle L. 2005. A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems*, 18, 187-195. doi:10.1016/j.knosys.2004.10.002.

Frunza O, Inkpen D, Matwin S, Klement W, O'Blenis P. 2011. Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, 51, 17-25. doi:10.1016/j.artmed.2010.10.005

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 10-18.

Hempel S, Shetty KD, Shekelle PG, Rubenstein LV, Danz MS, Johnsen B, Dalal SR. 2012. *Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations*. Rockville (MD): Agency for Healthcare Research and Quality (US). Available at http://www.ncbi.nlm.nih.gov/books/NBK109711/ (accessed 20 March 2014)

Mccallum A, Nigam K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*.

U.S. EPA. 2008. Health and environmental research online (HERO). http://hero.epa.gov/ (accessed 20 March 2014)

U.S. EPA. 2013. Preamble. *Integrated Science Assessment of Ozone and Related Photochemical Oxidants (Final Report)*. U.S. Environmental Protection Agency, Washington, DC. EPA/600/R-10/076F. Available at http://oaspub.epa.gov/eims/eimscomm.getfile?p_download_id=511347 (accessed 20 March 2014)

Wang P, Morgan AA, Zhang Q, Sette A, Peters B. *et al*. 2007. Automating document classification for the Immune Epitope Database. *BMC Bioinformatics*, 8, 269. doi: 10.1186/1471-2105-8-269

Yu W, Clyne M, Dolan SM, Yespuprivya A, Wulf A, Liu T, Khoury MJ, Gwinn M. 2008. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, 9, 205. doi: 10.1186/1471-2105-9-205.

353 ## Figures and Tables

354 **Fig. 1.** ISA literature evaluation process.

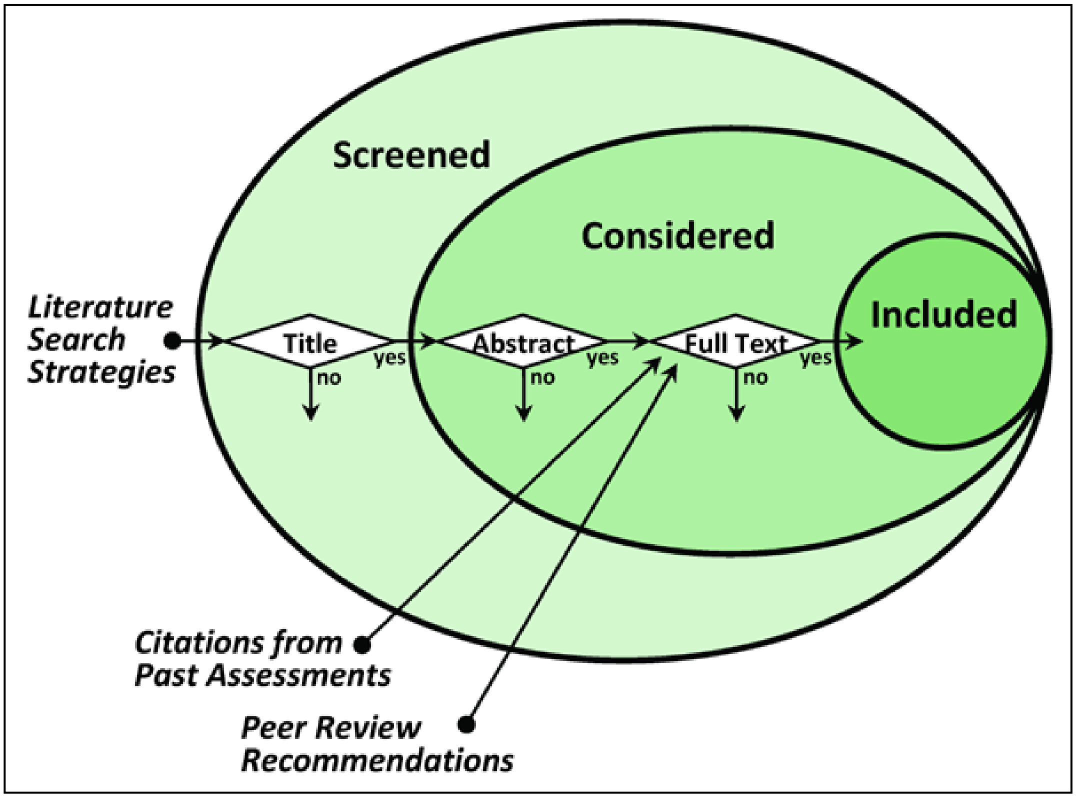
Screened
Considered
Included
Literature Search Strategies
Title
Abstract
Full Text
yes
no
yes
no
yes
no
Citations from Past Assessments
Peer Review Recommendations

Table 1. Summary of dataset, with counts of documents.

Topic	Topic Abbreviation	Number of documents
Atmospheric Science	At	355
Epidemiology	Ep	528
Exposure Science	Es	317
Toxicology	Tx	326
Other	Oth	8090

Table 2. Multi-class results, without Other documents (Multi-1)

At-P	*Ep-P*	*Es-P*	*Tx-P*		prec	rec
314	2	34	5	*At-T*	0.880	0.885
1	**496**	25	6	*Ep-T*	0.938	0.939
37	25	**243**	12	*Es-T*	0.786	0.767
5	6	7	**308**	*Tx-T*	0.931	0.945
				overall	0.891	0.892

True positives are **emphasized**.

Table 3. Multi-class results, with Other documents (Multi-2)

At-P	*Ep-P*	*Es-P*	*Tx-P*	*Oth-P*		prec	rec
303	3	33	3	13	*At-T*	0.388	0.854
1	**494**	26	2	5	*Ep-T*	0.865	0.936
34	30	**235**	10	8	*Es-T*	0.723	0.741
4	6	7	**269**	40	*Tx-T*	0.760	0.825
439	38	24	70	**7519**	*Oth-T*	0.991	0.929
					Overall (without Oth)	0.702	0.853

Table 4. Single-class results, with comparison to multi-class results (from Tables 2 and 3)

	Single-1		Single-2		Multi-1		Multi-2	
	prec	rec	prec	rec	prec	rec	prec	rec
At	0.773	0.930	0.349	**0.935**	**0.880**	0.885	0.388	0.854
Ep	0.876	0.964	0.629	**0.981**	**0.938**	0.939	0.865	0.936
Es	0.694	0.845	0.315	**0.915**	**0.786**	0.767	0.723	0.741
Tx	0.909	**0.951**	0.709	0.859	**0.931**	0.945	0.760	0.825

Highest recall and precision value for each class is **emphasized**.

Table 5. NO_x multi-class categories, highest ranked terms in each class, with probabilities that any random term picked from a document is the given term

At		Ep		Es		Tx		Oth	
nox	0.0148	air	0.0292	exposure	0.0206	exposure	0.0261	oxide	0.0061
emissions	0.0143	pollution	0.0218	concentrations	0.0182	ppm	0.0196	nitric	0.0057
nitrogen	0.0119	exposure	0.0155	indoor	0.0174	nitrogen	0.0155	study	0.0050
air	0.0091	pm	0.0149	air	0.0153	dioxide	0.0142	induced	0.0048
model	0.0081	asthma	0.0096	personal	0.0114	exposed	0.0138	treatment	0.0043
ozone	0.0075	effects	0.0093	pm	0.0103	lung	0.0107	activity	0.0040
concentrations	0.0069	study	0.0089	outdoor	0.0091	cells	0.0094	results	0.0040
emission	0.0067	pollutants	0.0088	nitrogen	0.0088	effects	0.0091	effect	0.0040
hono	0.0065	ci	0.0087	levels	0.0085	rats	0.0090	effects	0.0038
measurements	0.0065	levels	0.0080	study	0.0079	pulmonary	0.0079	ii	0.0038
results	0.0048	children	0.0074	exposures	0.0078	nitric	0.0077	acid	0.0038
high	0.0045	dioxide	0.0072	dioxide	0.0076	oxide	0.0071	group	0.0035
combustion	0.0043	results	0.0072	ambient	0.0074	mice	0.0069	increased	0.0035
atmospheric	0.0043	health	0.0072	traffic	0.0065	air	0.0062	cells	0.0035
data	0.0043	associations	0.0069	concentration	0.0060	animals	0.0059	levels	0.0034
oxides	0.0042	increase	0.0068	pollution	0.0059	alveolar	0.0057	expression	0.0033
study	0.0042	ambient	0.0065	model	0.0058	effect	0.0052	water	0.0031
formation	0.0041	risk	0.0065	data	0.0057	increased	0.0051	high	0.0031
production	0.0041	mortality	0.0064	measurements	0.0053	concentration	0.0051	sildenafil	0.0031
observed	0.0039	respiratory	0.0064	urban	0.0052	lungs	0.0050	nitrate	0.0031

Table 6. NO_x multi-class categories, top ranked terms where probability is > 3x the probability of that term in any of the other categories

At		Ep		Es		Tx		Oth	
nox	0.0148	pollution	0.0218	indoor	0.0174	ppm	0.0196	treatment	0.0043
emissions	0.0143	asthma	0.0096	personal	0.0114	exposed	0.0138	expression	0.0033
emission	0.0067	ci	0.0087	outdoor	0.0091	lung	0.0107	sildenafil	0.0031
hono	0.0065	associations	0.0069	homes	0.0034	rats	0.0090	structure	0.0027
combustion	0.0043	risk	0.0065	samplers	0.0033	pulmonary	0.0079	complexes	0.0024
atmospheric	0.0043	mortality	0.0064	cooking	0.0019	mice	0.0069	properties	0.0023
oxides	0.0042	association	0.0057	sampler	0.0017	animals	0.0059	erectile	0.0023
chemistry	0.0033	daily	0.0056	heaters	0.0016	alveolar	0.0057	growth	0.0021
diesel	0.0033	birth	0.0046	indoors	0.0014	lungs	0.0050	complex	0.0020
fuel	0.0032	methods	0.0045	heating	0.0013	inhalation	0.0042	hydrogen	0.0019
observations	0.0032	visits	0.0040	inside	0.0013	macrophages	0.0041	chimpanzees	0.0019
satellite	0.0031	hospital	0.0039	street	0.0012	guinea	0.0029	crystal	0.0019
tropospheric	0.0029	cardiovascular	0.0035	hydroxyl	0.0011	pigs	0.0026	dysfunction	0.0019
engine	0.0028	disease	0.0034	diffusive	0.0011	lavage	0.0024	fe	0.0018
instrument	0.0023	admissions	0.0032	formaldehyde	0.0011	lipid	0.0023	cu	0.0017
omi	0.0023	age	0.0031	outdoors	0.0010	resistance	0.0022	gene	0.0017
atmosphere	0.0022	inflammation	0.0027	uk	0.0010	damage	0.0022	iii	0.0016
coal	0.0021	conclusions	0.0025	integrated	0.0009	inhaled	0.0021	synthase	0.0016
mixing	0.0021	diameter	0.0025	housing	0.0009	epithelial	0.0019	metal	0.0015
lightning	0.0021	diseases	0.0025	drivers	0.0009	infection	0.0019	ca	0.0014