# Machine Learning and Text Mining to Evaluate Relevance and Quality of Systematic Reviews on Health Outcomes in Humanitarian Emergencies

Federico Paoletti, Nicola Trendel, Lev Tankelevitch, Pantelis Antonoudiou
University of Oxford

correspondence: federico.paoletti.fp@gmail.com

March 14, 2018

**Abstract**

**Systematic reviews (SRs) summarise evidence on the effectiveness of healthcare practices, and are a valuable tool to implement successful health interventions in humanitarian emergencies. The process of screening for relevant, high-quality evidence is often time consuming due to the vast number of published SRs, but can be potentially simplified via automation. The aims of this consulting project were to 1) implement a proof-of-principle machine learning algorithm to classify SRs as being relevant or irrelevant to health outcomes in humanitarian emergencies, and 2) automatically characterise SRs based on pre-determined quality criteria. To tackle 1), We implemented and optimised a Bernoulli Naive Bayes classifier based on Cochrane SR titles and abstracts that distinguishes relevant and non-relevant SRs (mean recall = 81 %, mean specificity = 71 %, mean accuracy = 77 %). To tackle 2), we implemented a set of text mining rules that identify key information in each SR abstract relating to six pre-defined quality criteria: the types of studies included, the number of studies, the total number of participants, the number of databases searched, the number of independent reviewers, and whether a risk of bias assessment was performed. We then applied these rules to compare the quality of relevant and non-relevant SRs. The results point to the promising use of machine learning and text mining to reduce manual screening time of SRs and facilitate the use of evidence-based medicine in humanitarian emergencies.**

# Introduction

Systematic reviews (SRs) of primary research in healthcare and health policy evaluate the effectiveness of prevention, diagnosis, treatment and rehabilitation interventions, and are widely regarded as the cornerstone of evidence-based medicine [4]. SRs on health outcomes in humanitarian emergencies are considered valuable tools during and after crises, although their use in decision-making is often scarce due to inadequate access [7]. Providing access to relevant, high-quality SRs involves time-consuming manual screening of SR titles, abstracts and often full texts.

Machine learning and text mining techniques have been proposed as valuable tools to automate the screening of primary research articles for inclusion in SRs and reduce manual screening time [12, 10]. However, it is unclear whether such methods can also be applied to automate the screening of high-quality SRs relevant to health outcomes in humanitarian emergencies. In this report, we aimed to apply machine learning and text mining to facilitate the screening of high-quality SRs relevant to health outcomes in humanitarian emergencies. In particular, our aims were to implement a proof-of-principle machine learning algorithm to classify SRs for relevance, and a set of text mining rules to evaluate the quality of SRs based on six pre-defined quality criteria.

Machine learning algorithms can be trained to categorise texts, and algorithm training requires a set of correctly classified texts. We trained our algorithm to distinguish relevant and non-relevant SRs using Cochrane SRs due to their standardised abstract section and use of language. We also developed and applied text mining rules to extract information from Cochrane SRs relevant to our pre-defined quality criteria. We simplified our analysis by only focusing on titles and abstracts (Fig. 1).

# Data Extraction

To obtain a set of relevant SRs, we extracted all Cochrane SR titles and abstracts found in the Evidence Aid website using a Python-based webcrawler script (i.e. 171 relevant SRs; WebCrawler_Main.py, Appendix C). To obtain a set of non-relevant SRs, we downloaded all Cochrane SR titles and abstracts and excluded the relevant SRs as well as method-based reviews (6835 non-relevant SRs). SR titles and abstracts were converted into comma-separated value (.csv) files, where each SR represents a row and each column represents an abstract header: title, background, objectives, search methods, selection criteria, data collection and analysis, main results, authors' conclusions (Process_RelevantSRs.py, Process_AllCochraneSRs.py, Appendix C). For both classification and quality assessment, we only considered SR titles and abstracts.

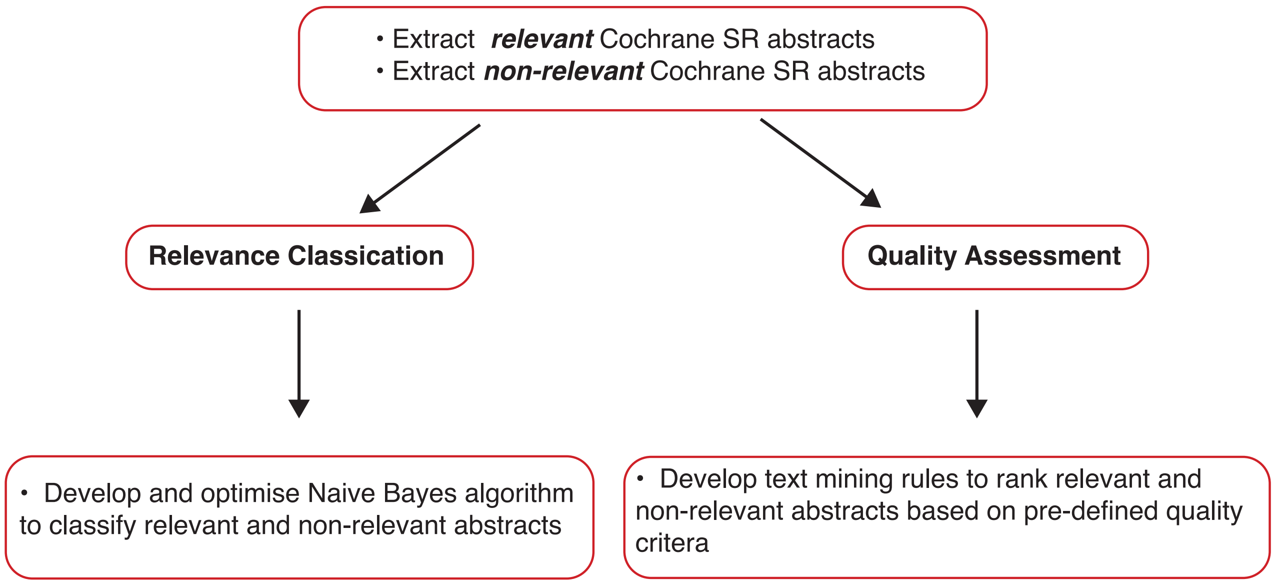

• Extract *relevant* Cochrane SR abstracts
• Extract *non-relevant* Cochrane SR abstracts
Relevance Classication
Quality Assessment
• Develop and optimise Naive Bayes algorithm to classify relevant and non-relevant abstracts
• Develop text mining rules to rank relevant and non-relevant abstracts based on pre-defined quality critera

Figure 1: Outline of the report aims. Relevant and non-relevant Cochrane SR abstracts were starting points for both relevance classification and quality assessment.

## Relevance Classification

Bernoulli Naive Bayes (BNB) is a simple, supervised machine learning algorithm for text categorisation. Here we used BNB to automatically determine whether a SR belongs to the "relevant" or "non-relevant" category. [1, 6]. The algorithm uses Bayes' Rule to evaluate the probability of a SR belonging to the "relevant" or "non-relevant" category, given the presence and absence of specific words called features [1, 6]. In our BNB implementation, words are included in the feature set if they satisfy a 1) minimum word character length and 2) a relative word frequency cutoff within the entire set of texts. It does not regard correlations between multiple words.

We used the Python Natural Language Toolkit (NLTK) to build a BNB classifier (Appendix C). We evaluated the performance of our algorithm using three standard metrics [10]: accuracy, recall and specificity (Table 1).

Table 1: Metrics used to evaluate classifier performance. TR: true relevant. FR: false relevant. TNR: true non-relevant. FNR: false non-relevant.

| Metric | Equation | Meaning |
|---|---|---|
| Recall | TR / (TR + FNR) | % of correctly classified relevant SRs |
| Specificity | TNR / (TNR + FR) | % of correctly classified non-relevant SRs |
| Accuracy | (TR + TNR) / (TR + TNR + FR + FNR) | bulk % of correctly classified SRs |

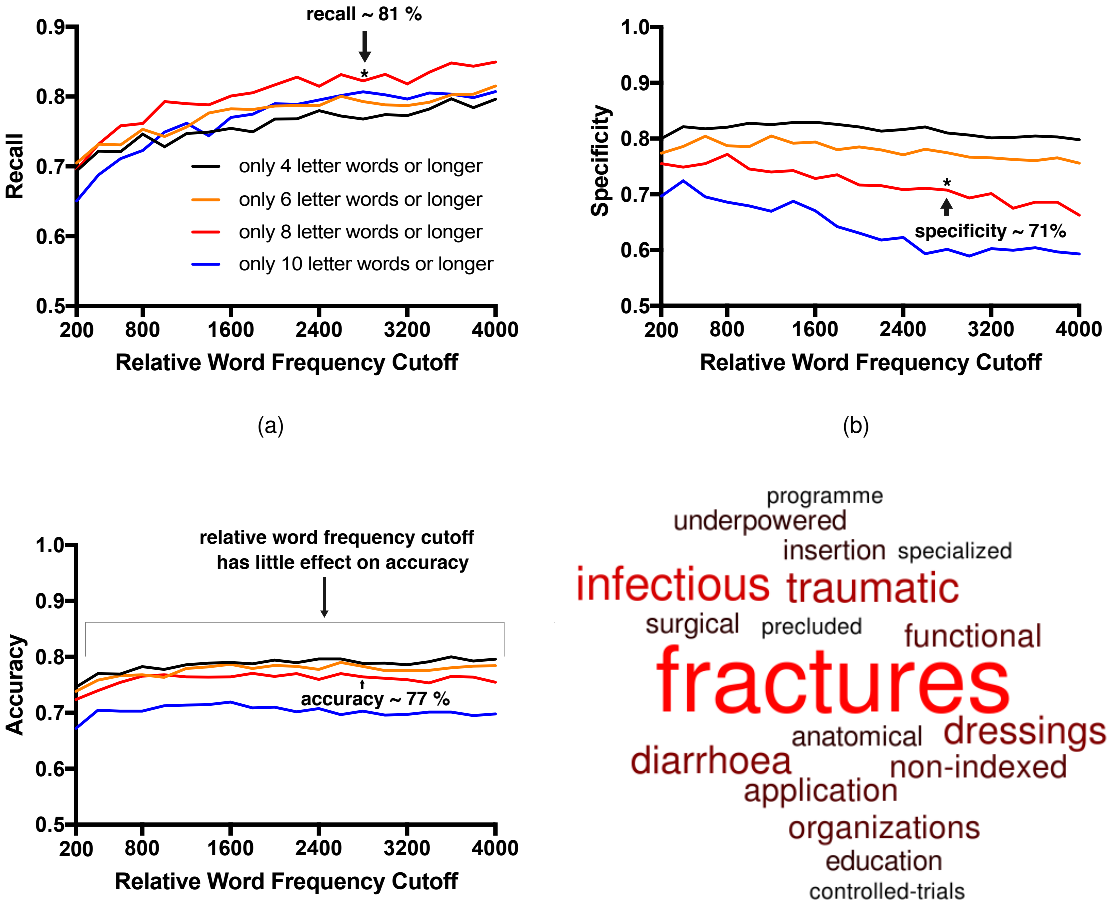
To optimise our BNB classifier, we varied three parameters: 1) the ratio of relevant to non-relevant SRs (R:NR ratio) for classifier training, 2) the minimum word character length, and 3) the relative word frequency cutoff. We varied the R:NR ratio because it affects the fraction of words from relevant and non-relevant SRs that are included in the feature set (e.g. a higher number of non-relevant SRs leads to a higher proportion of words from non-relevant SRs in the feature set). We assessed the impact of the R:NR ratio using 1:1, 1:2 and 1:4 R:NR ratios. We varied the minimum word character length to test whether the inclusion of shorter or longer words would improve performance. To assess the impact of minimum word character length, we tested performance using minimum word lengths of 4, 6, 8 and 10. We varied the relative word frequency cutoff to test the impact of including more frequent (low cutoff) or less frequent (high cutoff) words as features on algorithm performance. To assess the impact of the relative word frequency cutoff, we used frequency cutoffs spanning from 200 (i.e. include only the top 200 most frequent words as features) to 4000 (i.e. include only the top 4000 most frequent words as features).

For each parameter set, we evaluated performance 100 times using the following steps: we chose a random set of non-relevant SRs from the original 6835 set, according to the chosen R:NR ratio (e.g. 1:1 R:NR ratio - all 171 relevant SRs plus 171 randomly chosen non-relevant SRs). We then randomly allocated 100 SRs for algorithm testing and the remaining SRs for algorithm training.

We found that a 1:1 R:NR ratio (i.e. balanced data sets: 171 relevant SRs and 171 non-relevant SRs) maximises recall compared to a 1:2 (171 relevant SRs, 342 non-relevant SRs) or a 1:4 R:NR ratio (171 relevant SRs, 684 non-relevant SRs) (Fig. S1). Using this R:NR ratio, we maximised recall to 81 %, kept specificity above 70 % and accuracy around 77 % on average with a minimum word character length of 8 and a relative word frequency cutoff of 2800 (Fig. 2a - 2c). An 80 % recall can be interpreted as 20 % of relevant SRs being on average mislabeled as non-relevant, while a 70 % specificity as 30 % of non-relevant SRs being on average mislabeled as relevant. We also observed 1) a tradeoff between recall and specificity when varying the minimum word character length and the relative word frequency cutoff (Fig. 2a - 2b), and 2) a negligible effect of relative word frequency cutoff on accuracy (Fig. 2c).

We then used these parameters to quantify and visualise the most informative words that distinguish relevant SRs (Fig. 2d). We saw that the words "fractures", "infectious", "traumatic", "diarrohoea" and "dressings" are most often found in relevant SRs, and suggests fracture and infectious disease management are topics that distinguish relevant SRs from non-relevant SRs. Other less informative words such as "application", "specialized", "programme" and "functional" may simply be words that were more associated with relevant SRs by chance. Context analysis would be required to fully determine why such words are more associated with relevant SRs.

Overall, these results suggest that a BNB classifier can be used to classify SRs for relevance to the humanitarian sector with relatively high accuracy and reduce manual screening time. However, the obvious trade-off in using this classifier with its current performance metrics is that, on average, around 20% of relevant SRs will be mislabeled as non-relevant. Nonetheless, we assume it is safest to have high recall and a moderate specificity as this lowers the chances that a potentially relevant SR is missed by the algorithm. This is particularly useful in the case that the user decides to only double-check SRs classified as relevant. It is likely that providing more relevant SR abstracts for classifier training will enhance recall, specificity and accuracy, therefore improving classification reliability.



(a)

(b)

(c)

(d)

Figure 2: Optimisation of the Bernoulli Naive Bayes classifier. (a) Recall plotted against the relative word frequency cutoff for minimum word lengths of 4, 6, 8 and 10. (b) Specificity plotted against the relative word frequency cutoff for minimum word lengths of 4, 6, 8 and 10. (c) Accuracy plotted against the relative word frequency cutoff for minimum word lengths of 4, 6, 8 and 10. (a - c) Each data point represents the mean of 100 randomly selected test data sets, using a constant 1:1 R:NR ratio (i.e. 171 relevant SRs and 171 non-relevant SRs). In a, the asterisk indicates a significantly higher recall for the minimum word character length of 8 compared to 4, 6 and 10 for a relative word frequency cutoff of 2800, as determined by one-way ANOVA followed by a Tukey's multiple comparisons test (F = 4.168, p = 0.008). In b, the asterisk indicates a significantly higher specificity for the minimum word character length of 8 compared to 10, and a significantly lower specificity compared to 4 and 6, for a relative word frequency cutoff of 2800, as determined by one-way ANOVA followed by a Tukey's multiple comparisons test (F = 153.7, p < 0.0001). Error bars are omitted for simplicity. (d) Word cloud representing the 18 words most frequently associated with relevant SRs compared to non-relevant SRs. The size and red colour relate to how much more likely that word is to be found in relevant SRs compared to non-relevant SRs (generated using WordItOut).

## Quality Assessment

We have developed six criteria to evaluate the quality of SRs. These criteria are not meant to provide a complete assessment of quality, but rather serve as starting points that can be quickly, automatically, and reliably extracted from SR abstracts:

- **High quality study types**: this asks whether the SR considers only randomised controlled trials, only repeated measures studies, or a mixture of several types. Importantly, though RCTs are generally considered as the highest quality type of study, other types may be more appropriate for specific questions. How each type is weighted would depend on the focus of the SR. For our purposes, we associate high quality with the presence of priority study types (Appendix B).
- **Number of studies**: This considers the total number of studies included within the SR.
- **Number of participants**: This considers the total number of participants included within the SR (i.e. the sum of all participants within each included study).
- **Is risk of bias (ROB) assessment performed?**: This asks whether a ROB assessment is performed, and we associate high quality with the presence of a risk of bias assessment [5]. Importantly, it does not consider the actual outcome of such an assessment, given the limited amount of information available in the abstract.
- **Number of databases**: This evaluates the number of databases searched, providing information on the search scope of the SR.
- **Were independent reviewers used?**: This considers whether at least two people were used to independently review studies for relevance and extract data. Cochrane has pushed for this to be the standard in recent years, but this is not always the case. We associate high quality with the use of at least two independent reviewers.

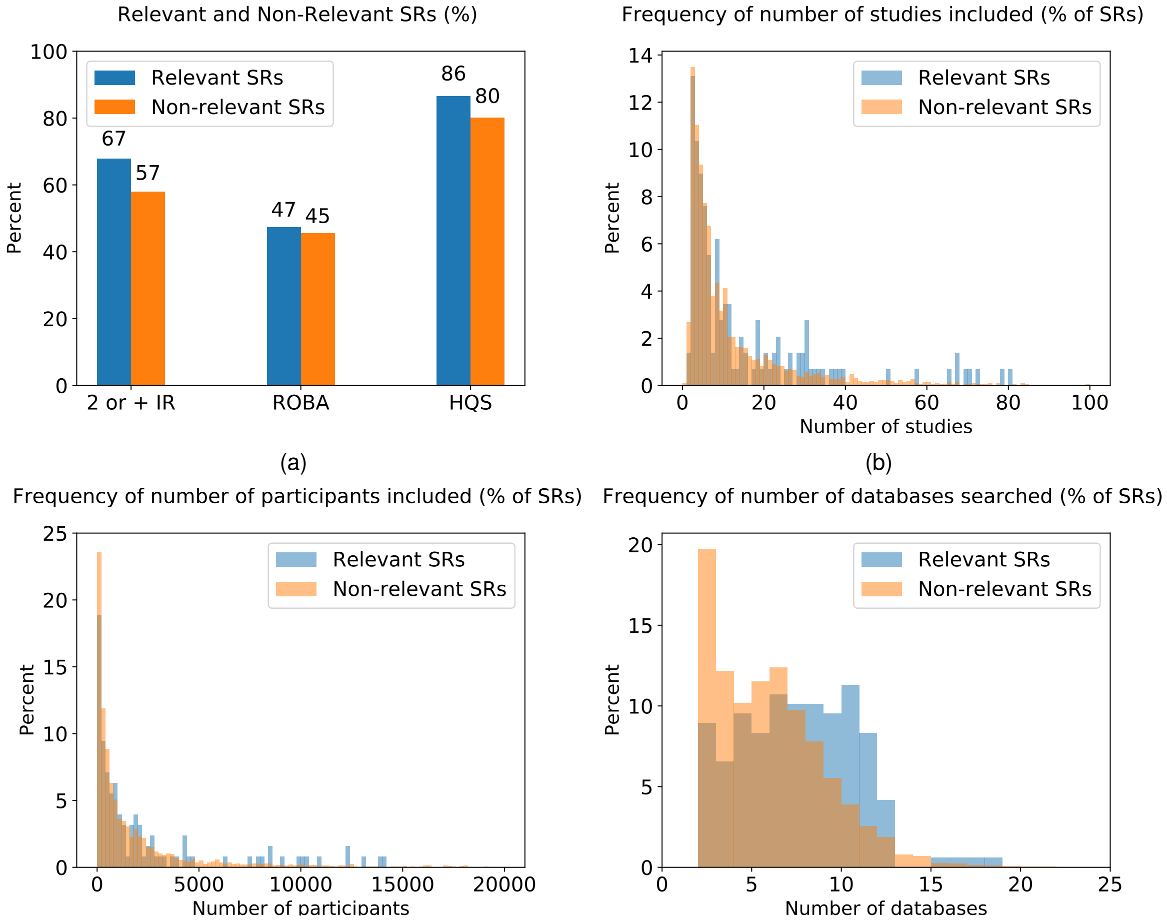
We have also developed text mining rules that automatically gather information from specific abstract sections based on our six quality criteria. It is important to highlight that we do not aim to directly assess SR quality, but rather to provide key information via text mining for users to assess SR quality (Table 2).

Table 2: Abstract sections searched and output format for each quality criterion. ROB: risk of bias.

| Criterion | Section Searched | Output Format |
|---|---|---|
| High quality study types | Selection criteria | yes / no |
| Number of studies | Main results | (raw number) |
| Number of participants | Main results | (raw number) |
| ROB assessment | Data collection and analysis | yes / no |
| Number of databases | Search methods | (raw number) |
| independent reviewers | Data collection and analysis | yes / no |

For each quality criteria, we appended a column to each SR abstract. This allows the potential user to quickly rank SR abstracts within the .csv files based on any combination of quality criteria.

We then compared relevant and non-relevant SRs based on these quality criteria (Fig. 3). We observed a slightly higher percentage of relevant SRs compared to non-relevant SRs include high quality studies, two or more independent reviewers and assess ROB, although this difference is not dramatic (Fig. 3a). In addition, the frequencies of number of studies and number of participants do not appear different between relevant and non-relevant SRs (Fig. 3b - 3c), and a larger percentage of relevant SRs compared to non-relevant SRs appeared to search between 7 and 12 databases (3d). However, statistical analysis of distributions was not carried out and such differences could simply be due to differences in sample sizes.



(a)

(b)

(c)

(d)

Figure 3: (a) The percent of relevant and non-relevant SRs that include 2 or more indendent reviewers, a ROB assessment and high quality study types. 2 or + IR: two or more independent reviewers. ROBA: ROB assessment. HQS: high quality study types. (b) The frequency of number of studies per relevant and non-relevant SRs, plotted as percent of total number of SRs. (c) The frequency of number of participants per relevant and non-relevant SRs, plotted as percent of total number of SRs. (d) The frequency of number of databases searched per relevant and non-relevant SRs, plotted as percent of total number of SRs. (a - d) For all analyses, the entire set of relevant and non-relevant SRs were used (i.e. 171 relevant SRs, 6835 non-relevant SRs).

# Conclusions

For this consulting project, our aim was to 1) apply machine learning to automatically screen SRs for relevance to health outcomes in humanitarian emergencies, and 2) automatically characterise SRs based on pre-determined quality criteria.

To tackle 1), we implemented and optimised a proof-of-principle BNB classifier using the Cochrane SRs from the Evidence Aid website as relevant SRs and all other Cochrane SRs as non-relevant SRs. Our optimal classifier on average has a recall of 81 %, a specificity of 71 %, and an overall accuracy of 77 %.

These results suggest that machine learning can potentially play a pivotal role in reducing the manual screening time required to gather key information relevant to the humanitarian sector. However, the relatively small set of relevant SRs (171) limits algorithm performance. Further manual screening would be required to generate a larger set of relevant SRs, which could then be used to train the classifier to reach a higher recall, specificity and accuracy. Another approach may be to use a different algorithm such as support vector machines (SVMs), which represent a popular alternative [10, 2]. SVMs have been shown in some instances to have higher recall for text classification compared to BNB classifiers [9].

To tackle 2), we developed a set of text mining rules to provide information on SR quality based on six pre-determined criteria. These rules can be used to evaluate and rank SRs based on any combination of quality criteria. We then applied these rules to compare the performance of relevant and non-relevant SRs based on the six criteria.

Our results suggest text mining rules can be used to rapidly characterise the quality of SRs, as a complementary tool to machine learning for SR classification. However, our rules would potentially require further refinement to include additional quality criteria such as, for example, publication bias assessment.

The main challenge of implementing quality criteria in an automated way is the extraction of relevant information from a SR's abstract or main text when such information is reported in a non-standardised way (i.e. written using different vocabulary, sentence constructions, etc.). For example, even something as clear-cut as PICOS (Population, Intervention, Comparator group, Outcome, Study design) can be reported in a wide variety of ways, making it difficult to extract this fundamental information from each abstract [8].

Without standardised reporting, reliable automatic information extraction requires natural language processing, an advanced algorithmic approach to parsing information written in natural, non-standardised language. This is a worthwhile endeavour, but beyond the scope of this project.

Cochrane has made important advancements to standardise the production and communication of SRs with the use of standardised research methods and formatting of sections within SRs. However, SRs remain unstandardised in the way that most key information is reported within a given section. This makes it challenging to ensure that the relevant information is extracted reliably and without manual intervention. Increased adoption of the PRISMA for Abstracts Checklist would enable even easier information extraction from abstracts [3].

Finally, future work based on this report could aim to use full texts instead of only titles and abstracts to improve relevance classification and quality assessment. A promising task would also be to develop a simple user interface that combines our classifier and text mining applications to simultaneously classify and evaluate the quality of input SRs.

# References

[1] Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1):4–20, 2010.

[2] Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew S.C. Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error. *bioRxiv*, page 255760, 2018.

[3] Elaine M. Beller, Paul P. Glasziou, Douglas G. Altman, Sally Hopewell, Hilda Bastian, Iain Chalmers, Peter C. Gøtzsche, Toby Lasserson, and David Tovey. PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts. *PLoS Medicine*, 10(4), 2013.

[4] JPT Higgins and S (editors) Green. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. *The Cochrane Collaboration*, 2011.

[5] Julian P T Higgins, Douglas G. Altman, Peter C. Gøtzsche, Peter Jüni, David Moher, Andrew D. Oxman, Jelena Savović, Kenneth F. Schulz, Laura Weeks, and Jonathan A C Sterne. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ (Online)*, 343(7829):1–9, 2011.

[6] Mita K. Dalal and Mukesh A. Zaveri. Automatic Text Classification: A Technical Review. *International Journal of Computer Applications*, 28(2):37–40, 2011.

[7] Bonnix Kayabu and Mike Clarke. The Use of Systematic Reviews and Other Research Evidence in Disasters and Related Areas: Preliminary Report of a Needs Assessment Survey. *PLOS Current Disasters*, 2013.

[8] Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P.A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ (Clinical research ed.)*, 339, 2009.

[9] Sridhar Mocherla, Alexander Danehy, and Christopher Impey. Evaluation of Naive Bayes and Support Vector Machines for Wikipedia. *Applied Artificial Intelligence*, 31(9-10):733–744, 2017.

[10] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):1–22, 2015.

[11] Cochrane Effective Practice and Organisation of Care. What study designs should be included in an EPOC review and what should they be called? 2017.

[12] Guy Tsafnat, Paul Glasziou, Miew K. Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. Systematic review automation technologies. *Systematic Reviews*, 3(1):1–15, 2014.

# Contributions

- Federico Paoletti extracted the relevant SRs (WebCrawler_Functions.py, WebCrawler_Main.py, Appendix C), developed the BNB classifier (Classifier_Functions.py, Classifier_Main.py, Appendix C) and wrote and edited the report
- Nicola Trendel developed the text mining rules (Quality_Functions.py, Quality_Visualise.py, Quality_Main.py, Appendix C) and edited the report
- Lev Tankelevitch developed the list of quality criteria and edited the report
- Pantelis Antonoudiou processed the relevant and non-relevant SRs (Process_RelevantSRs.py, Process_AllCochraneSRs.py, Appendix C) and edited the report

# Appendices

## A Impact of R:NR ratio on Classifier Performance

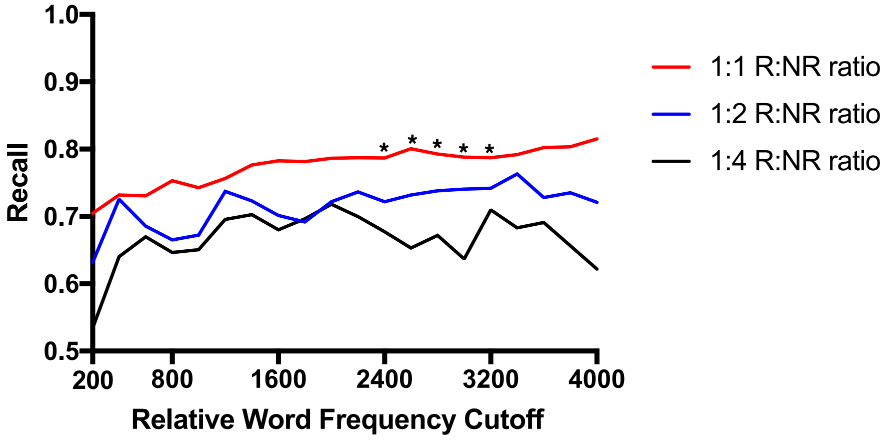


Figure S1: Recall plotted against the relative word frequency cutoff for 1:1, 1:2 and 1:4 R:NR ratios, with a constant mininimum word character length of 6. Each data point represents the mean of 100 (1:1 R:NR ratio) or 25 (1:2, 1:4 R:NR ratios) randomly selected test data sets. The asterisks indicate a significantly higher recall for the 1:1 R:NR ratio compared to the 1:2 and 1:4 ratios for the 2400, 2600, 2800, 3000 and 3200 relative word frequency cutoff values as determined by one-way ANOVAs followed by Tukey's multiple comparisons tests (for all ANOVAs, $p < 0.05$). Error bars are omitted for simplicity.

## B Types of possible studies included in SRs

Stars indicate high-quality priority types (from Cochrane EPOC guidelines [11]).

- Randomised controlled trial (RCT) OR randomised trial *
- Non-randomised controlled trial (NRCT) OR non-randomised trial OR controlled clinical trial OR quasi-randomised controlled trial *
- Cluster randomised trials *
- Non-randomised cluster trials
- Controlled before-after study (CBA) *
- Interrupted-time-series study (ITS) *
- Repeated measures study (RMS) *
- Before-after study
- Opinion paper
- Non-comparative study
- Retrospective case-control study
- Prospective case-control study
- Retrospective cohort study
- Prospective cohort study
- Non-concurrent cohort study

# C Data and Scripts

Relevant and non-relevant SR data files are available here.
All scripts are available at: https://github.com/FedericoPaoletti/EvidenceAid.

- **WebCrawler_Functions.py**: Functions required to extract the website links to the relevant SRs from the Evidence Aid website.
- **WebCrawler_Main.py**: Extracts links to all Cochrane database relevant SRs from the Evidence Aid website and produces a raw text file for each linked SR.
- **Process_RelevantSRs.py**: Processes the individual text files produces by WebCrawler_Main.py to generate a single .csv file, where each SR is a row and each column is an SR section: Title, Background, Objectives, Search methods, Selection criteria, Data collection and analysis, Main results, Authors' conclusions.
- **Process_AllCochraneSRs.py**: Processes a single text file containing all Cochrane database SR abstracts (obtained by applying the wildcard string "*" in the Cochrane search engine). Produces a single .csv file, where each SR is a row and each column is an SR section: Title, Background, Objectives, Search methods, Selection criteria, Data collection and analysis, Main results, Authors' conclusions. Relevant SRs and method-based reviews were then manually excluded.
- **Classifier_Functions.py**: Functions required to run the BNB classification algorithm.
- **Classifier_Main.py**: Applies functions from Classifier_Functions.py to repeatedly test classifier performance based on the N:NR ratio, the minimum word character limit and the relative word frequency cutoff (Fig. 2). Outputs .csv files with accuracy, false negative rate and false positive rate for each randomly sampled data set, according to the N:NR ratio.
- **Quality_Functions.py**: Functions required to individually isolate information regarding the six quality criteria. Outputs a .csv file where each SR is a row and each column is an SR section, with additional columns for each quality criteria.
- **Quality_Main.py**: Applies text mining functions from Quality_Functions.py to produce a .csv file where each SR is a row and each column is an SR section, with additional columns containing information for each quality criteria.
- **Quality_Visualise.py**: Plotting script used to generate Figure 3.