Predicting Publication Inclusion for Diagnostic Accuracy Test Reviews Using Random Forests and Topic Modelling

A.J. van Altena[1] and S.D. Olabarriaga[1]

Department of Epidemiology, Biostatistics and Bioinformatics
Academic Medical Center of the University of Amsterdam
{a.j.vanaltena, s.d.olabarriaga}@amc.uva.nl

Abstract Finding all relevant publications to perform a systematic review can be a time consuming task, especially in the field of diagnostic test accuracy. Therefore, the CLEF eHealth lab 'technologically assisted reviews in empirical medicine' was established to create a basis of comparison between various methods.
In this paper we describe a method submitted to the lab. This method consists of a topic model used to extract features and a random forest to classify the relevant papers.
Classifier performance shows and average decrease of 33.3% in workload (i.e., documents to read) when aiming for a 95% recall and 24.9% for 100% recall. However, there is a large variety in workload reduction (79.3% to 0.9%) between the diagnostic test accuracy reviews.

1 Introduction

Finding the right publications to include in a systematic review can be a time consuming task in the medical research field, especially in Diagnostic Test Accuracy (DTA) reviews. This type of research aims to summarise all evidence on a specific topic by analysing primary research, for example to study the accuracy of Lyme borreliosis tests [16]. Because systematic reviewers aim to retrieve all relevant publications, their search queries have to be very inclusive (i.e., broad). The number of results that these searches yield can range from a few hundreds to hundreds of thousands, while the searched publications (inclusions) account for only a very small part (often less than 1%). Sometimes the search strategy can be narrowed down by applying the filters that publication databases — such as PubMed, Scopus, or Ovid — provide. DTA differs from other types of systematic reviews because search filters that can select the correct type of publications are not consistent enough to deliver trustworthy output.

Many methods have been proposed to lighten the burden on systematic reviewers. With the increased popularity of machine learning for text mining, applying such techniques seems a logical step. However, the task of identifying publications for inclusion is a difficult task because the available data is mostly unstructured text.

In 2015 a study identified 44 different text mining and machine learning methods [20]. However, there are at least two issues that can make a researcher that performs systematic reviews reluctant to apply these methods: (a) the comparison between the different methods is difficult because there is no de facto performance measure; and (b) even when the workload can be greatly reduced (up to 70%), there is no guarantee of a perfect recall of all relevant publications.

To work towards solving these issues the 'technologically assisted reviews in empirical medicine' lab [15] was started as a subsidary of the CLEF eHealth labs [10]. In this lab a dataset of approximately 50 DTA studies with close to 270.000 publications was released. For 20 DTA studies the inclusion and exclusion labels were known to enable method development. To compete in the lab the labels of the other 30 studies had to be predicted.

In this paper we describe the method that we applied to this problem. To extract features from the publications the unsupervised text mining method 'Topic Modelling' was used. The features were then fed into a 'Random Forest' to classify the unknown publications.

2 Methods

In this section we describe feature extraction with topic modelling (TM), classification through Random Forests (RF), and how stability of results was assessed. More details about TM, our approach, and implementation can be obtained from our earlier work [3] and from the code [1, 2].

2.1 Feature extraction

For extracting features from the corpus TM was applied [5, 6]. TM constructs topics (i.e., lists of ordered words) by considering each word in a document and estimates two latent variables, namely topic-to-document (θ) and word-to-topic (ϕ). When two words appear together in many documents, they have a higher chance of appearing in the same topic (through the word-to-topic relationship). Also, all documents with those words have a strong topic-to-document relation to that specific topic. Note also that each document and word may have relationships with multiple topics, which is useful in the case of (bio)medical research where publications may contain many concepts (e.g., research field, methods applied, etc.).

The pre-processing, TM fitting, and post-processing steps are implemented in two packages, respectively using the PHP [1] and R [2, 21] languages.

Pre-processing consisted of preparing the documents for ingestion into the R environment and cleaning the text. Preparing for ingestion was performed using `article miner` [1]. This PHP package retrieved articles from PubMed through the public API using provided PubMed IDs. The titles and abstracts of all articles were parsed into a single CSV, the hyphens in hyphenated words were replaced by underscores to assist in further cleaning steps. Corpus cleaning was executed using the R `tm` package [9]. Processing consisted of removing punctuation, numbers, whitespace, and stop words taken from the SMART list [18, 22] (e.g., about, the, which)[1].

Fitting was performed using the same approach as in our previous work [3]. Multiple topic models were fitted with input parameters that were based on literature and previous experience. The number of topics (T) has to be provided as an input to the method, so a range of $T \in \{25, 50, 75\}$ was chosen to generate three models. Furthermore, the inputs α and β (can be considered 'smoothing' parameters for the θ and ϕ distributions, for more details see [23]) were set at $\alpha = 50/T$ and $\beta = 0.01$, and models were run for 500 iterations [23, 7]. TM results were post-processed to determine θ, which is not calculated directly by the applied TM implementation. Each of these steps was implemented in R using the `tm` and `topicmodels` packages [11, 9, 21].

2.2 Classifier

To determine whether documents should be considered inclusion or exclusion the features extracted with TM (i.e., θ matrix) were used as input to a Random Forest (RF). The RF method was chosen because of its suitability for binary outcomes (i.e., inclusion or exclusion). Training and analysis of RF outcomes was implemented using the `caret` R package [14]. The number of trees was set at 800 and determined by examining the error by number of trees graph on larger test runs (i.e., 1500 trees). Choosing the optimal number of sampled parameters per tree was done by the `caret` package using the `tuneGrid` setting. The search grid was set in increments of 10 up to the size of the input TM (i.e., number of topics, T) and included T when $T \bmod 10 \neq 0$. For example, when $T = 75$ the grid was $\{10, 20, 30, 40, 50, 60, 70, 75\}$. Performance was assessed using ROC curve and F_1-measure, where the latter expresses the average between recall and precision as follows:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

2.3 Resources

All runs were performed on cloud servers with a varying number of cores and RAM. Test runs used a larger number of cores and RAM because one model had to be trained for each T (three in total). Our method benefits from more cores as the applied packages allow parallelism, and as each TM can be trained individually. Furthermore, `caret` supports parallelising the cross-folds that are performed inside the `train` function using the `registerDoMC` function. Lastly, titles and abstracts of documents were retrieved from PubMed using the Entrez API.

[1] The full list can be found at [17]

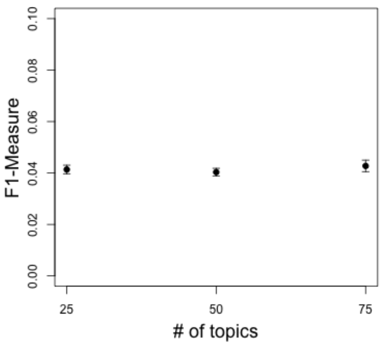

Figure 1. Mean F1-measures and their standard deviation over the cross-folds for each model based on training data.

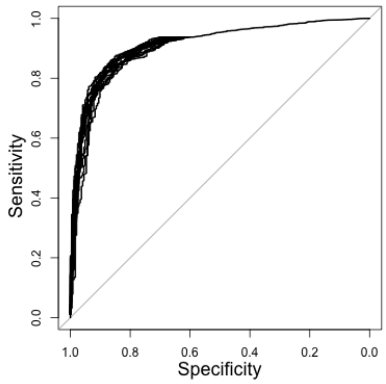

Figure 2. Receiver Operating Characteristic (ROC) curves for each model based on training data, where $T \in \{25, 50, 75\}$.

3 Results

In this section we describe the results of the training runs and the test runs for the CLEF eHealth lab. The purpose of the training runs was to fine-tune our method, whereas the test run was submitted to compete in the lab.

3.1 Corpus

Not all documents could be retrieved through the Entrez API. In the training set 38 documents are missing, and abstracts were missing for 17 included documents. The test set had 7 documents missing, it is unknown how many abstracts were missing from included documents.

3.2 Training

To achieve the optimal TM and RF settings various training runs were performed. Three different settings for T were tried to optimise the TM. For each TM a RF was trained and tested. The resulting F_1-measures are shown in Figure 1. While the individual F_1-measures are poor due to the class imbalance in the input data, little difference is visible between the different values of T. Furthermore, ROC curves for each RF are shown in Figure 2.

Optimisation of the number of trees was done according to the reported error rate (data not shown). A steep drop in error is visible between 1 and 200 trees, and the error rate remains at a plateau after 200 until 1500 trees are reached.

3.3 Testing

Results of the test run are shown in Tables 1 and 2, and Figure 3, organized by workload reduction (i.e., Work Save over Sampling, WSS). Performance outcomes are split into two groups based on whether WSS at 95% recall (WSS 95) is greater than at 100% recall (WSS 100) or not – respectively Table 1 and Table 2. This split was done to better represent the results. The group where WSS 100 is greater then WSS 95 has a smaller number of relevant documents, therefore performance outcomes act more erratically (see Figure 3-left).

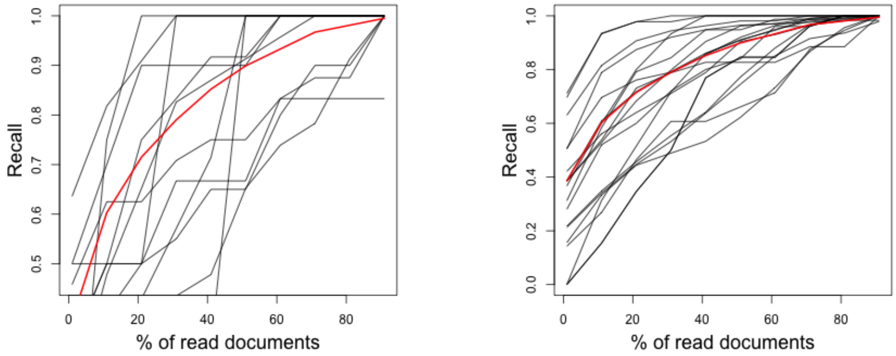

Figure 3. Recall curve for WSS 100 > WSS 95 (left) and WSS 95 > WSS 100 (right). The red line indicates the average recall curve over all DTA studies.

4 Discussion

Little variation was shown in RF performance in Figures 1 and 2. However, because fitting large TMs (i.e., many documents and topics) consumes a high amount of RAM, our implementation was limited at approximately $T = 75$. Bigger TMs failed with out of memory errors on the largest servers available. Other implementations employ an online training method [13], which is implemented in [12] and circumvents the problem of out of memory errors by loading a subset of documents into memory. Therefore, while the performance of the RFs was stable, further fine-tuning of the TMs would be necessary to find the optimal features for classifying.

The test run performance shows that a considerable workload reduction (WSS) can be achieved for both 100% and 95% recall of relevant documents. When considering the WSS at 100% recall our method has an acceptable performance (>10% decrease in workload) in 18 out of 30 reviews. At 95% recall

Table 1. Classification results for the test set, when WSS at 95% is larger than at 100% recall. *Topic ID*: unique study identifier; *# Docs*: number of documents in the study; *# Relevant*: number of relevant (i.e., included) documents; *# Found*: number of relevant documents correctly classified; *Last Relevant*: position of the last relevant document in the output of the classifier; *WSS 100 and WSS 95*: WSS at respectively 100% and 95% recall; *AP*: Area under the precision-recall curve; *Norm. AP*: AP normalised by the optimal area; *Reliability*: see [8].

Classifier results, WSS 95 > WSS 100

Topic ID	# Docs	# Relevant	# Found	Last Relevant	WSS 100	WSS 95	Norm. AP	AP	Reliability
CD009551	1911	46	1911	863	0.548	0.744	0.927	0.216	0.469
CD008782	10507	45	10507	3358	0.680	0.741	0.926	0.059	0.476
CD010276	5495	54	5494	4060	0.261	0.567	0.888	0.068	0.422
CD010783	10905	30	10905	4991	0.542	0.551	0.823	0.013	0.592
CD009135	791	77	791	781	0.013	0.462	0.830	0.281	0.319
CD009925	6531	460	6529	6379	0.023	0.440	0.880	0.335	0.032
CD009519	5971	104	5970	4801	0.196	0.382	0.779	0.049	0.240
CD008803	5220	99	5219	4267	0.183	0.374	0.784	0.064	0.252
CD009579	6455	138	6455	6032	0.066	0.297	0.807	0.080	0.177
CD009372	2248	25	2248	1868	0.169	0.290	0.731	0.029	0.640
CD008081	970	26	970	706	0.272	0.278	0.705	0.071	0.630
CD011145	10872	202	10872	9689	0.109	0.247	0.784	0.073	0.110
CD010339	12807	114	12801	12115	0.054	0.204	0.747	0.028	0.218
CD009185	1615	92	1615	1425	0.118	0.182	0.650	0.088	0.271
CD009647	2785	56	2785	2764	0.008	0.081	0.586	0.029	0.411
CD010772	316	47	316	316	0.000	0.061	0.670	0.234	0.463
CD010653	8002	45	8001	7553	0.056	0.060	0.570	0.009	0.476
CD010023	981	52	981	972	0.009	0.018	0.748	0.238	0.433
ALL	117562	1857	117548	2913	0.249	0.333	0.761	0.129	0.544

Table 2. Classification results for the test set, when WSS at 100% is larger than at 95% recall, see Table 1.

Classifier results, WSS 100 > WSS 95

Topic ID	# Docs	# Relevant	# Found	Last Relevant	WSS 100	WSS 95	Norm. AP	AP	Reliability
CD010633	1573	4	1573	326	0.793	0.743	0.848	0.010	0.925
CD010775	241	11	241	75	0.689	0.726	0.923	0.385	0.812
CD010386	626	2	625	198	0.684	0.634	0.842	0.172	0.958
CD012019	10317	3	10317	5861	0.432	0.382	0.590	0.001	0.943
CD010860	94	7	94	54	0.426	0.376	0.724	0.160	0.873
CD008760	64	12	64	42	0.344	0.544	0.869	0.518	0.797
CD009786	2065	10	2065	1379	0.332	0.282	0.842	0.036	0.826
CD010173	5495	23	5494	3885	0.293	0.271	0.747	0.010	0.661
CD010705	114	23	114	105	0.079	0.046	0.586	0.220	0.661
CD010896	169	6	169	162	0.041	0.000	0.667	0.098	0.890
CD010542	348	20	348	340	0.023	0.010	0.638	0.248	0.694
CD007431	2074	24	2074	2030	0.021	0.000	0.710	0.039	0.650

this number increases to 22 out of 30 reviews. The classifier has a good performance (>50% decrease in workload) for respectively 6 and 8 reviews out of 30 (at 100% and 95% recall).

WSS varies wildly among the various DTA studies, as shown in Tables 1 and 2. There can be multiple reasons, one of which being the similarity of documents within a single DTA study. When topics of documents are relatively similar to each other, the classifier's score assigned to each document will be less distinctive. This may result in relevant documents being far apart in the ranking, thereby introducing more false positives. Another reason is that there could be a large difference between the topics of the documents. When the topics in relevant documents from a certain DTA study do not line up with the topics found in the DTA studies used for training, the classifier cannot make the distinction between relevant and non-relevant documents.

TM was chosen in our method because it identifies topics that are shared between documents. Therefore, it can be employed to find similarities between documents. However, this may also assist in building better search queries. For example, by finding the variable importance of the RF (using the `varImp` function of the `caret` package), the most important topics can be identified which distinguish between inclusion and exclusion in DTA reviews. Exploring and interpreting these topics could further specify the search query by suggesting search terms, either to include or exclude publications.

Finally, TM and RF can be employed in an unsupervised manner which relieves the reviewers from the task of providing training data to the method. The future of automation will likely rely on a compound method consisting of various classification techniques. We think the method proposed in this study contributes to systematic review automation by making an initial ordering of documents. While documents are being read and included or excluded an online method can further refine the reading order of documents.

4.1 Related research

Both Bekhuis et al. and Mo et al. [4, 19] report on the use of TM as a feature in predicting systematic review inclusion. In both cases the systematic reviews are not specifically DTA related.

Bekhuis et al. reports that classification performance outcomes for DTA reviews are better when compared to non-DTA reviews. This is likely due to the fact that DTA reviews focus on a very specific topic which is easier to capture in features. From the results of Bekhuis et al. it is apparent that while recall is relatively high for classifiers based on TM features the precision is often lacking. This observation can also be seen in the F1-measure presented in this paper (see Figure 1). Therefore, finding a feature which increases the precision of the classification method will massively affect performance measures such as F1 and will also drop workload (i.e., documents to read).

Mo et al. compares methods using either bag-of-words or TM features. They report that TM yields a better recall which is an highly important metric when considering systematic reviews where reviewers want to find all relevant documents.

It is difficult to compare the employed methods directly because the experiment designs and reported performance measures vary. This is one of the difficulties systematic reviewers encounter when they consider various classification systems, which is also reported in [20]. The performance measures reported in this paper are standardised according to the CLEF eHealth lab, which should contribute towards better understanding of classification methods.

Acknowledgements

This work was carried out on the High Performance Computing Cloud resources of the Dutch national e-infrastructure with the support of SURF Foundation. Furthermore, we would like to thank P.D. Moerland, A.H. Zwinderman, and M.M.G. Leeflang for their contributions and advice.

References

[1] A. J. van Altena. *Article Miner.* 2017. URL: `https://github.com/AMCeScience/article-miner`.

[2] A. J. van Altena. *R-CLEF.* 2017. URL: `https://github.com/Flythe/R-CLEF`.

[3] A. J. van Altena, P. D. Moerland, A. H. Zwinderman and S. D. Olabarriaga. ‘Understanding big data themes from scientific biomedical literature through topic modeling’. In: *Journal of Big Data* 3.1 (2016), p. 23.

[4] T. Bekhuis, E. Tseytlin, K. J. Mitchell and D. Demner-Fushman. ‘Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence’. In: *PloS one* 9.1 (2014), e86277.

[5] D. M. Blei. ‘Probabilistic topic models’. In: *Communications of the ACM* 55.4 (2012), pp. 77–84.

[6] D. M. Blei, A. Y. Ng and M. I. Jordan. ‘Latent dirichlet allocation’. In: *the Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[7] J. Chuang, S. Gupta, C. Manning and J. Heer. ‘Topic model diagnostics: Assessing domain relevance via topical alignment’. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013, pp. 612–620.

[8] G. V. Cormack and M. R. Grossman. ‘Engineering quality and reliability in technology-assisted review’. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* ACM. 2016, pp. 75–84.

[9] I. Feinerer, K. Hornik and D. Meyer. ‘Text mining infrastructure in R’. In: *Journal of Statistical Software* 25.5 (2008), pp. 1–54. URL: `http://www.jstatsoft.org/v25/i05/`.

[10] L. Goeuriot et al. 'CLEF 2017 eHealth Evaluation Lab Overview'. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*. Lecture Notes in Computer Science. Springer, 2017.

[11] B. Grün and K. Hornik. 'topicmodels: An R package for fitting topic models'. In: *Journal of Statistical Software* 40.13 (2011), pp. 1–30. URL: `http://www.jstatsoft.org/v40/i13/`.

[12] M. Hoffman. *onlineldavb*. 2017. URL: `https://github.com/blei-lab/onlineldavb`.

[13] M. Hoffman, F. R. Bach and D. M. Blei. 'Online learning for latent dirichlet allocation'. In: *advances in neural information processing systems*. 2010, pp. 856–864.

[14] M. K. C. from Jed Wing et al. *caret: Classification and Regression Training*. R package version 6.0-71. 2016.

[15] E. Kanoulas, D. Li, L. Azzopardi and R. Spijker. 'Overview of the CLEF Technologically Assisted Reviews in Empirical Medicine'. In: *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017*. CEUR Workshop Proceedings. CEUR-WS.org, 2017.

[16] M. Leeflang et al. 'The diagnostic accuracy of serological tests for Lyme borreliosis in Europe: a systematic review and meta-analysis'. In: *BMC infectious diseases* 16.1 (2016), p. 140.

[17] D. D. Lewis, Y. Yang, T. G. Rose and F. Li. -. URL: `http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop` (visited on 20/11/2015).

[18] D. D. Lewis, Y. Yang, T. G. Rose and F. Li. 'RCV1: A New Benchmark Collection for Text Categorization Research'. In: *J. Mach. Learn. Res.* 5 (Dec. 2004), pp. 361–397. ISSN: 1532-4435. URL: `http://dl.acm.org/citation.cfm?id=1005332.1005345`.

[19] Y. Mo, G. Kontonatsios and S. Ananiadou. 'Supporting systematic reviews using LDA-based document representations'. In: *Systematic reviews* 4.1 (2015), p. 172.

[20] A. OMara-Eves, J. Thomas, J. McNaught, M. Miwa and S. Ananiadou. 'Using text mining for study identification in systematic reviews: a systematic review of current approaches'. In: *Systematic reviews* 4.1 (2015), p. 5.

[21] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: `https://www.R-project.org/`.

[22] G. Salton. 'The SMART retrieval systemexperiments in automatic document processing'. In: (1971).

[23] M. Steyvers and T. Griffiths. 'Probabilistic topic models'. In: *Handbook of Latent Semantic Analysis* 427.7 (2007), pp. 424–440.