LIMSI@CLEF eHealth 2017 Task 2: Logistic Regression for Automatic Article Ranking

Christopher Norman[1 2], Mariska Leeflang[2], and Aurélie Névéol[1]

[1] LIMSI, CNRS, Université Paris Saclay, F-91405 Orsay
`firstname.lastname@limsi.fr`
[2] Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
`m.m.leeflang@uva.nl`

Abstract. This paper describes the participation of the LIMSI-MIROR team at CLEF eHealth 2017, task 2. The task addresses the automatic ranking of articles in order to assist with the screening process of Diagnostic Test Accuracy (DTA) Systematic Reviews. We used a logistic regression classifier and handled class imbalance using a combination of class reweighting and undersampling. We also experimented with two strategies for relevance feedback. Our best run obtained an overall Average Precision of 0.179 and Work Saved over Sampling @95% Recall of 0.650. This run uses stochastic gradient descent for training but no feature selection or relevance feedback. We observe high performance variation within the queries in the test set. Nonetheless, our results suggest that automatic assistance is promising for ranking the DTA literature as it could reduce the screening workload for review writer by 65% on average.

Keywords: Evidence Based Medicine, Information Storage and Retrieval, Review Literature as Topic, Supervised Machine Learning

1 Introduction

Systematic reviews seek to gather all available published evidence for a given topic and provide an informed analysis of the results. This work constitutes some of the strongest forms of scientific evidence. Systematic reviews are an integral part of evidence based medicine in particular, and serve a key role in informing and guiding public and institutional decision-making. Systematic reviews for Diagnostic Test Accuracy (DTA) studies have been shown particularly challenging compared to other types of reviews because of the difficulty in defining search strategies offering adequate levels of sensitivity and specificity [8]. For this reason, there is a need to particularly investigate automation strategies to assist DTA systematic review writers in the time-consuming screening process.

Methods for automating the screening process in systematic reviews have been actively researched over the years [6], with promising results obtained using a range of machine learning methods. However, previous work has not addressed DTA studies.

This paper describes the work underlying our participation in the CLEF 2017 eHealth Task 2 [10, 4]. This work is part of an ongoing effort on providing automatic assistance for the screening process in systematic reviews addressing a variety of topics, including DTA studies.

The remainder of this paper is organized as follows; Section 2 presents the datasets used for system development. Section 3 provides an overview of our system and describes each component. Finally, section 4 reports our results and section 5 provides an analysis of our methods and participation in the task.

2 Datasets

The task relied on a corpus comprising 50 DTA systematic review topics associated with the full list of articles retrieved by an expert query and assessed for inclusion based on title and abstract or full text. The corpus was split into a development dataset comprising 20 topics and a test set comprising the remaining 30 topics. Our classifier was trained on the development dataset and evaluated on the test dataset. We have also used a dataset of systematic reviews on drug class efficacy due to Cohen et al. [1] to develop the methods applied in this task. Several groups have been using this dataset in the past [1, 5], which gives us a way to compare our results with previous work, although we can of course only do by using the same evaluation metrics and training modes as previous work.

For both the CLEF and Cohen datasets we know the inclusion decisions based on the abstracts, as well as the inclusion decisions based on the full text. We thus have two definitions of positive examples, depending on whether we use the abstract decisions or full text decisions as the gold standard.

We use a tripartite labeling to reflect this:

- **No (N)** is the set of articles that were excluded based on the abstract
- **Maybe (M)** is the set of articles that were preliminarily included based on the abstract, but later excluded based on the full text
- **Yes (Y)** is the set of articles that were included based on both the abstract and the full text, and later used in the meta-analysis

Table 1 shows a breakdown of the distribution of examples for each class the CLEF and Cohen datasets used in our work.

Following the work of Cohen et al.[2], we also distinguish between two modes of training:

- **Intertopic** training uses articles from a different topic (systematic review) for training
- **Intratopic** training uses articles from the current topic (systematic review) for training

		Absolute number			Relative number		
Dataset	Topic	Y	M	N	Y	M	N
Cohen	CalciumChannelBlockers	100	180	938	8.21%	14.78%	77.01%
	ACEInhibitors	41	142	2361	1.61%	5.58%	92.81%
	BetaBlockers	42	260	1770	2.03%	12.55%	85.42%
	Opiods	15	33	1867	0.78%	1.72%	97.49%
	OralHypoglycemics	136	3	364	27.04%	0.60%	72.37%
	Statins	85	88	3292	2.45%	2.54%	95.01%
	SkeletalMuscleRelaxants	9	25	1609	0.55%	1.52%	97.93%
	Antihistamines	16	76	218	5.16%	24.52%	70.32%
	ProtonPumpInhibitors	51	187	1095	3.83%	14.03%	82.15%
	Triptans	24	194	453	3.58%	28.91%	67.51%
	NSAIDS	41	47	305	10.43%	11.96%	77.61%
	ADHD	20	64	767	2.35%	7.52%	90.13%
	AtypicalAntipsychotics	146	218	756	13.04%	19.46%	67.50%
	UrinaryIncontinence	40	38	249	12.23%	11.63%	77.61%
	Estrogens	80	0	288	21.74%	0.00%	78.26%
	Total	846	1555	16333	4.52%	8.30%	87.18%
CLEF (train)	CD007394	47	48	2450	1.85%	1.89%	96.27%
	CD007427	17	106	1398	1.12%	6.97%	91.91%
	CD008054	41	233	2940	1.28%	7.25%	91.47%
	CD008643	7	4	15065	0.05%	0.03%	99.93%
	CD008686	5	2	3946	0.13%	0.05%	99.82%
	CD008691	20	53	1243	1.52%	4.03%	94.45%
	CD009020	12	150	1422	0.76%	9.47%	89.77%
	CD009323	9	113	3757	0.23%	2.91%	96.85%
	CD009591	41	103	7847	0.51%	1.29%	98.20%
	CD009593	24	54	14844	0.16%	0.36%	99.48%
	CD009944	64	53	1064	5.42%	4.49%	90.09%
	CD010409	41	35	43287	0.09%	0.08%	99.82%
	CD010438	3	36	3211	0.09%	1.11%	98.80%
	CD010632	14	18	1472	0.93%	1.20%	97.87%
	CD010771	1	47	274	0.31%	14.60%	85.09%
	CD011134	49	166	1738	2.51%	8.50%	88.99%
	CD011548	1	108	12591	0.01%	0.85%	99.14%
	CD011549	1	1	12699	0.01%	0.01%	99.98%
	CD011975	60	559	7582	0.73%	6.82%	92.45%
	CD011984	28	426	7738	0.34%	5.20%	94.46%
	Total	485	2315	146568	0.32%	1.55%	98.13%
CLEF (test)	CD007431	47	9	2050	2.23%	0.43%	97.34%
	CD008081	10	16	944	1.03%	1.65%	97.32%
	CD008760	9	3	52	14.06%	4.69%	81.25%
	CD008782	34	11	10460	0.32%	0.10%	99.57%
	CD008803	99	0	5121	1.90%	0.00%	98.10%
	CD009135	19	58	714	2.40%	7.33%	90.27%
	CD009185	23	69	1523	1.42%	4.27%	94.30%
	CD009372	10	15	2223	0.44%	0.67%	98.89%
	CD009519	46	58	5867	0.77%	0.97%	98.26%
	CD009551	16	30	1865	0.84%	1.57%	97.59%
	CD009579	79	59	6317	1.22%	0.91%	97.86%
	CD009647	17	39	2729	0.61%	1.40%	97.99%
	CD009786	6	4	2055	0.29%	0.19%	99.52%
	CD009925	55	405	6071	0.84%	6.20%	92.96%
	CD010023	14	38	929	1.43%	3.87%	84.70%
	CD010173	10	13	5472	0.18%	0.24%	99.58%
	CD010276	24	30	5441	0.44%	0.55%	99.02%
	CD010339	9	105	12689	0.07%	0.82%	99.11%
	CD010386	1	1	623	0.16%	0.16%	99.68%
	CD010542	8	12	328	2.30%	3.45%	94.25%
	CD010633	3	1	1569	0.19%	0.06%	99.75%
	CD010653	0	45	7957	0.00%	0.56%	99.44%
	CD010705	18	5	91	15.79%	4.39%	79.82%
	CD010772	11	36	269	3.48%	11.39%	85.13%
	CD010775	4	7	230	1.66%	2.90%	95.44%
	CD010783	11	19	10875	0.10%	0.17%	99.72%
	CD010860	4	3	87	4.26%	3.19%	92.55%
	CD010896	3	3	163	1.78%	1.78%	96.45%
	CD011145	48	154	10670	0.44%	1.42%	98.14%
	CD012019	1	2	10314	0.01%	0.02%	99.97%
	Total	639	1250	115698	0.54%	1.06%	98.39%
	Total (train + test)	1124	3565	262266	0.42%	1.34%	98.24%

Table 1: The distribution of class labels in each dataset.

3 Method

We first give an overview of our system, which relies on logistic regression, in section 3.1. Further details about the system are given in sections 3.2–3.5, including features, strategies to handle class imbalance and implement relevance feedback.

3.1 Overview

We have tried the following two classifiers:

- **Classifier 1** uses logistic regression trained using stochastic gradient descent on all features
- **Classifier 2** uses standard logistic regression trained using standard methods on a subset of the features, and with additional preprocessing to improve the throughput

We have tried three approaches to relevance feedback:

- **no relevance feedback**
- **abrupt** uses intertopic ranking until a sufficient number of relevant and non-relevant articles have been identified, and then switches to using intratopic ranking based on the identified articles
- **gradual** initially uses intertopic ranking, and gradually improves the model using both Y and M identified through relevance feedback

In total, we have submitted the following four runs to the CLEF evaluation:

- **no_AF_full** uses classifier 1 with no relevance feedback
- **no_AF** uses classifier 2 with no relevance feedback
- **abrupt** uses classifier 2 with `abrupt` relevance feedback
- **gradual** uses classifier 2 with `gradual` relevance feedback

3.2 Classification approach

We are currently using two classification systems. Both use logistic regression but differ in how the model is optimized and the amounts and types of pre- and postprocessing that is performed. Both methods use implementations provided by sklearn [7].

Our first method, which is used in `no_AF_full` tends to work well for intertopic classification on previous datasets (see table 3), presumably because it generalizes better. This system uses logistic regression trained using stochastic gradient descent. The only preprocessing done is the normalization of numerals.

Our second method, which is used in `no_AF`, `abrupt`, and `gradual` uses standard methods for training (liblinear). This version tends to work well on intratopic classification on previous datasets (see table 3), but does not scale as well with data volume. We therefore need to do additional preprocessing to reduce the number of features and keep running times down. We thus remove features with variance less than a predefined threshold, we only consider n-grams with high mutual information with the target class in the training set, we normalize numerals, and we extract the principal components from the resulting data.

Principal component analysis tends to reduce overfitting in our experiments, and it also drastically reduces the time it takes to train and apply the classifier, which is mostly important when we use relevance feedback.

3.3 Features

For all classifiers we extract n-grams ($n \leq 5$) from the titles and abstracts. We also extract publication type, journal names, author assigned keywords, MeSH terms, and backward references, where these are available. The backward references are only available for references pointing to articles available in Pubmed Central, and this feature set is therefore fairly sparse.

Not all feature sets are useful for identifying DTA studies, but the current model has been constructed such that irrelevant features should not adversely effect the performance. All the feature sets have been shown to be useful on some domain. For instance MeSH terms might not be useful for DTA studies, but we have previously found them to be useful in identifying topics related to drug efficacy.

3.4 Class imbalance

Class imbalance can be handled using undersampling, or by class reweighting. We are currently using a combination of both these approaches.

Class weights We set the weight for the positive class to 80 for the initial intertopic classifier. We have determined this to be a reasonable weight experimentally using the Cohen dataset.

For the `gradual` relevance feedback we also attached higher weights to the intratopic training examples identified through relevance feedback.

Undersampling In order to reduce the effects of the class imbalance we undersample the training set to include an equal number of Y, M, and N. However, by doing so we end up with only around 1500 training samples. PCA yields at most the same number of principal components as we have input samples, and 1500 is generally too few principal components to build an accurate classifier. For the second model we therefore perform undersampling in two steps; We first select a maximum of 500 Y, 1000 M, and 1500 N that we feed into the feature extraction pipeline, which thus determines the number of features in our model. We then select a smaller undersample to use for training.

We take a new undersample in each iteration of relevance feedback.

3.5 Relevance Feedback

We use two schemes for relevance feedback. For both schemes we retrain the classifier each time we retrieve relevance feedback.

abrupt trains an initial intertopic classifier on the training dataset and ranks the test dataset in descending order of confidence. The system then iteratively asks for feedback for the top ranked results. When enough positive and negative examples have been identified, the system switches to using a classifier trained on the examples identified from relevance feedback. Additional examples are added to the intratopic classifier as they are discovered.

The idea behind this system is that on some topics in Cohen we can train highly performing intratopic classifiers using very small amounts of data, and we have observed that even trained on small amounts of data these sometimes outperform intertopic classifiers by a large margin. In these cases it might make sense to switch to intratopic classification as soon as we can.

We set the minimum number of positive examples to 4, and the minimum number of negative examples to 10.

gradual trains an initial intertopic classifier using the training set and ranks the test set in descending order of confidence. The system then iteratively asks for feedback for the top ranked result. Articles queried for relevance feedback are then added to the model as they are queried, but with higher weights than the intertopic examples. The model thus starts out as an intertopic classifier, but gradually turns into an intratopic classifier as more targeted data is added to the model. Since the intratopic examples identified through relevance feedback are given higher weights, these will eventually drown out the original classifier, provided enough examples exist to be discovered.

Besides using Y and N, we also use intratopic M as positive examples, with lower weights than intratopic Y, but higher than intertopic Y. The reasoning behind this is that we often encounter M earlier than Y, and in greater numbers, in particular on topics with very few Y. We have observed on other datasets that we can sometimes improve performance by using both Y and M as positive examples, when the number of Y is very low.

After the number of Y found is larger than 40, we stop using M as positive examples.

Reasonable parameter settings were identified experimentally on the Cohen dataset.

3.6 Use of the CLEF development dataset

We do not split the training data into separate training and validation splits, since we do not have the necessary number of Y to do this without hurting the performance of the classifier. We do however use a small set of samples that overlaps with the training set for validation. The performance we observe on this validation suffers from severe overfitting, but we can observe when the model fails to build a classifier on the current undersample. In such cases we can observe an AUROC < 0.5 even on the training set. In these cases we simply discard the classifier and try again with a new undersample. We observe that this improves performance dramatically when we have a very small amount of training data (approximately four or less positive examples).

4 Results

We present a comparison with previous work on the Cohen dataset for WSS@95 in table 2 and for AUC in table 3. Results from previous literature are taken from Khabsa et al. [5], and Cohen et al. [2]. Exact intertopic AUC scores are not explicitly reported by Cohen et al. and have instead been extracted from Figure 1 in their paper The majority of these results, with the exception of one result by Cohenet al. [2] use intratopic classification.

We present our results on the CLEF dataset for average precision in table 4, normalized average precision in table 5, WSS@95 in table 6, and in aggregate in table 7. The results in these tables correspond to those submitted as official runs. For comparison, we also calculate a baseline by evaluating each metric on the data ordered randomly. This has been repeated 1000 times and we report the average and standard deviation.

We also report the mean, standard deviation, minimum and maximum WSS@95 and AUC over ten runs for a selection of topics in the CLEF dataset in table 8.

5 Discussion

5.1 Datasets

One of the topics in the CLEF dataset, CD010653, has no Y. While we can still calculate performance scores relative to M, this topic might arguably have been omitted from the test data. One of the topics, CD008803, similarly has no M. This also happens to be the topic with the largest number of Y.

As a general tendency, we can observe that the relative number of Y / M / N in the CLEF dataset varies dramatically across topics. At the one end we have one topic consisting of 14.06% Y (CD008760), and one topic consisting of 15.79% Y (CD010705). At the other end we have three topics with a mere 0.01% Y (CD011548, CD011549, and CD012019). Most topics in the CLEF dataset have a very small number of Y compared to Cohen, both in terms of relative and absolute numbers. Several topics have a large number of M however (CD007427, CD008054, CD009020, CD009323, CD009591, 011134, CD011548, CD0011975, CD011984, CD009925, CD10339, CD011145). Curiously, more topics in the training set have a large number of M than in the test set, despite this comprising a smaller number of topics.

Topic	no_AF_full	no_AF	VP	CNB	RF
CalciumChannelBlockers	.398	**.408**	<.100	.234	.287
ACEInhibitors	**.629**	.517	.318	.523	.447
BetaBlockers	**.511**	.427	.284	.367	.361
Opiods	.590	**.641**	<.190	.554	.455
OralHypoglycemics	.111	**.153**	<.050	.080	.074
Statins	.436	**.573**	.242	.315	.400
SkeletalMuscleRelaxants	**.429**	.179	-.050	.265	.371
Antihistamines	.149	**.157**	.080	.148	.030
ProtonPumpInhibitors	.307	**.320**	<.180	.229	.288
Triptans	.303	**.312**	.030	.279	**.312**
NSAIDS	.537	**.600**	.352	.528	.404
ADHD	.616	.530	**.668**	.622	.447
AtypicalAntipsychotics	.210	**.234**	.140	.206	.199
UrinaryIncontinence	**.422**	.365	.260	.290	.411
Estrogens	.292	**.475**	.140	.375	.180

Table 2: Comparison in terms of WSS@95% with previous literature using Voting Perceptrons, Complement Naive Bayes, and Random Forests, as reported by Khabsa et al. [5]. We here only have state of the art metrics for the intratopic case.

	Intertopic			RF	Intratopic			
Topic	no_AF_full	no_AF	Cohen	gradual	no_AF_full	no_AF	Cohen	Khabsa
CalciumChannelBlockers	.759	**.773**	.712	.862	.825	.868	**.873**	.870
ACEInhibitors	**.817**	.782	.806	.899	.917	.925	.946	**.951**
BetaBlockers	**.837**	.832	.801	.860	.863	.871	.891	**.893**
Opiods	.885	**.902**	.856	.936	.905	.893	.897	**.913**
OralHypoglycemics	**.657**	.581	.573	.753	.568	.768	**.781**	.734
Statins	**.826**	.798	.773	.797	.873	**.922**	.900	.915
SkeletalMuscleRelaxants	.826	.823	**.836**	.812	.740	.527	.738	**.794**
Antihistamines	**.652**	.600	.620	.752	.650	.655	**.722**	.701
ProtonPumpInhibitors	**.823**	.790	.793	.886	.826	.860	.860	**.880**
Triptans	.819	.796	**.823**	.804	.792	.808	**.909**	.894
NSAIDS	**.912**	.828	.899	.922	.861	.935	**.951**	.933
ADHD	.591	**.606**	.469	.740	.908	.897	.924	**.951**
AtypicalAntipsychotics	**.759**	.645	.653	.855	.779	.803	**.835**	.818
UrinaryIncontinence	**.887**	.875	.851	.888	.784	.885	**.890**	.862
Estrogens	**.693**	.649	.588	.879	.689	**.912**	.887	.840

Table 3: Comparison in terms of AUC with previous literature using Support Vector Machines (Cohen) and Random Forests (Khabsa), as reported by Khabsa et al. [5], and Cohen et al. [2]. Exact intertopic AUC scores are not explicitly reported by Cohen et al. and have instead been extracted from Figure 1 in their paper.

	Y||MN					YM||N				
	w/o RF		w/ RF		baseline	w/o RF		w/ RF		baseline
Topic	no_AF_full	no_AF	abrupt	gradual		no_AF_full	no_AF	abrupt	gradual	
CD007431	0.047	0.016	0.026	0.013	0.010 ± 0.005	0.065	0.026	0.044	0.019	0.015 ± 0.005
CD008081	0.146	0.099	0.087	0.046	0.016 ± 0.010	0.114	0.097	0.060	0.041	0.032 ± 0.009
CD008760	0.790	0.516	0.569	0.835	0.169 ± 0.052	0.886	0.644	0.734	0.807	0.210 ± 0.050
CD008782	0.057	0.231	0.032	0.042	0.004 ± 0.002	0.060	0.242	0.040	0.050	0.005 ± 0.002
CD008803	0.181	0.131	0.147	0.120	0.020 ± 0.003	0.181	0.131	0.147	0.120	0.020 ± 0.002
CD009135	0.382	0.217	0.149	0.324	0.030 ± 0.009	0.485	0.349	0.266	0.493	0.102 ± 0.012
CD009185	0.041	0.049	0.080	0.027	0.018 ± 0.006	0.139	0.096	0.135	0.085	0.060 ± 0.007
CD009372	0.122	0.189	0.078	0.081	0.007 ± 0.006	0.080	0.107	0.056	0.060	0.014 ± 0.004
CD009519	0.031	0.022	0.034	0.020	0.009 ± 0.002	0.059	0.051	0.067	0.038	0.019 ± 0.002
CD009551	0.199	0.140	0.222	0.157	0.011 ± 0.006	0.287	0.259	0.259	0.284	0.027 ± 0.006
CD009579	0.172	0.105	0.257	0.286	0.013 ± 0.002	0.253	0.157	0.259	0.286	0.022 ± 0.002
CD009647	0.038	0.024	0.019	0.026	0.008 ± 0.004	0.052	0.040	0.034	0.068	0.022 ± 0.004
CD009786	0.028	0.024	0.012	0.008	0.006 ± 0.007	0.034	0.055	0.190	0.014	0.008 ± 0.007
CD009925	0.114	0.080	0.044	0.077	0.010 ± 0.002	0.334	0.168	0.151	0.285	0.071 ± 0.003
CD010023	0.089	0.058	0.051	0.085	0.020 ± 0.008	0.303	0.273	0.168	0.222	0.058 ± 0.009
CD010173	0.014	0.008	0.010	0.001	0.003 ± 0.004	0.025	0.014	0.015	0.003	0.006 ± 0.003
CD010276	0.072	0.055	0.032	0.003	0.006 ± 0.003	0.108	0.100	0.057	0.007	0.011 ± 0.003
CD010339	0.018	0.067	0.021	0.035	0.001 ± 0.002	0.043	0.046	0.020	0.040	0.010 ± 0.001
CD010386	0.053	0.083	0.091	0.167	0.009 ± 0.023	0.031	0.044	0.050	0.085	0.010 ± 0.017
CD010542	0.082	0.145	0.190	0.038	0.036 ± 0.021	0.131	0.158	0.188	0.110	0.068 ± 0.018
CD010633	0.015	0.010	0.010	0.002	0.006 ± 0.014	0.071	0.028	0.023	0.003	0.006 ± 0.010
CD010653	-	-	-	-	-	0.011	0.016	0.012	0.005	0.006 ± 0.002
CD010705	0.240	0.220	0.389	0.312	0.174 ± 0.037	0.250	0.247	0.444	0.380	0.214 ± 0.036
CD010772	0.117	0.035	0.069	0.086	0.048 ± 0.020	0.211	0.155	0.214	0.343	0.158 ± 0.021
CD010775	0.187	0.101	0.170	0.069	0.034 ± 0.031	0.623	0.462	0.433	0.258	0.062 ± 0.023
CD010783	0.071	0.037	0.020	0.009	0.002 ± 0.003	0.044	0.103	0.051	0.026	0.004 ± 0.002
CD010860	0.188	0.139	0.135	0.032	0.070 ± 0.042	0.168	0.126	0.134	0.047	0.104 ± 0.042
CD010896	0.347	0.093	0.248	0.239	0.037 ± 0.033	0.213	0.100	0.154	0.163	0.054 ± 0.028
CD011145	0.027	0.009	0.023	0.011	0.005 ± 0.001	0.108	0.044	0.058	0.038	0.019 ± 0.002
CD012019	0.003	0.002	0.003	0.002	0.001 ± 0.008	0.003	0.002	0.002	0.001	0.001 ± 0.001
Average	0.179	0.145	0.143	0.146	0.027 ± 0.003	0.133	0.100	0.111	0.109	0.047 ± 0.003

Table 4: Average precision score for all topics in the CLEF dataset.

	Y||MN					YM||N				
	w/o RF		w/ RF		baseline	w/o RF		w/ RF		baseline
Topic	no_AF_full	no_AF	abrupt	gradual		no_AF_full	no_AF	abrupt	gradual	
CD007431	0.825	0.673	0.762	0.704	0.503 ± 0.074	0.773	0.684	0.769	0.700	0.501 ± 0.060
CD008081	0.907	0.872	0.801	0.695	0.504 ± 0.091	0.801	0.751	0.653	0.603	0.506 ± 0.057
CD008760	0.963	0.895	0.933	0.955	0.518 ± 0.098	0.976	0.917	0.927	0.920	0.536 ± 0.082
CD008782	0.942	0.983	0.351	0.876	0.501 ± 0.050	0.939	0.977	0.360	0.888	0.500 ± 0.044
CD008803	0.944	0.889	0.898	0.915	0.505 ± 0.029	0.944	0.889	0.898	0.915	0.504 ± 0.028
CD009135	0.962	0.875	0.856	0.959	0.503 ± 0.068	0.875	0.841	0.744	0.897	0.524 ± 0.033
CD009185	0.790	0.676	0.677	0.744	0.500 ± 0.060	0.779	0.603	0.618	0.687	0.514 ± 0.031
CD009372	0.947	0.970	0.817	0.853	0.500 ± 0.092	0.815	0.839	0.714	0.749	0.501 ± 0.059
CD009519	0.865	0.802	0.862	0.807	0.498 ± 0.041	0.851	0.792	0.838	0.766	0.503 ± 0.028
CD009551	0.960	0.961	0.892	0.946	0.503 ± 0.072	0.945	0.953	0.862	0.930	0.506 ± 0.043
CD009579	0.902	0.784	0.871	0.913	0.505 ± 0.032	0.902	0.827	0.821	0.875	0.505 ± 0.025
CD009647	0.747	0.774	0.674	0.830	0.500 ± 0.071	0.720	0.706	0.628	0.835	0.504 ± 0.038
CD009786	0.918	0.854	0.756	0.736	0.503 ± 0.118	0.895	0.858	0.743	0.762	0.501 ± 0.093
CD009925	0.947	0.822	0.695	0.839	0.502 ± 0.038	0.883	0.674	0.616	0.753	0.518 ± 0.013
CD010023	0.890	0.806	0.780	0.879	0.504 ± 0.077	0.872	0.864	0.780	0.889	0.513 ± 0.039
CD010173	0.929	0.805	0.882	0.379	0.495 ± 0.091	0.901	0.770	0.766	0.383	0.502 ± 0.062
CD010276	0.956	0.938	0.882	0.279	0.503 ± 0.057	0.940	0.904	0.801	0.345	0.503 ± 0.038
CD010339	0.887	0.860	0.777	0.873	0.506 ± 0.094	0.816	0.760	0.580	0.764	0.504 ± 0.028
CD010386	0.971	0.982	0.984	0.992	0.512 ± 0.290	0.820	0.686	0.804	0.531	0.510 ± 0.202
CD010542	0.794	0.748	0.793	0.650	0.503 ± 0.099	0.692	0.606	0.696	0.676	0.512 ± 0.066
CD010633	0.846	0.873	0.830	0.315	0.503 ± 0.171	0.884	0.903	0.869	0.414	0.500 ± 0.145
CD010653	-	-	-	-	-	0.688	0.753	0.721	0.497	0.500 ± 0.043
CD010705	0.713	0.621	0.867	0.802	0.531 ± 0.068	0.655	0.611	0.868	0.817	0.544 ± 0.059
CD010772	0.805	0.455	0.581	0.679	0.504 ± 0.089	0.661	0.485	0.551	0.719	0.537 ± 0.041
CD010775	0.956	0.893	0.601	0.847	0.496 ± 0.146	0.982	0.947	0.762	0.914	0.508 ± 0.084
CD010783	0.935	0.935	0.926	0.916	0.501 ± 0.089	0.918	0.941	0.848	0.870	0.502 ± 0.052
CD010860	0.832	0.840	0.837	0.217	0.498 ± 0.141	0.697	0.656	0.667	0.152	0.514 ± 0.111
CD010896	0.855	0.648	0.721	0.904	0.501 ± 0.168	0.756	0.691	0.605	0.733	0.503 ± 0.115
CD011145	0.860	0.736	0.792	0.750	0.500 ± 0.042	0.868	0.751	0.724	0.723	0.504 ± 0.020
CD012019	0.964	0.960	0.962	0.946	0.505 ± 0.286	0.917	0.767	0.806	0.542	0.497 ± 0.160
Average	0.890	0.825	0.795	0.766	0.504 ± 0.022	0.839	0.780	0.735	0.708	0.509 ± 0.014

Table 5: Normalized average precision score for all topics in the CLEF dataset.

	Y||MN					YM||N				
	w/o RF		w/ RF		baseline	w/o RF		w/ RF		baseline
Topic	no_AF_full	no_AF	abrupt	gradual		no_AF_full	no_AF	abrupt	gradual	
CD007431	0.621	0.298	0.356	0.415	0.071 ± 0.078	0.297	0.079	0.356	0.323	0.030 ± 0.052
CD008081	0.452	0.260	0.056	0.391	0.042 ± 0.082	0.430	0.138	0.283	0.365	0.023 ± 0.048
CD008760	0.731	0.512	0.591	0.575	0.023 ± 0.075	0.731	0.575	0.591	0.575	0.075 ± 0.087
CD008782	0.767	0.873	-0.037	0.476	0.036 ± 0.046	0.706	0.857	-0.039	0.476	0.013 ± 0.035
CD008803	0.787	0.584	0.528	0.612	0.009 ± 0.023	0.787	0.584	0.528	0.312	0.009 ± 0.023
CD009135	0.759	0.457	0.739	0.783	0.048 ± 0.067	0.439	0.403	0.035	0.580	0.012 ± 0.026
CD009185	0.377	0.073	0.096	0.500	0.031 ± 0.057	0.377	0.026	0.024	0.114	0.013 ± 0.025
CD009372	0.654	0.844	0.051	0.170	0.040 ± 0.083	0.353	0.461	0.139	0.170	0.025 ± 0.050
CD009519	0.597	0.294	0.442	0.624	0.011 ± 0.034	0.483	0.219	0.336	0.291	0.006 ± 0.022
CD009551	0.856	0.866	0.584	0.834	0.070 ± 0.075	0.757	0.838	0.368	0.667	0.014 ± 0.037
CD009579	0.531	0.153	0.327	0.522	0.012 ± 0.027	0.580	0.275	0.203	0.351	0.008 ± 0.021
CD009647	0.321	0.469	0.124	0.577	0.058 ± 0.073	0.240	0.243	0.028	0.499	0.018 ± 0.033
CD009786	0.799	0.656	0.134	0.248	0.098 ± 0.124	0.621	0.656	0.234	0.248	0.041 ± 0.085
CD009925	0.810	0.346	0.050	0.277	0.022 ± 0.033	0.469	0.0	-0.040	-0.037	0.002 ± 0.010
CD010023	0.714	0.649	0.572	0.693	0.085 ± 0.085	0.492	0.474	0.515	0.662	0.024 ± 0.034
CD010173	0.777	0.476	0.555	0.078	0.039 ± 0.083	0.671	0.271	0.174	0.078	0.034 ± 0.057
CD010276	0.811	0.803	0.486	-0.031	0.031 ± 0.050	0.719	0.511	0.217	-0.031	0.024 ± 0.034
CD010339	0.193	0.114	0.077	0.330	0.052 ± 0.095	0.346	0.192	0.026	0.219	0.011 ± 0.023
CD010386	0.920	0.931	0.932	0.940	0.461 ± 0.289	0.616	0.337	0.571	0.019	0.286 ± 0.237
CD010542	0.464	0.171	0.099	0.191	0.056 ± 0.097	0.232	0.065	0.099	0.191	0.041 ± 0.061
CD010633	0.637	0.741	0.584	0.050	0.203 ± 0.197	0.637	0.741	0.584	0.050	0.143 ± 0.164
CD010653	-	-	-	-	-	0.218	0.272	0.227	0.050	0.014 ± 0.035
CD010705	0.213	0.187	0.625	0.503	0.040 ± 0.064	0.064	0.161	0.564	0.494	0.014 ± 0.048
CD010772	0.532	0.070	0.247	0.153	0.108 ± 0.101	0.077	0.001	-0.041	-0.009	0.006 ± 0.031
CD010775	0.867	0.726	0.170	0.701	0.140 ± 0.163	0.867	0.813	0.174	0.718	0.109 ± 0.098
CD010783	0.856	0.906	0.819	0.842	0.124 ± 0.106	0.701	0.381	0.340	0.072	0.015 ± 0.043
CD010860	0.578	0.695	0.716	0.046	0.134 ± 0.155	0.237	0.067	-0.050	-0.039	0.065 ± 0.106
CD010896	0.517	0.098	0.151	0.684	0.192 ± 0.196	0.518	0.098	0.151	0.051	0.086 ± 0.121
CD011145	0.497	0.342	0.228	0.175	0.011 ± 0.034	0.446	0.327	0.108	0.103	0.004 ± 0.015
CD012019	0.914	0.909	0.912	0.896	0.455 ± 0.286	0.797	0.369	0.534	0.276	0.195 ± 0.190
Average	0.640	0.500	0.390	0.457	0.093 ± 0.023	0.497	0.348	0.241	0.271	0.015 ± 0.016

Table 6: Work saved over sampling at 95% recall for all topics in the CLEF dataset.

	Y||MN					YM||N				
	w/o RF		w/ RF		baseline	w/o RF		w/ RF		baseline
Topic	no_AF_full	no_AF	abrupt	gradual		no_AF_full	no_AF	abrupt	gradual	
WSS@95	0.640	0.500	0.390	0.457	0.093 ± 0.023	0.497	0.348	0.241	0.271	0.045 ± 0.016
WSS@100	0.591	0.420	0.350	0.407	0.112 ± 0.022	0.412	0.261	0.173	0.195	0.056 ± 0.015
last_rel	1678	2263	2619	2384	3393.7 ± 118.1	2250	2993	3414	3406	3749.7 ± 68.8
NCG@10	0.517	0.407	0.357	0.346	0.081 ± 0.010	0.475	0.367	0.316	0.350	0.092 ± 0.006
NCG@20	0.802	0.639	0.644	0.685	0.180 ± 0.015	0.717	0.554	0.518	0.601	0.192 ± 0.008
NCG@30	0.908	0.783	0.753	0.789	0.280 ± 0.018	0.825	0.674	0.609	0.698	0.291 ± 0.010
NCG@40	0.946	0.843	0.814	0.832	0.379 ± 0.020	0.887	0.746	0.678	0.763	0.391 ± 0.011
NCG@50	0.972	0.890	0.842	0.881	0.479 ± 0.020	0.929	0.800	0.727	0.816	0.491 ± 0.011
NCG@60	0.984	0.921	0.886	0.911	0.579 ± 0.020	0.955	0.851	0.789	0.853	0.591 ± 0.011
NCG@70	0.990	0.942	0.911	0.937	0.679 ± 0.018	0.976	0.903	0.834	0.889	0.691 ± 0.011
NCG@80	0.997	0.960	0.939	0.959	0.778 ± 0.016	0.987	0.930	0.878	0.918	0.791 ± 0.007
NCG@90	0.998	0.987	0.965	0.980	0.878 ± 0.013	0.996	0.964	0.920	0.943	0.890 ± 0.007
NCG@100	1.000	0.998	1.000	1.000	0.977 ± 0.006	1.000	0.999	0.998	0.997	0.990 ± 0.002
norm_area	0.890	0.825	0.795	0.766	0.504 ± 0.022	0.839	0.780	0.735	0.708	0.509 ± 0.014
ap	0.133	0.100	0.111	0.109	0.027 ± 0.003	0.179	0.145	0.143	0.146	0.047 ± 0.003

Table 7: Aggregate performance for each ranking metric.

	WSS@95								AUC							
	no_AF_full				no_AF				no_AF_full				no_AF			
Topic	mean	std	min	max	mean	std	min	max	mean	std	min	max	mean	std	min	max
CD008760	.723	.034	.653	.762	.666	.080	.481	.764	.949	.012	.933	.971	.937	.020	.899	.962
CD010386	.899	.012	.883	.921	.932	.006	.923	.942	.949	.012	.933	.971	.982	.006	.973	.992
CD010705	.085	.036	.025	.143	.047	.027	.011	.099	.696	.013	.683	.705	.595	.028	.572	.632
CD012019	.920	.006	.907	.929	.923	.007	.903	.927	.962	.010	.949	.982	.973	.007	.899	.979
CD010339	.250	.085	.084	.415	.438	.139	.265	.607	.884	.013	.029	.864	.903	.039	.798	.923

Table 8: The average, standard deviation, mininum, and maximum WSS@95 and AUC over ten iterations on a subset of the topic in CLEF for our systems `no_AF` and `no_AF_full`.

The number of N also varies wildly, from 52 up to 43287. Compared to the Cohen dataset we also have a smaller minimum number of N, as well as much larger maximum number.

If we compare the training and test sets, the training set contains almost double the absolute number of M, many more N, but fewer Y.

5.2 Performance

While relevance feedback sometimes gives an improvement in performance, relevance feedback often seems to only confuse the system (tables 4–7). This should be contrasted with our experiments on the Cohen dataset, where the same implementation reliably yields an improvement (table 3), and generally yields performance intermediate between intertopic and intratopic classification, as one would expect. There are perhaps better approaches to relevance feedback than ours, which can reliably improve upon the baseline, but it might also be that there is simply little to gain from relevance feedback on several of the topics. Of particular note, we should not expect any improvements by using RF on topics such as CD010386, CD010633, CD010860, CD010896, and CD012019, that have a low absolute number of Y and M. It is also worth pointing out that our `abrupt` scheme requires at least 4 Y before switching to the intratopic model, and any differences between `no_AF` and `abrupt` on these topics can thus only be due to chance.

We can see an improvement on the topic CD010705 when using relevance feedback (tables 4–7). This topics is also the topic with the highest percentage of Y at 15.79%. We do not see any improvement for CD008760, the other topic with a high percentage of Y (14.06%), but this may be due to the initial classifier having much higher performance.

We can observe that `gradual` outperforms `abrupt` on topic CD008760, despite this topic having only 3 M, which is probably too low a numbe for `gradual` to have an advantage. The simplest explanation for this is likely random chance.

It is however easy to see that relevance feedback does not appear to lead to an improvement for our system. For instance `abrupt` outperforms `no_AF` 15 times out of 30, and `gradual` outperforms `no_AF` only 10 times out of 30 (tables 4).

Of course, it seems unlikely for relevance feedback to be useful for those topics where the number of positives is extremely low, even in theory. In particular, if there is only one relevant article, as is the case for CD012019 and CD010386, then relevance feedback cannot really add any value to the classification. Any successful use of relevance feedback on such topics would necessarily have to use the negative examples.

We get better performance for `no_AF_full` than `no_AF`. We have however generally observed that this difference is generally reversed for intratopic classification, which is what we should end up with when we after relevance feedback, but it is possible that we would get better performance if we were to use `no_AF_full` as a base for our relevance feedback experiments, since we would start with a much better initial classifier.

Ordinarily, screeners would be free to choose the order in which they screen each article, and may proceed for instance in alphabetical or chronological order. For the purposes of our baseline, we assume that any such order ordinarily available to screeners would be indistinguishable from random order on average.

5.3 Metrics

Average Precision has been selected as the main metric for this task as it was previously found particularly adapted to evaluate retrieval performance for highly imbalanced datasets [9, 3]. However, these studies rely on common assumptions that we value high precision at the top of the ranking, whereas for systematic review screening we value recall almost exclusively. Of particular note, average precision heavily penalizes rankings where the top few results are non-relevant, even if the ranking manages to place all relevant articles in the upper percentiles of the ranking.

Furthermore, average precision is strongly correlated with the number of positives in the topic, with most of the cases where we achieve ap > 0.2 are for topics with high prevalence. While this is to be expected, it means that average precision makes it difficult to compare performance across topics, since we can see a strong correlation with the prevalence of relevant articles in the topic (tables 1, 4–7). Similarly, Mean Average Precision will likely be dominated by the results on the topics with many relevant articles and a small number of total candidates, i.e. arguably the topics which are the least representative systematic reviews of DTA studies, and where automated methods are likely the least useful.

5.4 Reliability of the Experiments

Our classification method is stochastic, and thus does not produce deterministic results that are always the same every time we run on the same input data. To gauge the reliability of the experiment we repeat it ten times for a subset of the topics and calculate the standard deviations, as well as examine the minimum and maximum values (table 8).

We can generally observe a fairly large variability for topics with a small total number of candidates, such as CD008760 and CD010705, and for topics with a comparably smaller proportion of Y, such as CD010339. When we consider topics with a large number of candidates we can observe a large variability for the CD012019, but small variability for CD010386. We might speculate that small topic size and a small relative number of Y is correlated with larger variability, but it is clear that the variability for some topics is quite large, regardless of the underlying causes and mechanisms. The standard deviation can be as large as .139, which is large enough that it casts doubts about the reliability of the results. Furthermore, the minimum and maximum values are much more skewed towards extreme values than we should expect from the standard deviations were the values normally distributed, suggesting that the distribution is heavy-tailed and skewed towards outliers.

Considering the above, we might suspect that the differences in performance in tables 4–7 are not significant. For instance `abrupt` outperforms `gradual` 17 times out of 30, but we do not know whether this means that `abrupt` is a better method, or if this is simply due to random chance. We might speculate that our gradual implementation works better for the cases where we have a sufficient number of M, but the experiment is ultimately too low-powered to draw conclusions. Future iterations of the campaign could consider whether performance should be computed as an average over multiple runs, in order to get more precise results for stochastic systems such as ours.

We can however see smaller variability in the mean performance across all topics, which might suggest that these are more reliable estimates. However, these give little indication as to how the performance depends on topic composition.

5.5 General Remarks on the Shared Task Model

The Shared Task Model is typically implemented in evaluation campaigns that seek to perform a community-wide technical evaluation of systems addressing a particular task. A Shared Task thus offers an evaluation paradigm that includes: 1/a specific definition of the task and evaluation metrics 2/an implementation through the dissemination of datasets and evaluation tools and 3/the execution of the evaluation in a controlled setting where participants have access to data at the same time and are evaluated blindly by an independent third party. As outlined below, this year the TAR task was not conducted according to the Shared Task Model.

In this iteration of the evaluation campaign, the final set of evaluation metrics was decided only shortly before participants were required to freeze their systems. One of the expected outcomes of evaluation campaigns such as this is indeed the discussion of the relative merits of the various metrics to be used. However, changing the target metric close to the submission deadline means that some participants may have optimized for different metrics than those ultimately used for evaluation.

The gold standard labeled test data was distributed directly to the participants at the begining of the test phase. This is explained by the lack of an assessor through which participants could receive relevance feedback as has been the case in e.g. TREC Total Recall. While common labeled test collections are routinely used for research, this procedure is unusual in a shared task setting where participants are typically asked to process a test dataset while being blind to the gold standard associated with the dataset. This could alternatively have been accomplished in part by requiring the submission of runs without relevance feedback before the distribution of the gold standard labels.

Another feature of the shared task model is the computation of performance metrics for all participants by a common, independent party which ensures that all participations are evaluated using the exact same conditions. This confers a stronger reliability in the comparability and reproducibility of results. At the time of writing, while a common evaluation tool has been released, the performance reported by participants has been self-computed without validation from the task organizers. In addition to result validation, it would also have been useful to receive an indication of the overall performance of the participants prior to the deadline for the submission of the working notes. This would have enabled a discussion about the relative performance of the system that is currently difficult to do without comparing with previous literature using external datasets.

6 Conclusions

Our best system is the one using logistic regression trained using stochastic gradient descent, using a minimum of preprocessing, and no relevance feedback. This system achieves a workload reduction of 64.0% on average, with a minimum workload reduction of 19.3%, and a maximum workload reduction of 92.0%. On average, we would have to screen 1678 articles per topic to retrieve all relevant articles. Overall there is a large variation in performance across topics however.

We do not generally see an improvement when using relevance feedback. For the topics where relevance feedback is hypothetically feasible we sometimes see an improvement, although the effect does not appear very reliable, and the low power of the experiment means that the results are unlikely to be significant.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

Bibliography

[1] Cohen, A.M., Hersh, W.R., Peterson, K., Yen, P.: Reducing Workload in Systematic Review Preparation Using Automated Citation Classification pp. 206–219 (2006)

[2] Cohen, A.M.: Optimizing feature representation for automated systematic review work prioritization. AMIA Annual Symposium proceedings pp. 121–5 (2008)

[3] Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240. ACM (2006)

[4] Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Overview of the CLEF technologically assisted reviews in empirical medicine

[5] Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., Ouzzani, M.: Learning to identify relevant studies for systematic reviews using random forest and external information. Machine Learning 102(3), 465–482 (2016)

[6] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews 4(1), 5 (2015)

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12(Oct), 2825–2830 (2011)

[8] Petersen, H., Poon, J., Poon, S.K., Loy, C.: Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities. JMIR medical informatics 2(1), e11 (2014)

[9] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one 10(3), e0118432 (2015)

[10] Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Spijker, R., Névéol, A., Zuccon, G., Palotti, J.R.M.: Overview of the CLEF ehealth evaluation lab 2017. 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)