UWaterlooMDS at the TREC 2019 Decision Track

MUSTAFA ABUALSAUD[1], FUAT C. BEYLUNIOĞLU[2], MARK D. SMUCKER[2], and P. ROBERT DUIMERING[2],

[1] David R. Cheriton School of Computer Science, University of Waterloo

[2] Department of Management Sciences, University of Waterloo

In this report, we discuss the experiments we conducted for the TREC 2019 Decision Track. This year, our goal was to investigate the effect of document credibility on the quality of automatic runs. To address credibility, we combined scores from a spam classifier and a credibility classifier trained to detect non-trustworthy websites. The results from both classifiers were then used to modify a baseline BM25 ranking. In addition to the automatic runs, we also submitted manual runs using the HiCAL [2] system. Our manual runs modify a baseline BM25 ranking using manually judged documents found using the system.

1 INTRODUCTION

One of the goals of the TREC 2019 Decision Track is to drive researchers to implement retrieval methods that promote correct information over incorrect information. This is motivated by previous research showing that incorrect information presented on a search engine result page (SERP) can drive the user to make incorrect and potentially harmful decisions [4]. To evaluate the performance of retrieval methods, submitted runs are evaluated not only on relevance, but also on correctness and credibility of documents returned.

According to the track's assessing guidelines, relevance is assessed in a manner similar to previous tracks at TREC. Assessing the correctness of information in a document is arguably harder to determine than relevance. Correctness depends on the correspondence between the claims made in the retrieved document and those of an independent source deemed to represent the truth. Credibility, on the other hand, is related to the intentions of the information content provider and the person reading the document. A content provider wants their content to be regarded as trustworthy and accepted by readers, whereas the reader wants the information to be *true* [5, page 5]. We hypothesize that a credible document, whether it is judged as credible because it is from a trustworthy source or because of its quality, is most likely to promote correct information over incorrect information. Based on these arguments, we focus our effort on retrieving relevant and credible documents and anticipate that such effort will also be useful for retrieving correct information.

We submitted automatic and manual runs to the track. Both types of runs depended on a BM25 ranking, and then modified the BM25 ranked documents to filter out documents likely to be non-relevant or non-credible. For automatic runs, our method of determining credibility was based on a combination of scores from a spam classifier and a credibility classifier trained to detect non-trustworthy documents. Manual runs were constructed using the HiCAL system, with two assessors manually judging documents for each of the track's topics.

2 SUBMITTED RUNS

2.1 Automatic Runs

Our runs used three types of scores on (i) relevance, (ii) credibility and (iii) spaminess. Relevance scores were based on the BM25 retrieval method with default parameters as implemented in Anserini[1]. To assess credibility, we trained a logistic regression classifier on a health corpus subsetted from the ClueWeb12 dataset. Lastly, we filtered spam using spaminess scores proposed by Cormack et al. [3]. While spaminess scores based on a classifier capturing features that signal whether or not a document is spam, the credibility classifier aims to capture the

[1] https://github.com/castorini/anserini

Table 1. List of submitted runs and their description

Run identifier	Type	Judging Precedence	Description
UWaterMDS_BM25	Automatic	3	A baseline BM25 run based on Ansereni's default parameters (k1:0.9, b:0.4).
UWatMDS_BM25_ZS	Automatic	1	SPAMSCORE > 10, BM25 * (1+Z) ELSE 0
UWatMDS_BM25_Z	Automatic	2	BM25 * (1+Z)
UWatMDS_BMZBS10	Automatic	3	IF SPAMSCORE > 10, BM25 * (1+Z*2) ELSE 0
UWatMDS_BMF_C90	Automatic	4	IF P(Doc == NonCredible) < 0.90, BM25 ELSE 0
UWatMDS_BMF_C95	Automatic	5	IF P(Doc == NonCredible) < 0.95, BM25 ELSE 0
UWatMDS_BMF_S30	Automatic	Other	IF SPAMSCORE > 30 BM25, ELSE 0
UWatMDSBM25_HC1	Manual	1	Used HiCAL to find relevant documents, then return remaining unjudged BM25 documents.
UWatMDSBM25_HC2	Manual	2	Used HiCAL to find relevant documents, then return remaining unjudged BM25 documents, with documents ranking pushed higher if HiCAL classifier also found the same document.
UWatMDSBM25_HC3	Manual	1	Used HiCAL to find relevant documents, then return remaining unjudged BM25 documents, with a document ranking pushed if the classifier also found it. Here, the classifier was trained on a cleaner (only html body, no <a> tags) collection.

tone that signals whether or not the document is trustworthy. The spam and credibility scores are used to adjust the ranking by elevating the position of credible information relative to the baseline BM25 ranking.

Note that both spaminess and credibility classification rely on similar learning algorithms. As with the classifier used to generate spam scores in Cormack et al. [3], the credibility classifier employs logistic regression inputting all available character 4-grams in a document, coded based on binary features (i.e. presence or absence) rather than frequencies of occurrence as we will detail below.

2.1.1 Training and Test Collection. We prepared two different corpora subsets of ClueWeb12-B13 for (i) training the credibility classifier, and (ii) and measuring the algorithms' performance to determine final runs. For the former, we defined a set of 25 topics, and a number of queries to retrieve documents from ClueWeb12 for each topic. The created topics are similar to the track's topics but are on a different set of medical interventions and treatments. The topics cover a variety of health issues from cancer to diabetes and scoliosis, and were chosen based on different levels of controversy, from lower (e.g. exercise for scoliosis) to higher (e.g. vaccines for hepatitis B) and different target groups (such as vinpocetine for dementia, antioxidants for female subfertility). We then constructed a set of queries in the form of "[treatment] for [issue]" and its variation using synonyms and different modifiers (e.g. "antidepressants for tinnitus", "can antidepressants help tinnitus", "antidepressants for ringing in the ear"). We used Anserini with default BM25 parameters to retrieve the top 1000 documents per query and filtered out malicious pages with an open source anti-virus software, ClamAV, resulting in 40753 unique documents.

For the second corpus, we selected topics based on their popularity to ensure sufficient credible and non-credible content (e.g., "acupuncture for autism", "antibiotics for otitis media", "pilates for lower back pain", "lycopene for prostate cancer", "green tea cancer"). We retrieved 1000 documents per topic using the procedure described above.

2.1.2 Annotating the Corpora. To prepare an annotated corpus, we used HiCAL[2], a system for high-recall retrieval, to assess documents on their credibility. Given a seed query, HiCAL initially ranks the collection and

returns documents that are most-likely relevant to the seed query. Although HiCAL was designed for relevance, it was effective in finding credible (or non-credible) web pages. As HiCAL searches patterns from pages previously judged as credible (or non-credible), we could reach documents having similar page design, narrative or simply some obvious phrases (e.g. "<img src="/images/imgHonCode_petit.jpg">").

For each topic, we started a new session with an initial query in the form of "[treatment] [issue]" related to the topic. However we did not restrict the set of documents to a certain topic; rather we ran the session over the pool of all documents. As the assessment continued, since the algorithm searched for *similar* patterns, it could retrieve any page somehow related to the previous judgments independent of topic relevance and based solely on credibility[2]. In this way, we could easily detect certain non-credible pages such as forums, blogs or spam pages that are algorithmically generated by filling in a HTML template. Note that if the user labels non-credible instances as positive, HiCAL gradually retrieves more non-credible documents than credible ones. Therefore for each session, we ran HiCAL for assessing either credibility, or non-credibility or both (in separate sessions), but later merged the data by reverting signs of non-credibility assessments. This procedure yielded 2452 non-credible and 1081 credible non-duplicating documents.

For the set of 5 topics, however, we evaluated both credibility and relevance of pages to be able to test the runs' ranking performance. If a page is non-relevant it was labeled as 0; if not then it was given a score between 0 to 2, with 0 for "non-credible", 1 for "cannot decide" and 2 for "credible". For each topic, we assessed at least 200 out of 1000 documents using HiCAL, which yielded 187 / 5000 relevant instances. The only exception was Topic 5 for which, instead of using HiCAL, we assessed the top 200 out of 1000 documents based on BM25 scores. This procedure gives our baseline, UWaterMDS_BM25, a relative advantage because the only way to outperform it would be by rearranging credible documents at the top 200, rather than replacing with other credible ones from the remaining 800 documents. More specifically if BM25 ranks of 77 out of 200 true positives are distorted, 2the precision cannot recover by increasing ranks of unlabelled positive instances in the other 800. This approach for Topic 5 allowed us to observe how precise other algorithms are in lowering positions of negative instances.

We want to make an important note about the first collection here. As the authors of this study are also members of the team who designed and prepared the Decision Track, there might be similarities between a few topics in our training collection and the track's collection. The training collection is prepared to include a wide range of health documents to be able to represent the task. Although our labeling procedure disregarded relevance and thus is not restricted to annotate credibility of any certain topic, such influence might have improved the classifier's accuracy in some cases.

2.1.3 Credibility Classifier and Scores. The supervised classifier aims to detect deep patterns in web pages that signals credibility independent from the topic. These patterns can be a combination of features that gives all sorts of information from page design to colors and content. We trained a logistic regression model over raw documents by first converting text to lower case, then tokenizing into all sequential character 4-grams. For example the word "HonCode" is parsed into "honc", "onco", "ncod" and "code". We trained the classifier under binary setting, i.e. 1 if the feature is present, 0 if absent.

Each document d_i in the training collection was converted into a binary vector of N features, $\mathbf{X_i} = \langle 1; X_{i1}; \ldots; X_{ij} \rangle^T$ for $X_{ij} \in \{0; 1\}$; $j \in \{1; \ldots; N\}$ and for $\{i \in 1; \ldots; D\}$ documents. We fit standard logistic regression model:

$$P(d_i \in Credible) = \frac{1}{1 + e^{-Z_i}}; \text{ for } Z_i = \mathbf{w}^T \mathbf{X}_i$$

for $\mathbf{w}^T = \langle w_0; w_1; \ldots; w_N \rangle$.

With the above model trained on the full set of 25 topics, we computed probabilities, $P(d_i \in Credible)$, for each document in the test set of 5 topics and later for the entire TREC Decision collection.

[2]For example a session may start with scoliosis, immediately change into otitis media or dementia and visit the full range of topics.

2.1.4 Spam Scores. Spam may not attempt directly to deceive the content consumer but is an obvious type of a non-credible document. Spam also influences search engine ranks through various deceptive means, thereby improving the rank of a page regardless of the content's relevance. For example, content spam modifies the page simply by adding more keywords to increase its score for a given query.

To detect with spam documents, Cormack et al. [3] developed a spam model on the ClueWeb09 dataset which was later used to generate spam scores for ClueWeb12 collection[3]. The dataset contains spamminess percentiles ranging between 0 to 99 with 0 being the most *spammy*. We used these scores to evaluate the spaminess of each document.

2.1.5 Runs. In our runs, we aimed first to find all relevant documents and then adjust their positions with respect to their credibility. By combining the relevance, credibility and spam scores we prepared several test runs with different parameters and evaluated on the 5 test topics described above. Then we determined the best-performing algorithms to generate the final runs.

All the runs were prepared by transforming our baseline, UWaterMDS_BM25, by adding credibility reward and filtering highly spammy documents. Below we describe these transformations and present their performance results.

2.1.6 `UWaterMDS_BM25`. In this run, we used the BM25 retrieval algorithm as implemented in Ansereni. The parameters used for this run were the default parameters set by Ansereni. We used the query field of the topic as input to the algorithm.

2.1.7 `UWatMDS_BM25_ZS`, `UWatMDS_BM25_Z`, `UWatMDS_BMZBS10` . In the test runs, using the credibility classifier probabilities to filter out documents or adjust their ranks directly (e.g. by computing $BM25 \quad P^1d\ 2\ Credible^o$) resulted in discarding many useful documents. Besides the probabilities are not appropriate for linear transformations for re-ranking documents. Therefore we transformed the probabilities to Z-scores using logit function and used them to linearly combine with other scores:

$$p = P^1d\ 2\ Credible^o = \frac{1}{1 + e^{\ Z}}; \text{ and } Z = \ln \frac{p}{1 \quad p}$$

To combine with BM25, we rescaled Z to 0-1. As we aim to improve BM25 scores to favour credibility, we added credibility reward proportionate to relevance as $BM25 \quad ^1 1 + Z^o$. If Z is close to 0 (non-credible) then the score remains unchanged whereas if it is close to 1, it sums up to double. Using this approach, the interaction between credibility and relevance prevents a non-relevant but credible document from occupying a higher position in the results. We also added a parameter, β to Z score to control the relative weight of credibility judgments:

$$BM25 \quad ^1 1 + \beta Z^o$$

Further, we used spam scores (SPAM) as filters. In the test runs it filtered out too many necessary documents when the threshold was set to 70. The best performances were reached when $10 < SPAM < 40$ varying with respect to the test topic. Hence we used the spam filter with a threshold of 10 to filter out "junk" documents and adjusted scores using the above rule.

After a series of trials with $SPAM\ 2\ f10; 20; : : : ; 70g$ and $\beta\ 2\ f1; 1{:}1; : : : ; 2g$ we decided on the rules listed below, which generated relatively higher precision and more stable results.

1. `UWatMDS_BM25_ZS = IF SPAM > 10,` $BM25 \quad ^1 1 + Z^o$ `ELSE 0`
2. `UWatMDS_BM25_Z =` $BM25 \quad ^1 1 + Z^o$
3. `UWatMDS_BMZBS10 = IF SPAM > 10,` $BM25 \quad ^1 1 + Z \quad 2^o$ `ELSE 0`

[3]Spam scores are available at https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/

Table 2. Mean average precision (map) and geometric map of the methods on our self-created tuning topics.

Topic Rel / Ret	1 9 / 1000	2 70 / 1000	3 17 / 1000	4 14 / 1000	5* 77 / 1000	All 187 / 5000	All (gm_map)
BM25	0.160	0.377	0.065	**0.016**	**0.514**	0.226	0.126
BM25_Z	0.345	0.456	0.132	0.013	0.470	0.283	0.165
BM25_ZS	**0.346**	0.462	0.141	0.015	0.393	0.271	0.167
BMZBS10	0.336	**0.463**	0.140	0.014	0.384	0.268	0.165
BMF_C90	0.255	0.298	**0.157**	0.005	0.170	0.177	0.100
BMF_C95	0.362	0.302	0.141	0.004	0.146	0.191	0.097
BMF_S30	0.163	0.380	0.083	0.010	0.321	0.191	0.111

2.1.8 `UWatMDS_BMF_C90`, `UWatMDS_BMF_C95` *and* `UWatMDS_BMF_S30`. When combined with relevance scores as filters, the logistic regression classifier's computed $P^1d\ 2\ NonCredible^o$ and spam filter could also improve results in some cases. To determine parameters, we generated runs for $SPAM\ 2\ f10; 20; : : : ; 70g$ and $P^1d\ 2\ NonCredible^o < p_0$ for $p_0\ 2\ f0{:}89; 0{:}90; : : : ; 0{:}99g$, and chose the following rules based on their performances:

4. `UWatMDS_BMF_C90 = IF` $P^1d\ 2\ NonCredible^o < 0{:}90$ `BM25, ELSE 0`
5. `UWatMDS_BMF_C95 = IF` $P^1d\ 2\ NonCredible^o < 0{:}95$ `BM25, ELSE 0`
6. `UWatMDS_BMF_S30 = IF SPAM > 30 BM25, ELSE 0`

2.1.9 Test Performances. Table 2 presents the performance of automatic runs on 5 test topics. As shown in the table, the credibility classifier improves BM25 scores. BM25 filtered with 90% level (`UWatMDS_BMF_C90`) yielded inconsistent results, outperforming all other algorithms for Topic 3, but only marginally improving on BM25 baseline for Topic 1. The combinations of BM25, Z and SPAM (`UWatMDS_BM25_Z`, `UWatMDS_BM25_ZS`, `UWatMDS_BM25ZBS10`) performed better than standard BM25, and even doubling its precision for some topics.

For the 5th topic, as expected, other algorithms performed worse than `UWatMDS_BM25`. As described earlier in Section 2.1.2, assessment of this topic gives BM25 an advantage and a better algorithm would precisely reduce the ranks of non-credible or non-relevant documents. As Table 2 shows, spam and credibility classifiers reduced precision sharply when used as filters (`UWatMDS_BMF_C90`, `UWatMDS_BMF_C95`, `UWatMDS_BMF_S30`) suggesting that these algorithms filtered out many true positive documents. On the other hand, `UWatMDS_BM25_Z` caused the least distortion to the overall position of positive instances.

2.2 Manual Runs

We submitted 3 manual runs for the track. We used the HiCAL system for manually assessing documents. Currently, HiCAL only supports rendering structured documents (e.g. news articles, tweets, etc). We extended HiCAL to render HTML based documents, as shown in Figure 1. We provided the topic's query as the seed query for HiCAL's classifier. To mitigate the cold-start problem and for the classifier to "learn" the topic of interest, we modified the system to make the assessor start the assessment process with the top 20 documents returned by BM25. Judgments on the first 20 documents were then used to retrain the classifier, and the most likely relevant document is shown to the assessor. The assessment is stopped once the assessor has judged 200 documents.

2.2.1 Collection (Documents Set). The current version of HiCAL operates entirely on memory. For HiCAL to work, the entire ClueWeb12-B13 collection (1.95 TB) needs to be loaded into memory. This requirement was not feasible with our resources at the time of assessing. Instead, we operated on a subset of the collection that was manually constructed. To construct the subset, we collected the top 10,000 documents returned by BM25 for each

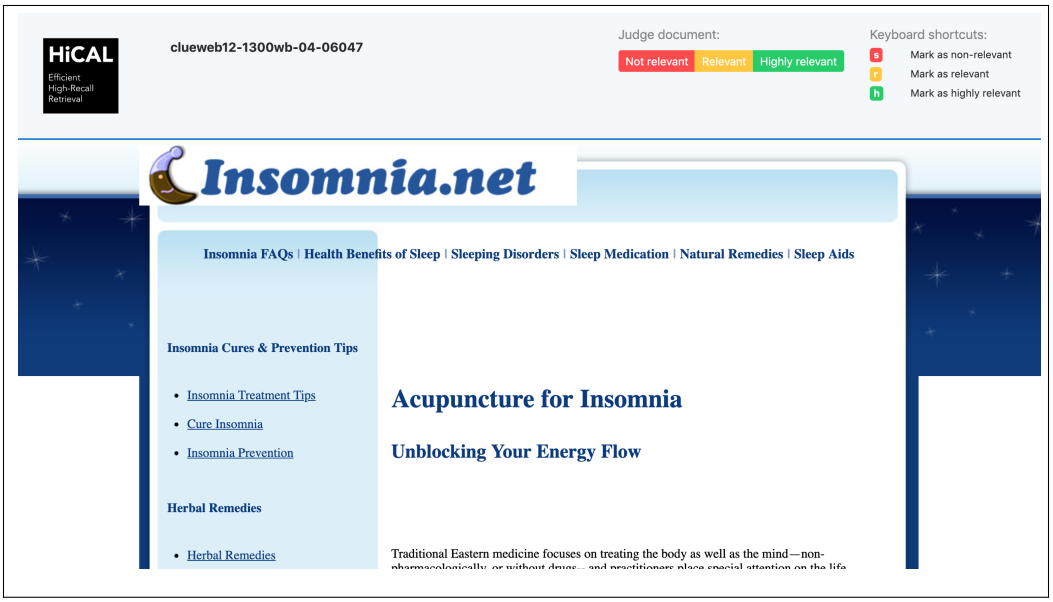

Fig. 1. HiCAL's [2] assessment interface modified to render HTML-based collections. Here, an example document on the topic "acupuncture for insomnia" is shown to the user. Users can judge the document using a three-level relevance scale provided on the top of the page. Once a user judges a document, HiCAL shows the next likely relevant document based on an active-learning classifier.

of the 51 queries in the topics file, resulting in 356,635 unique documents. We also added 10,000 51 random documents to allow tf-idf features in HiCAL to be more representative of the entire collection. In total, there were 866,635 documents in the constructed subset of the collection. Training HiCAL's classifier was done on the raw HTML document without any pre-processing.

2.2.2 Assessors. In total, there were 51 topics. Assessment of the topics was divided between two graduate students who are involved in this task. One assessor judged 15 topics and the other judged 36 topics. We randomly assigned the topics to each assessor.

2.2.3 Judging Criteria. As important as relevance is for this task, other criteria are also important, namely the credibility of the document and the correctness of the information. While credibility is possible to judge (based on the assessor's knowledge of reliable sources of information, for example), it is much harder to judge the accuracy of information medical treatments towards health issues, especially for assessors who are not medical experts and without any medical accreditation. Our relevance scale, therefore, is modified to account for relevance and credibility only. Another option we considered was having separate judging criteria for relevancy and credibility. This would slow down the assessment process, however, and due to time limits, we opted for a single relevance scale.

Before the assessing process, both assessors were informed of the following information when judging:

> **Highly Relevant:** Documents that contain information about the medical treatment and the health issue and appear to be from reliable sources or contain reliable content (e.g., mention of doctor or clinic name, or providing reliable references).

Relevant: Documents that contain information about the medical treatment and the health issue, but unsure of the credibility of the webpage.

Not Relevant: Documents that are either off topic, or completely not trustworthy (e.g., spam documents).

While documents judged as relevant may be considered not credible by NIST assessors, judging them as not relevant could hurt HiCAL's classifier. In particular, a relevant document may contain relevant terms that are useful for the training step in HiCAL. Having the classifier trained on such documents is useful to find other, potentially highly relevant documents.

2.2.4 Number of Relevant Documents. Figure 2 indicates the number of relevant documents found using HiCAL. We expected highly relevant documents (e.g. documents we consider relevant and credible) to be harder to find. Across all topics, a total of 535 highly relevant and 1345 relevant documents were found.

One of the challenges we encountered by using HiCAL for finding relevant documents was the quality of documents returned by the classifier. While HiCAL has been shown to be successful in achieving high recall on datasets such as news articles [1, 6], using the system on a web collection, such as ClueWeb, introduced unanticipated challenges. For example, while using HiCAL, both assessors agreed that the classifier returned many documents that were completely off topic but from a website with a document they judged as relevant earlier during the assessment process. A possible explanation is due to the nature of the HTML documents. Boilerplate content associated with web pages judged as relevant could negatively affect the quality of the classifier.

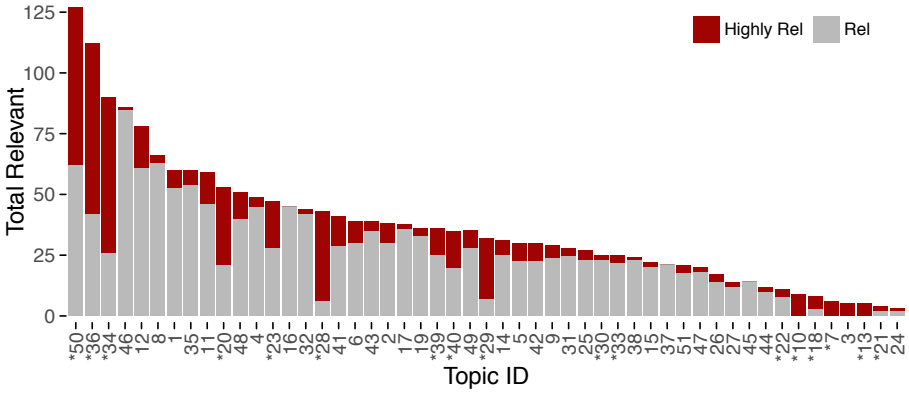

Fig. 2. Number of highly relevant and relevant documents found for each topic using HiCAL, sorted by the total value. Star symbol next to topic IDs are used to separate topics assessed by the two assessors.

2.2.5 `UWatMDSBM25_HC1`. To construct this run, we first append documents judged as highly relevant, then by relevant documents, sorted in reverse chronological order of appearance. The remainder of the list was filled by unjudged BM25 documents, sorted by their associated score.

2.2.6 `UWatMDSBM25_HC2`. We filled the list with our known judged relevant document the same way as in `UWatMDSBM25_HC1`. The remainder of the list was filled using a combination of BM25 ranking and a HiCAL classifier built for each topic. For each topic with 20 relevant documents, we built a final classifier to score each document in the sub-collection. We use the top 1,000 most-likely relevant documents returned by the classifier to alter the ranking produced by BM25. The premise behind this approach is that documents that appeared both in the top 1,000 most likely relevant documents set by the classifier and in BM25 top 1,000 documents should be ranked higher than those that appeared only in BM25's set. This approach allows potentially relevant documents

Table 3. Performance of the runs under the track's evaluation measures. Bold values indicate highest value in the column. While the result here show that our baseline run UWaterMDS_BM25 (gray shaded) performs best in CAM, our analysis in Section 3.1 show that CAM may not reflect the true performance of runs with respect to the track's goal. Numbers in parentheses indicate percent of change over the baseline score.

Run	MAP	nDCG@10	NLRE	CAM
Manual Runs				
UWatMDSBM25_HC1	0.403 (+6.99%)	0.450 (-9.67%)	0.997 (+0.14%)	0.536 (-2.14%)
UWatMDSBM25_HC2	0.391 (+3.91%)	0.450 (-9.67%)	0.998 (+0.18%)	0.534 (-2.57%)
UWatMDSBM25_HC3	0.411 (+9.14%)	0.450 (-9.67%)	0.998 (+0.25%)	0.539 (-1.66%)
Automatic Runs				
UWatMDS_BM25_Z	0.345 (-8.40%)	0.443 (-11.15%)	0.997 (+0.10%)	0.547 (-0.18%)
UWatMDS_BM25_ZS	0.310 (-17.51%)	0.430 (-13.72%)	0.997 (+0.11%)	0.510 (-6.96%)
UWatMDS_BMF_C90	0.156 (-58.50%)	0.425 (-14.78%)	0.999 (+0.33%)	0.309 (-43.60%)
UWatMDS_BMF_C95	0.170 (-54.86%)	0.445 (-10.75%)	0.999 (+0.33%)	0.334 (-39.04%)
UWatMDS_BMF_S30	0.285 (-24.15%)	0.500 (+0.28%)	0.998 (+0.20%)	0.456 (-16.74%)
UWatMDS_BMZBS10	0.283 (-24.89%)	0.392 (-21.36%)	0.997 (+0.13%)	0.492 (-10.21%)
UWaterMDS_BM25	0.376	0.499	0.996	0.548

that are low in BM25 ranked list to be pushed higher in the list if they were found to be most-likely relevant by the classifier. As the classifier is also trained with non-relevant documents, unjudged non-relevant documents that are ranked high in BM25 can be pushed even lower, effectively increasing the quality of the list. Documents judged as non-relevant are automatically skipped.

Training the classifier with few documents is not practical. Therefore, topics with less than 20 documents have the same ranking as in UWaterMDS_BM25.

2.2.7 UWatMDSBM25_HC3. In an effort to determine the effect of boilerplate content in a webpage, and whether simple pre-processing of the collection can improve the quality of HiCAL's classifier, we processed our sub-collection to remove all HTML related tags, remove content of all <a> tags, and only keep the content of <body> tag. Following this step, the construction of the run was the same as UWatMDSBM25_HC2.

3 RESULT AND DISCUSSION

Table 3 shows the performance of our runs under the track's evaluation measures. NLRE and CAM are combination measures designed to combine different assessing aspects into a single score. Here, both NLRE and CAM used all three aspects to compute its final score. The combination measures proposed by the track's organizers make it difficult to interpret the performance of the runs in one aspect over the other (e.g., how does a run optimized for credibility perform in terms of credibility alone?). In the next section, we look into the performance of runs under different combinations of aspects and under all aspects by modifying the notion of relevance.

3.1 Evaluation Under Different Aspects

To understand how the runs perform in terms of correctness and credibility, we computed MAP scores for each aspect separately (relevance, credibility, and correctness). It is important to note that based on the track's assessing guidelines, treatment efficacy (from which the correctness of information is computed) and credibility

Table 4. MAP result using relevance, correctness, and credibility assessments separately. "All" indicate MAP score by enforcing documents to be considered valid only if it is relevant, correct and credible. UWaterMDS_BM25 (gray shaded) is our baseline run. Bold values indicate highest value in the column for manual and automatic runs. Numbers in parentheses indicate percent of change over the baseline score. Statistical significance over the baseline is computed for bolded values (* indicates statistical significance at p<0.05).

Run	Relevance	Correctness	Credibility	All
		Manual Runs		
UWatMDSBM25_HC1	0.403 (+6.99%)	0.167 (+27.49%)	0.283 (+10.07%)	0.128 (+32.47%)
UWatMDSBM25_HC2	0.391 (+3.91%)	0.163 (+24.81%)	0.277 (+7.86%)	0.125 (+30.08%)
UWatMDSBM25_HC3	**0.411 (+9.14%)**	**0.171 (+31.01%)**	**0.289 (+12.49%)**	**0.131 (+36.20%)**
		Automatic Runs		
UWatMDS_BM25_Z	**0.345 (-8.40%)**	**0.122 (-6.58%)**	**0.306 (+19.06%)**	**0.119 (+23.86%)**
UWatMDS_BM25_ZS	0.310 (-17.51%)	0.113 (-13.25%)	0.291 (+13.22%)	0.117 (+21.47%)
UWatMDS_BMF_C90	0.156 (-58.50%)	0.074 (-43.42%)	0.202 (-21.28%)	0.095 (-1.87%)
UWatMDS_BMF_C95	0.170 (-54.86%)	0.080 (-39.13%)	0.217 (-15.75%)	0.101 (+5.08%)
UWatMDS_BMF_S30	0.285 (-24.15%)	0.108 (-17.38%)	0.241 (-6.26%)	0.100 (+3.84%)
UWatMDS_BMZBS10	0.283 (-24.89%)	0.104 (-20.44%)	0.275 (+7.08%)	0.111 (+14.73%)
UWaterMDS_BM25	0.376	0.131	0.257	0.096

Table 5. nDCG@10 result using relevance, correctness, and credibility assessments separately. "All" indicate nDCG@10 score by enforcing documents to be considered valid only if it is relevant, correct and credible. UWaterMDS_BM25 (gray shaded) is our baseline run. Bold values indicate highest value in the column for manual and automatic runs. Numbers in parentheses indicate percent of change over the baseline score. No statistical significance (at significance level of 0.05) was found over baseline.

Run	Relevance	Correctness	Credibility	All
		Manual Runs		
UWatMDSBM25_HC1	0.450 (-9.79%)	0.232 (+23.76%)	0.420 (+6.06%)	0.180 (+33.48%)
UWatMDSBM25_HC2	0.450 (-9.79%)	0.232 (+23.76%)	0.420 (+6.06%)	0.180 (+33.48%)
UWatMDSBM25_HC3	0.450 (-9.79%)	0.232 (+23.76%)	0.420 (+6.06%)	0.180 (+33.48%)
		Automatic Runs		
UWatMDS_BM25_Z	0.443 (-11.28%)	0.180 (-4.06%)	0.452 (+14.04%)	0.175 (+29.92%)
UWatMDS_BM25_ZS	0.430 (-13.84%)	0.181 (-3.31%)	0.446 (+12.60%)	0.177 (+31.25%)
UWatMDS_BMF_C90	0.425 (-14.90%)	0.193 (+3.15%)	0.444 (+12.12%)	0.178 (+32.37%)
UWatMDS_BMF_C95	0.445 (-10.88%)	0.201 (+7.26%)	**0.463 (+16.86%)**	**0.183 (+36.01%)**
UWatMDS_BMF_S30	**0.500 (+0.14%)**	**0.209 (+11.59%)**	0.426 (+7.62%)	0.161 (+19.52%)
UWatMDS_BMZBS10	0.392 (-21.47%)	0.163 (-12.76%)	0.418 (+5.50%)	0.162 (+20.49%)
UwaterMDS_BM25	0.499	0.187	0.396	0.135

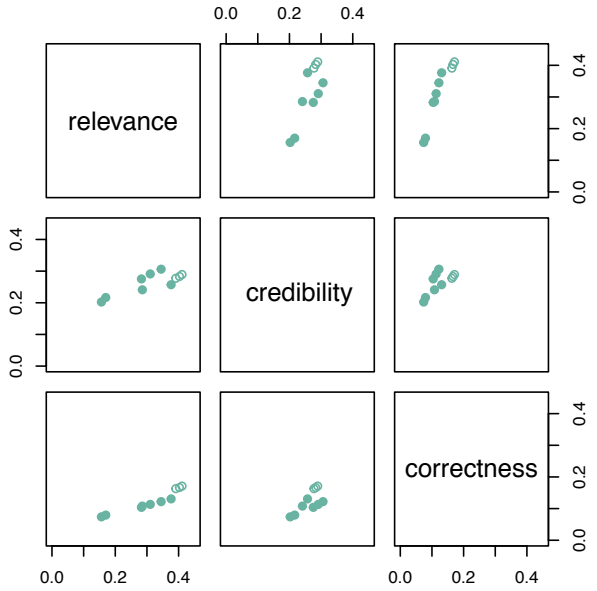

Fig. 3. MAP result using relevance, correctness, and credibility assessments separately (c.f. Table 4). Halo circles indicate manual runs.

are only assessed if the document is relevant. Therefore, by considering each of the aspects separately, we are actually measuring relevance with regards to another aspect. For example, in measuring performance in terms of correctness, a document is relevant only if it is relevant to the topic and contains correct information (similarly for credibility). Figure 3 and Table 4 show MAP evaluated using relevance, as provided by NIST, and correctness and credibility evaluated separately using the provided assessments from NIST (qrels). Here, we assume any document that is not relevant or has not been judged in terms of relevance, to be incorrect and not credible, following the same assumption made by the track organizers for evaluating NLRE and CAM measures.

Our automatic run, `UWatMDS_BM25_Z`, which uses BM25 scores and Z-scores from our credibility classifier, has the highest performance in terms of credibility. It is interesting to note that credibility does not seem to indicate correctness, as the same run performs below expected in terms of correctness. We leave this question for future work.

In terms of correctness, our manual runs perform best. Because both assessors were heavily involved in the creation of the track's topics and knew the efficacy of the various treatments, it is possible that their knowledge of the topics may have influenced their judgments towards correct documents.

One of the task's goal is to drive search engines to provide correct and credible information to users. To obtain another view on performance, we computed MAP with relevance defined such that a document is relevant if and only if it is 1) relevant to the topic, 2) provides correct information and 3) is considered a credible document. Note that by enforcing such a rule, we effectively disregard documents that are relevant and credible but contain incorrect information, which we argue are documents that are most harmful to show to users. Column "All" in Table 4 shows the MAP scores when such a requirement is enforced. Here, we see that our baseline run, `UWaterMDS_BM25`, which ranks first in terms of CAM, underperforms and is ranked second to lowest among our runs. Our manual runs, as well as `UWatMDS_BM25_Z` automatic run seem to perform best. In light of these results, NLRE and CAM may not appropriately reflect the true performance of runs with respect to the track's goals.

MAP scores in Table 4 provide information of overall performance. To highlight the quality of the produced ranking from a user perspective, we computed nDCG@10 scores for all runs using different types of assessments as we did in Table 4. The nDCG@10 scores for all runs are shown in Table 5.

The scores for manual runs are the same because each manual run used the documents found using HiCAL and were manually judged as highly relevant or relevant as the initial set of documents in the ranked list, and modified the remaining non-judged documents based on the methods described in Section 2.2.5-2.2.7. In terms of relevance only, nDCG@10 scores are slight lower than the baseline. If we consider correctness and credibility assessment separately, manual runs improve over the baseline. Similarly for "All", where a documents is valid only if it is relevant, correct and credible.

When using relevance assessments only, both manual and automatic runs underperform compared to the baseline. The only exception is `UWatMDS_BMF_S30`, which does not negatively influence nDCG@10 scores in terms of relevance alone, and can also improve nDCG@10 scores in terms of correctness more than the other automatic runs and more than the baseline by 11.59%. To find more credible information, `UWatMDS_BMF_C95` (which uses a credibility filter with a 95% threshold) performs slightly better than re-ranking methods (e.g. `UWatMDS_BM25_Z`) and manual runs. When considering a document to be valid only if it is relevant, correct, and credible (column "All"), all runs improve over the baseline, with most contribution is made by `UWatMDS_BMF_C95` with 36% improvement, though none of the differences are statistically significant at 5% significant level.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-03642), and in part by the University of Waterloo.

REFERENCES

[1] Mustafa Abualsaud, Gordon V. Cormack, Nimesh Ghelani, Amira Ghenai, Maura R. Grossman, Shahin Rahbariasl, Haotian Zhang, and Mark D. Smucker. 2017. UWaterlooMDS at the TREC 2018 Common Core Track. In *TREC*.

[2] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2018. A System for Efficient High-Recall Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 1317–1320. https://doi.org/10.1145/3209978.3210176

[3] Gordon V Cormack, Mark D Smucker, and Charles LA Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval* 14, 5 (2011), 441–465.

[4] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 209–216.

[5] Adam Wierzbicki. 2018. *Web Content Credibility*.

[6] Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Angshuman Ghosh, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2017. UWaterlooMDS at the TREC 2017 Common Core Track. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*. https://trec.nist.gov/pubs/trec26/papers/UWaterlooMDS-CC.pdf