# Building Systematic Reviews Using Automatic Text Classification Techniques

**Oana Frunza**, **Diana Inkpen**, and **Stan Matwin**
School of Information Technology and Engineering
University of Ottawa Ottawa, ON, Canada, K1N 6N5
{ofrunza,diana,stan}@site.uottawa.ca

## Abstract

The amount of information in medical publications continues to increase at a tremendous rate. Systematic reviews help to process this growing body of information. They are fundamental tools for evidence-based medicine. In this paper, we show that automatic text classification can be useful in building systematic reviews for medical topics to speed up the reviewing process. We propose a per-question classification method that uses an ensemble of classifiers that exploit the particular protocol of a systematic review. We also show that when integrating the classifier in the human workflow of building a systematic review, the per-question method is superior to the global classification method. We test several evaluation measures on a real dataset.

## 1 Introduction

Systematic reviews are the result of a tedious process which involves human reviewers to manually screen references of papers to determine their relevance to the review. This process often entails reading thousands or even tens of thousands of abstracts from prospective articles. As the body of available articles continues to grow, this process is becoming increasingly difficult.

Common systematic review practices stipulate that two reviewers are used at the screening phases of a systematic review to review each abstract of the documents retrieved after a simple query-based search. After a final decision is made for each abstract (the two reviewers decide if the abstract is relevant or not to the topic of review), in the next phase, further analysis (more strict screening steps) on the entire article is used in order to identify if the article is clinically relevant or not, to extract information, etc. A systematic review has to be complete, articles that are published on a certain topic and are clinically relevant need to be part of the review. This requires near-perfect recall since the accidental exclusion of a potentially relevant abstract can have a significantly negative impact on the validity of the overall systematic review (Cohen *et al*., 2006). Our goal in this paper is to propose an automatic system that can help human judges in the process of triaging articles by looking only at abstracts and not the entire documents. This decision step is known as the initial screening phase in the protocol of building systematic reviews, only the abstracts are used as source of information.

One reviewer will still read the entire collection of abstracts while the other will benefit from the help of the system; this reviewer will have to label only the articles that will be used to train the classifier (ideally a small proportion for obtaining a high workload reduction), the rest of the articles will be labeled by the classifier.

In the systematic review preparation, if at least one reviewer agrees to include an abstract, the abstract will have the labeled included and it will pass to the next screening phase; otherwise, it will be discarded. Therefore, the benefit of doubt plays an important role in the decision process. When we replace one reviewer with the automatic classifier, because we keep one human judge in the process, the confidence and reliability of the systematic review is still higher while the overall workload is reduced. The reduction is from the time required for two passes through the collection (for the two humans) to only one pass and the smaller part labeled by the reviewer which is assisted by the classifier. Figure 1 presents on overview of our proposed workflow.

The task that needs to be solved in order to help the systematic review process is a text classification task intended to classify an abstract as relevant or not relevant to the topic of review.

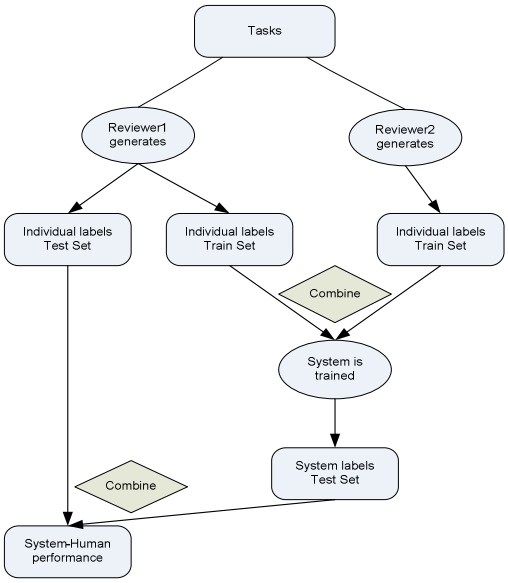


Figure 1. Embedding automatic text classification in the process of building a systematic review.

The hypothesis that guides our research is that it is possible to save time for the human reviewers and obtain good performance levels, similar to the ones obtained by humans. In this current study we show that we can achieve this by building a classification model that is based on the natural human workflow used for building systematic reviews. We show, on a real data set, that a human-machine system obtains the best results when an ensemble of classifiers is used as the classification model.

## 2 Related Work

The traditional way to collect and triage the abstracts that are part of a systematic review consists in using simple query search techniques based on MeSH[1] or keywords terms. The queries are usual Boolean-based and are optimized either for precision or for recall. The studies done by Haynes *et al.* (1994) show that it is difficult to obtain high performance for both measures.

The research done by Aphinyanaphongs and Aliferis (2005) is probably the first application of automatic text classification to the task of creating systematic reviews. In that paper, the authors experimented with a variety of text classification techniques using the data derived from the ACP Journal Club as their corpus. They found that support vector machine (SVM) was the best classifier according to a variety of measures. Further work for systematic reviews was done by Cohen *et al.* (2006). Their work is mostly focused on the elimination of non relevant documents. As their main goal is to save work for the reviewers involved in systematic review preparation, they define a measure, called work saved over sampling (WSS) that captures the amount of work that the reviewers will save with respect to a baseline of just sampling for a given value of recall. The idea is that a classifier returns, with high recall, a set of abstracts, and only those abstracts need to be read to weed out the non-relevant ones. The savings are measured with respect to the number of abstracts that would have to be read if a random baseline classifier was used. Such baseline corresponds to uniformly sampling a given percentage of abstracts (equal to the desired recall) from the entire set. In Cohen *et al.* (2006), the WSS measure is applied to report the reduction in reviewer's work when retrieving 95% of the relevant documents, but the precision was very low.

We focus on developing a classifier for systematic review preparation, relying on characteristics of the data that were not included in the Cohen et *al.'s* (2006), because the questions asked in the preparation of the reviews are not available, Therefore we cannot perform a direct comparison of results here. Also, the data sets that they used in their experiments are significantly smaller than the one that we used.

## 3 The Data Set

A set of 47,274 abstracts with titles were collected from MEDLINE[2] as part of a systematic review done by the McMaster University's Evidence-Based Practice Center using TrialStat Corporation's Systematic Review System[3], a web-based software platform used to conduct systematic reviews.

The initial set of abstracts was collected using a set of Boolean search queries that were run for the specific topic of the systematic review: "*the dissemination strategy of health care services for elderly people of age 65 and over*".

In the protocol applied, two reviewers work in parallel. They read the entire collection of 47,274 abstracts and answer a set of questions to determine if an abstract is relevant or not to the topic of review. Examples of questions present in the protocol: *Is this article about a dissemination strategy or a behavioral intervention?; Is this a primary study?; Is this a review?; etc.* An abstract is not considered to pass to the next screen-

[1] http://www.nlm.nih.gov/mesh/

[2] http://medline.cos.com

[3] http://www.trialstat.com/

ing phase, when the entire article is available, if the two reviewers respond negative to the same question for a certain abstract. All other cases of possible responses suggest that the abstract will be part of the next screening phase. In this paper we focus on the initial screening phase, the only source of information is the abstract and the title of the article, while having as the main objective an acceptable level of recall not to mistakenly exclude relevant abstracts.

From the entire collection of labeled abstracts only 7,173 are relevant, and the rest of 40,101 are non-relevant. Usually, in the process of building systematic reviews the number of non-relevant documents is much higher than the number of relevant ones. The initial retrieval query is purposefully very broad, so as not to miss any relevant papers.

## 4 Methods

The machine learning techniques that could be used in the process of automating the creation of systematic reviews need to take into account some issues that can arise when dealing with such tasks. *Imbalanced data* sets are usually what we deal with when building reviews, the proportion of relevant articles that end up being present in the review is significantly lower compared with the original data set. The benefit of doubt will affect the quality of the data used to train the classifier, since a certain amount of *noise* is introduced: abstracts that are in fact non-relevant can be labeled as being relevant in the first screening process. The relatively high number of abstracts involved in the process will make the classification algorithms deal with a high number of features and the *representation* technique should try to capture aspects pertaining of the medical domain.

### 4.1 Representation Techniques

In our current research, we use three representation techniques: bag-of-words (BOW), concepts from the Unified Medical Language System (UMLS), and a combination of both.

The **bag-of-words** representation is commonly used for text classification and we have chosen to use binary feature values. Binary feature values were shown to out-perform weighted values for text classification tasks in the medical domain as shown by Cohen et *al.* (2006) and binary values tend to be more stable in results than frequency values for a task similar to ours, as shown by Ma (2007).

We considered feature words delimitated by space and simple punctuation marks that appeared at least three times in the training data, were not part of a stop words list[4], and had a length greater than three characters. With these constraints, we extracted approximately 30,000 word features. No stemming was used.

**UMLS concepts** which are part of the U.S. National Library of Medicine[5] (NLM) knowledge repository are identified and extracted form the collection of abstracts using the MetaMap[6] system. This conceptual representation helped us overcome some of the shortcomings of BOW representation, and allowed us to use multi-word features, medical knowledge, and higher-level meanings of words in context. As Cohen (2008) shows, multi-word and medical concept representations are suitable for the task.

### 4.2 Classification Algorithms

As a classification algorithm we have chosen to use the complement naive Bayes (CNB) (Frank and Bouckaert, 2006) classifier from the Weka[7] tool. The reason for this choice is that the CNB classifier implements state-of-the-art modifications of the standard multinomial naïve Bayes (MNB) classifier for a classification task with highly skewed class distribution. As the systematic reviews data usually contain a large majority of not relevant abstracts, resulting in a skewness reaching even below 1%, it is important to use appropriate classifiers.

CNB modifies the standard MNB classifier by applying asymmetric word count priors, reflecting skewed class distribution (Drummond and Holte, 2003). We experimented with other classifiers from Weka as well (decision tress, support vector machine, instance-based learning, boosting, etc.), but the results obtained with CNB are better than the results of the other classifiers and this is why in the paper we report the results only for this classifier.

### 4.3 Global Text Classification Method

The first method that we propose in order to solve the text classification task that is intended to help a systematic review process is a straightforward machine learning approach. We trained a classifier, CNB is the one for which we will report the results, on a collection of abstracts and

[4] http://www.site.uottawa.ca/~diana/csi5180/StopWords
[5] http://www.nlm.nih.gov/pubs/factsheets/umls.html
[6] http://mmtx.nlm.nih.gov/
[7] www.cs.waikato.ac.nz/machine learning/weka/

then evaluated the classifier's performance on a separate test data set. The power of this classification technique stands in the ability to use a suitable classification algorithm and a good representation for the text classification task; Cohen *et al.* (2006) also used this approach. We randomly split the data set described in Section 3, into a training set and a test set. We used the first part of the split for training and the second one to evaluate the classifiers' performance in discriminating an abstract as having one of the two possible classes, **Included (relevant)** or **Excluded (non relevant)**. We decided to work with a training set smaller than the test set because ideally good results need to be obtained without using too much training data. We have to take into consideration that training a classifier for a particular topic, human effort is required to annotate at least part of the collection of abstracts.

From the collection of 47,274 abstracts 20,000 were randomly taken to be part of the training data set and the remaining 27,274 represents the test set. Table 1 presents a summary of the data along with the class distribution in the training and test data sets. We randomly sampled the data to build the training and test data sets, and the original distribution of 1:5.6 between the two classes holds in both sets.

| Data set | No. of abstracts | Class distribution Included : Excluded (ratio) |
|---|---|---|
| Training | 20,000 | 3,056 : 16,944 (1:5.6) |
| Testing | 27,274 | 4,117 : 23,157 (1:5.6) |

**Table 1**. Training and test data sets.

#### 4.3.1 Feature Selection

Using the global method, we performed experiments with several feature selection algorithms. We used only the BOW representation.

**Chi$^2$** is a measure that evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. We used different ratios of **Included** and **Excluded** class in the training data. We selected the top $k_1$ $CHI^2$ features that are exclusively included (appeared only in the training abstracts that are classified as **Included**) and the top $k_2$ $CHI^2$ features that are exclusively excluded (appeared only in the training abstracts that are classified as **Excluded**) and used them as a representation for our data set. We varied the $k_1$ and $k_2$ parameters from 10 to 150 for the parameter $k_1$ and from 5 to 150 for the parameter $k_2$. We used a minimum of 20 features and a maximum of 300.

**InfoGain** evaluates the worth of an attribute by measuring the information gain with respect to the class. We run experiments when we varied the number of selected features from 50 to 500. We used a number of 50, 100, 150, 250, 300 and 500 top features (words that were present in the training data).

Bi-Normal Separation (**BNS**) is a feature selection technique that measures the separation between the threshold occurrences of a feature in one of the two classes. The latter measure is described in detail in Forman (2002). We used a ratio of features that varies from 10 to 150 for the most representative features for the **Included** class and from 5 to 150 for the **Excluded** class. For some experiments the number of features for the **Included** class is higher than the number of features for the **Excluded** class. We have chosen to do so because we wanted to re-balance the imbalance of classes in the training data set. After selecting the number of **Included** and **Excluded** features, we used the combination to represent our entire collection of abstracts.

We used the implementation from the Weka package for the **Chi$^2$** and **InfoGain** and the BNS implementation done by Ma (2007).

### 4.4 Per-Question Classification Method

The second method that we propose for solving our task takes into account the specifics of the systematic review process. More exactly, it takes advantage of the set of questions that the reviewers use in the process of deciding if an abstract is relevant or not. These questions are created in the design step of the systematic review and almost all systematic reviews have them. By using these questions we better emulate how the human judges think and work when building systematic reviews.

We have chosen to use only the questions that have inclusion/exclusion criteria, there were also some opened answer questions involved in the review, because they are the ones that are important for reviewers to make a decision.

To collect training data for each question, we used the same training data set as in the previous method (but note that not all the abstracts have answers for all the questions; therefore the training set sizes differ for each question). We also kept the same test data set. Table 2 presents an overview of the questions and data sets used.

When we created a training data set for each question separately, we removed the abstracts for which we had a disagreement between the human experts – two different answers for a spe-

cific question, because they represent noise in the training data. We need to train classifiers only on reliable data, when possible. For each of the questions from Table 2, we trained a CNB classifier on the corresponding data set.

| Question (Training : Included class : Excluded class) |
|---|
| Q1 - Is this article about a dissemination strategy or a behavioural intervention? **(14,057:1,145:12,912)** |
| Q2 - Is the population in this article made of individuals 65-year old or older or does it comprise individuals who serve the elderly population needs (i.e. health care providers, policy makers, organizations, community)? **(15,005:7,360:7,645)** |
| Q3 - Is this a primary study? **(8,825:6,895:1,930)** |
| Q4 - Is this a review? **(6,429:5,640:789)** |

**Table 2.** Data sets for the per-question classification method.

We used the same representation for the per-question classifiers as we did for the global classifier: BOW, UMLS (the concepts that appeared only in the new question-oriented training data sets), and the combination BOW+UMLS. We used each trained model to obtain a prediction for each instance from the test set; therefore each test instance was assigned four prediction values of 0 or 1. The predictions have values of 0 or 1. To assign a final class for each test instance, from the prediction of all four classifiers, the class of a test instance is decided according to one of the following four schemes:

1. If any one vote is **Excluded**, the final class of a test instance is **Excluded**. This is a 1-vote scheme, i.e., any one classifier voted **Excluded**.

2. If any two votes are **Excluded**, the final class of a test instance is **Excluded**. This is a 2-vote scheme.

3. If any three votes are **Excluded**, the final class of a test instance is **Excluded**. This is a 3-vote scheme.

4. If all four votes are **Excluded**, the final class of a test instance is **Excluded.** This is a 4-vote scheme.

With the final prediction for each test instance, we computed the confusion matrix by comparing the predicted class with the actual class assigned by the reviewers, and we calculated the performance measures. When we combined of the classifiers, we gave each classifier an equal importance.

## 5 Evaluation Measures and Results

When performing the evaluation for the task of classifying an abstract into one of the two classes **Included (relevant)** or **Excluded (non relevant)**, two objectives are of great importance: **Objective 1** - ensure the completeness of the systematic review (maximize the number of relevant documents included); **Objective 2** - reduce the reviewers' workload (maximize the number of irrelevant documents excluded).

We observe that objective 1 is more important than objective 2 and this is why we decided to report recall and precision for the **Included** class. We also report F-measure, since we are dealing with imbalanced data sets.

Besides the standard evaluation measures, we report WSS[8] measure as well in order to give a clearer view of the results we obtain.

As baseline for our methods we consider: two extreme baselines and a random-baseline classifier that takes into account the distribution of the two classes in the training data set. The baselines results are: *Include_All* – a baseline that classifies everything in the majority class: Recall = 100%, Precision = 15%, F-measure = 26.2%; WSS = 0% *Exclude_All* – a baseline that classifies everything as **Excluded**: Recall = 0%, Precision = 100%, F-measure = 64.2%; WSS = 0% *Random baseline*: Recall = 8.9%, Precision = 15.4%, F-measure = 67.8%; WSS = 0.23%.

### 5.1 Results for the Global Method

In this subsection, we present the results that we obtained using our global method with the three representation techniques and CNB as classification algorithm. The results are reported for the test set that mentioned in Table 1. To get a clear image of the results, we show the numbers of the confusion matrix as well in Table 3. In this way

| | BOW | UMLS | BOW+UMLS |
|---|---|---|---|
| **True Inc.** | *2,692* | *2,793* | *2,715* |
| **False Inc.** | *5,022* | *8,922* | *5,086* |
| **True Exc.** | *18,135* | *14,235* | *18,071* |
| **False Exc.** | *1,425* | *1,324* | *1,402* |
| **Recall** | 65.3% | 67.8% | 65.9% |
| **Precision** | 34.9% | 23.8% | 34.8% |
| **F-measure** | 45.5% | 35.2% | 45.5% |
| **WSS** | 37.1% | 24.9% | 37.3% |

**Table 3**. Results for the global method.

[8] WSS = (TE + FE)/(TE + FE + TI + FI) – 1+ TI/(TI + FE) where T stands for true; F – false I – Included class; E- Excluded class.

the reader can understand the workload reduction when using classifiers to help the process of building systematic reviews.

The BOW features were identified following the guidelines presented in Section 3.4 and a number of 23,906 features were selected. To determine the UMLS concepts we used the MetaMap system described earlier in the paper. From the whole training abstracts collection, a number of 459 UMLS features were identified. Analyzing the results from Table 5, in terms of recall, the UMLS representation obtained the best recall results, 67.8% for the global method but much lower precision, 23.8% than BOW representation, 34.9%. The hybrid representation, BOW+ UMLS features had similar results with the BOW alone. Recall increased a bit for the hybrid representation compared to BOW alone, 0.6% but its value is still not acceptable. We conclude that the levels of recall, our main objective for this task, were not acceptable for a classifier to be used as replacement of a human judge in the workflow of building a systematic review. The levels of precision that we obtained with the global method are acceptable but they cannot substitute the low level of recall. Since our major focus is recall, we investigated more and we further improved our precision scores with the per-question classification method, whose results are discussed in the next subsection.

#### 5.1.1 Results for Feature Selection

Table 4 presents the results obtained with our feature selection techniques. We decided to report only representative results using CNB as a classifier and a specific representation setting. The number of features used in the experiment is presented in the round brackets. The first number represents the number of features extracted from the **Included** class data set while the second from the **Excluded** class data set.

| | **$Chi^2$ (150:150)** | **InfoGain (300)** | **BNS (10:8)** |
|---|---|---|---|
| **True Inc.** | *3,819* | *3,875* | *2,690* |
| **False Inc.** | *19,233* | *19,638* | *13,905* |
| **True Exc.** | *3,924* | *3,518* | *9,253* |
| **False Exc.** | *298* | *242* | *1,427* |
| **Recall** | 92.8% | 94.1% | 65.3% |
| **Precision** | 16.6% | 16.5% | 16.2% |
| **F-measure** | 28% | 28% | 25% |
| **WSS** | 8.2% | 7.9% | 4.5% |

**Table 4.** Representative results obtained for various feature selection techniques.

As we can see from the results, almost all abstracts were classified as **Included** resulting in good recall measures for the **Included** class, but at the same time a poor precision. We want to achieve good results for both measures since a low precision for the **Included** class will translate in more human effort in the next phase.

Similar experiments were performed when using Naïve Bayes as classifier. The results obtained were opposite to ones obtained for CNB, all abstracts were classified as **Excluded**. We believe that this is the case because the CNB classifier tries to compensate for the class imbalance and gives more credit to the minority class, while the Naïve Bayes classifier will let the majority class overwhelm the classifier, resulting in almost all abstracts being excluded.

Besides all the results presented in Table 4, we also tried to boost the representative features for the **Included** class hoping to re-balance the imbalance present in the training data set. To perform these experiments we selected the top k $CHI^2$ word features and then added to this set of features the top $k_1$ $CHI^2$ representative features only for the **Included** class. The parameter k varied from 50 to 100 and the parameter $k_1$ from 30 to 70. We performed experiments when using the original imbalanced training data set and using a balanced data set as well, with both CNB and Naïve Bayes classifier. The results that we obtained for these experiments were similar to the ones when we used the previous feature selection techniques. There was no significant difference in the results compared to the ones in Table 5.

### 5.2 Results for the Per-Question Method

The results for our second method using the four voting schemes are presented in Table 5.

Compared with the global method, the results obtained by the per-question method, especially the ones for 2 votes are the best so far in terms of the balance between the two objectives. A large number of abstracts that should be excluded are classified as **Excluded**, whereas wrongly excluding very few abstracts that should have been included (a lot fewer than in the case of the global classification method).

The 2-votes scheme performs better than the 1-vote schemes, because of potential classification errors. When the classifiers for two different questions (that look at two different aspects of the systematic review topic) are confident that the abstract is not relevant, the chance of correct prediction is higher; a balance between excluding an article and keeping it as relevant is achieved.

When using the classifiers for 3 or 4 questions, the performance goes down in terms of precision; a higher number of abstracts get classified as **Included** because some abstracts do not address all target question of the review topic.

| 1-Vote | BOW | UMLS | BOW+UMLS |
|---|---|---|---|
| **True Inc.** | *1,262* | *1,222* | *1,264* |
| **False Inc.** | *745* | *2,266* | *741* |
| **True Exc.** | *22,412* | *20,891* | *22,416* |
| **False Exc.** | *2,855* | *2,895* | *2,853* |
| **Recall** | 30.6% | 29.6% | 30.7% |
| **Precision** | 62.8% | 35.0% | 63.0% |
| **F-measure** | 41.2% | 32.1% | 41.2% |
| **WSS** | 23.2% | 16.8% | 23.3% |
| **2-Vote** | **BOW** | **UMLS** | **BOW+UMLS** |
| **True Inc.** | *3,181* | *2,603* | *3,283* |
| **False Inc.** | *9,976* | *9,505* | *10,720* |
| **True Exc.** | *13,181* | *13,652* | *12,437* |
| **False Exc.** | *936* | *1,514* | *834* |
| **Recall** | 77.2% | 63.2% | 79.7% |
| **Precision** | 24.1% | 21.5% | 23.4% |
| **F-measure** | 36.8% | 32.0% | 36.2% |
| **WSS** | 29.0% | 18.8% | 28.4% |
| **3-Vote** | **BOW** | **UMLS** | **BOW+UMLS** |
| **True Inc.** | *3,898* | *3,480* | *3,890* |
| **False Inc.** | *18,915* | *16,472* | *18,881* |
| **True Exc.** | *4,242* | *6,685* | *4,276* |
| **False Exc.** | *219* | *637* | *227* |
| **Recall** | 94.6% | 84.5% | 94.4% |
| **Precision** | 17.0% | 17.4% | 17.0% |
| **F-measure** | 28.9% | 28.9% | 28.9% |
| **WSS** | 11.0% | 11.3% | 11.0% |
| **4-Vote** | **BOW** | **UMLS** | **BOW+UMLS** |
| **True Inc.** | *4,085* | *3,947* | *4,086* |
| **False Inc.** | *21,946* | *20,869* | *21,964* |
| **True Exc.** | *1,211* | *2,288* | *1,193* |
| **False Exc.** | *32* | *170* | *31* |
| **Recall** | 99.2% | 95.8% | 99.2% |
| **Precision** | 15.6% | 15.9% | 15.6% |
| **F-measure** | 27.1% | 27.2% | 27.0% |
| **WSS** | 3.7% | 4.8% | 3.7% |

**Table 5**. Results for the per-question method for the **Included** class.

For the per-question technique the recall value peaked at 99.2% with the 4-vote method BOW and BOW+UMLS representation techniques when using CNB as classifier. In the same time the lowest values of precision for the per-question technique, 15.6% is obtained with the same experimental setting. It is important to aim for a high recall, but not to dismiss the precision values. The difference of even less than 2% in precision values can cause the reviewers to read additional thousands of documents, as observed in the confusion matrices for 2-vote, 3-vote and 4-vote methods in Table 5.

From the confusion matrix in Table 5 for the 2-vote method and the 3- and 4-vote method we observe the high difference in the number of documents a reviewer will have to read (the falsely included documents). The difference in precision from 24.1% for the 2-vote method to 15.6% for the 4-vote method makes the reviewer go through 11,988 additional abstracts.

The best value for the WSS measure for the per-question method is achieved by the 2-vote scheme. The result it is lower than the one obtained by the global method but the recall level is higher than for the global method. Therefore, we still keep as a potential winner the 2-vote scheme for the per-question classification technique.

### 5.3 Results for Human-Machine Workflow

In Figure 1, we envisioned the way we can use the automatic classifier in the workflow of building a systematic review. In order to determine the performance of the human-machine workflow that we propose, we computed the recall values when the human reviewer's labels are combined with the labels obtained from the classifier. The same labeling technique is applied as for the human-human workflow: if at least one decision for an abstract is to include it in the systematic review, then the final label is **Included**.

We also calculated the evaluation measures for the two reviewers. The evaluation measures for the human judge that is kept in the human-machine workflow, Reviewer 1 in Figure 1, are 64.29% for recall and 15.20% for precision. The evaluation measures for the reviewer that is to be replaced in the human-machine classification, Reviewer 2 in Figure 1 are 59.66% for recall and 15.09% for precision. The recall value for the two human judges combined is 85.26% and the precision value is 100%. As we can observe, the recall value for the second reviewer, the one that is replaced in the human-classifier workflow is low. Like the conceptual idea that stands behind all committees of judges, the power of many is stronger than the power of one. The results that we obtain when both reviewers are used are much higher than each of the results of them individually. In Table 6, we present precision and recall results for the symbiotic model, for both our methods. In these results we can clearly see that the 2-vote technique is superior to the other voting techniques and to the global method. For almost the same level of precision, the level or recall it is much higher. These observations support the fact that the extra effort spent in identifying the most suitable methodology pays off.

The fact that we keep a human in the loop makes our method acceptable as a workflow for building a systematic review. We always aim for a high level of recall while keeping in mind that a good level of precision is important; low precision means more human effort in the next stage of the systematic review.

| Method | BOW | UMLS | BOW+ UMLS |
|---|---|---|---|
| **Global** | 17.9/87.7% | 17.0/88.6% | 17.9/87.7% |
| **1-Vote** | 17.1/75.3% | 16.5/74.8% | 17.1/75.4% |
| **2-Vote** | 17.1/91.6% | 16.4/86.6% | 17.1/92.7% |
| **3-Vote** | 15.8/97.9% | 15.8/94.2% | 15.8/97.8% |
| **4-Vote** | 15.3/99.6% | 15.4/98.3% | 15.3/99.6% |

**Table 6**. Precision/recall results for the human-classifier workflow for the **Included** class.

## 6 Discussion

The global method achieves good results in terms of precision while the best recall is obtained by the per-question method.

The best results for the task were obtained using the per-question method with the 2-vote scheme. The results obtained by the 3-vote scheme UMLS representation are close to the 2-vote scheme, but looking at F-measure and WSS results the 2-vote scheme outperforms the 3-vote one. The clear distinction between the methods comes when we combined the classifiers with the human judge in the workflow of building reviews. The 2-vote scheme with or without UMLS features is the best method.

The per-question technique is more robust and it offers the possibility to choose the desired type of performance. If the reviewers are willing to read almost the entire collection of documents, knowing that the recall is high, then a 3 or 4-vote scheme can be the set-up (though the 3 or 4-vote method is not likely to achieve 100% recall because it is very rare that an abstract contain answers to three or four of the questions associated with the systematic review). If the reviewers will like to read a small collection being confident that almost all the abstracts are relevant, then a 1-vote scheme can be the set-up required. The per-question method confirms the fact that a committee or an ensemble of classifiers is better than one classifier; this conclusion is supported in the machine learning literature (Dietterich, 1997).

When we combine the human and the system results we obtain a major improved in terms of recall. We base our discussion for the human-machine results for the experiment that obtained the best results, the 2-vote scheme with a BOW+UMLS representation technique. When combining the human and classifier decisions, the precision level decreased a bit compared to the one that the machine obtained. We believe that this is the case because some of the abstracts that the classifier excluded were included by the first human reviewer and, with this decision process in place, the level of precision dropped.

Our goal of improving the recall level from the first level of screening is achieved, since when both the classifier and the human judge are integrated in the workflow, the recall level jumps from 79.7% to 92.7%.

We believe that the low level of precision that is obtained for the human reviewer, for the human-classifier workflow, and for the classifier, is due to the fact that we are running experiments for the first screening phase. In the next screening phase the entire article is available and more informed decisions can be made by the judges.

We believe that further investigations are required to fully replace a human reviewer with an automatic classifier but the results that we obtained with the per-question method encourage us to believe that this is a suitable solution for reaching our final goal.

## 7 Conclusions and Future Work

In this paper, we looked at two methods by which we envision the way automatic text classification techniques could help the workflow of building systematic reviews.

The first method is a straight-forward application of the representations and learning algorithms that capture the specifics of the data: medical domain, huge number of features, misclassification, and imbalanced of classes.

We showed that the specifics of the human protocol in which systematic reviews are built have a positive effect when deployed in an automatic way. We believe that the tedious process that is currently used for building systematic reviews can be lightened by the use of a classifier in combination with only one human judge. By having a human judge in the loop, we ensure that the workflow is reliable and that the system can be easily integrated in the workflow.

In future work we would like to look into ways of improving the results by the way we chose the training data set and by integrating more domain specific knowledge. We would also like to investigate ways by witch we can can update systematic reviews.

## References


Aphinyanaphongs Y. and Aliferis C. *Text Categorization Models for Retrieval of High Quality Articles.* Journal of the American Medical Informatics Association 2005; 12:207-216.

Cohen A.M. *Optimizing Feature Representation for Automated Systematic Review Work Prioritization.* Proceedings of the AMIA Annual Symposium 2008; 6:121-126.

Cohen A.M., Hersh W.R., Peterson K., Yen P.Y. *Reducing Workload in Systematic Review Preparation Using Automated Citation Classification.* Journal of the American Medical Informatics Association 2006; 13:206-219.

Dietterich, T. *Machine-Learning Research: Four Current Directions*. Artificial Intelligence Magazine. 18(4): 97-136 (1997)

Drummond C. and Holte R.C. *C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling.* Proceedings of the Twentieth International Conference on Machine Learning: Workshop on Learning from Imbalanced Data Sets (II), 2003.

Forman G. *Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification.* In the Joint Proceedings of the 13th European Conference on Machine Learning and the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), 2002.

Frank E. and Bouckaert R.R. *Naive Bayes for Text Classification with Unbalanced Classes.* In the Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, 2006, pp. 503-510.

Haynes R.B., Wilczynski N., McKibbon K.A., Walker C.J., Sinclair J.C. *Developing optimal search strategies for detecting clinically sound studies in MEDLINE.* Journal of the American Medical Informatics Association 1994; 1:447-58.

Kohavi R. and Provost F. *Glossary of Terms*. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process 1998; 30:271-274.

Ma Y. 2007. Text classification on imbalanced data: Application to Systematic Reviews Automation. M.Sc. Thesis. University of Ottawa.