Text Classification for literature search of research study designs

By: Eddie Daniel Galarza
Supervisor: Dr. Karin Verspoor

25 Credit Points Research Project
ABSTRACT

Information search is used every day in many fields of the human labors. The information found is needed as evidence for making decisions. In healthcare, information has also become a necessity to find the best path to invest into a health program. Some organizations like the Health Technology Assessment have developed alternatives to find this information provided by life sciences and biomedical records. A manual technique could give the support for finding concrete information like “health economic evaluations” by looking at the terms contained in the records. However, this technique does not provide high levels of accuracy.

In this research it is proposed an alternative to contrast manual techniques by using an automatic machine learning approach. Classification is the method from machine learning used to achieve the information search task. It has been proven that it can give acceptable levels of accuracy as well as providing proof that terms analysis can be used to determine the category of a record. Moreover, the use of a good classification algorithm and different combinations of features could give an efficient model that contrasts a manually developed technique.
I certify that

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

-the thesis is 7643 words in length (excluding text in images, table, bibliographies and appendices).
Contents

Introduction ................................................................................................................................. 8

1.1 Context and Motivation ...................................................................................................... 8

1.2 Aim and Scope of the Thesis ............................................................................................ 9

1.3 Outline of the Thesis ......................................................................................................... 10

The Dataset: Medline Health Economics Evaluations .............................................................. 11

2.1 Medline and PubMed ....................................................................................................... 11

2.2 Health Economic Evaluations ........................................................................................ 12

2.3 Medline Articles Structure and Selected Sections .......................................................... 13

Methodology and Feature Engineering ................................................................................... 18

3.1 Data Extraction ................................................................................................................ 18

3.2 File and Feature Pre-processing in Weka ........................................................................ 19

3.3 Defining the Training, Developing and Testing Data ....................................................... 20

3.4 Performing the classification ............................................................................................ 21

Experimentation, Results and Discussion ............................................................................. 23

4.1 Experimentation and Results ........................................................................................... 23

4.1.1 First Set of Experiments ............................................................................................. 23

4.1.1 Second Set of Experiments ......................................................................................... 27

4.1.3 Third Set of Experiments ........................................................................................... 29

4.1.4 Best Features in the Best Classification Model .......................................................... 30

4.4 Discussion ......................................................................................................................... 31

Conclusions and Future Work ............................................................................................... 33

4.1 Conclusions ..................................................................................................................... 33

4.2 Future Work ..................................................................................................................... 34
List of Tables

Table 2.1 Medline records fields and abbreviations............................................. 15

Table 4.1 Results for experimentation one for Naïve Bayes algorithm with unigrams, and unmodified terms........................................................................................................ 24

Table 4.2: Results for experimentation one for J48 algorithm with unigrams, and unmodified terms........................................................................................................ 24

Table 4.3: Results for experimentation one for SVM algorithm with unigrams, and unmodified terms........................................................................................................ 25

Table 4.4: Best results for experimentation one for SVM algorithm with bigrams, and unmodified terms........................................................................................................ 25

Table 4.5: Best results for experimentation one for SVM algorithm with bigrams, and stemmed terms........................................................................................................ 26

Table 4.6: Best results for experimentation one for SVM algorithm with bigrams, and stemmed terms........................................................................................................ 26

Table 4.7: Best results for experimentation one for SVM algorithm with bigrams, and removed stop words........................................................................................................ 27

Table 4.8: Best results for experimentation two with the first training data and testing on the development data........................................................................................................ 28

Table 4.9: Best results for experimentation two with the second training data and testing with the final testing data........................................................................................................ 28

Table 4.10: Best results for experimentation three with the first training data and testing on the development data........................................................................................................ 29

Table 4.11: Best results for experimentation three with the second training data and testing with the final testing data........................................................................................................ 30
List of Figures

Figure 2.1: PubMed Medline Sample Record......................................................... 16
Figure 3.1: Part of the code to extract MeSH terms........................................... 18
Figure 3.2: StringToWordVector filter option in Weka....................................... 20
Figure 4.1: Ranked attributes weight in the best classification model.................. 30
Chapter 1
Introduction

1.1 Context and Motivation

At the present time, lot of information coming from data is collected from many fields. This is why many computational tools have been created to help humans in extracting this information and get valuable knowledge. The analysis of data could be done for example in the health-care industry, where specialists provide reports to the organizations sponsoring a patient or a group of patients. Additionally, these reports could be analyzed to satisfy other needs like making decisions to improve relationships with customers. But humans for themselves are not able to analyze thousands of records which will be stored in a database. The need of a computational tool makes way in the human environment to provide analysis, as well as understanding the data that belongs to a report, this data will become information.

In the health care industry, the analysis of data has become popular and essential. The analysis of the data could for example, help insurers detect any fraud or abuse, help physicians identify better and effective treatments. Also this analysis could help identifying records of a certain type like an economic evaluation. This is where the thesis will focus, it will provide an alternative for identifying types of health related articles, and more specifically identifying health economics evaluations. This identification will be combined with a computational tool to improve the effectiveness of finding those types of reports.

An economic evaluation is made whenever it is necessary to make a decision on an investment. It is made to know if such investment is appropriate and feasible. The decision-making falls into the hands of high corporative members of an organization. For the health-care resources it should be observed if the correct policies are rational for the diffusion and the use of health interventions. It could for example determine the correct price to assign to a health product such as drugs. The correct evaluation could determine the value that the manufacturer will receive as well as making it reachable to the patients. (Le Pen, 2009).

The Health Technology Assessment (HTA) has developed various technological strategies to identify “health economics evaluations” records from other medical records. They found that the identification of economic studies and other economic information is often problematic, and the current economic evaluation and quality of life filters are not ideal. Hence, in 2009 with the support of the Canadian Agency for Drugs and Technologies in Health designed a system to perform the task of collecting relevant records. The research of this system was performed within the Medline and Embase databases which are bibliographic databases with life sciences and biomedical information. The research in question involved developing filters to recognize the “health economics evaluations”.
To develop the filters it was identified a set of know database records meeting the criteria of an economic evaluation. This set of relevant records were named as gold standards. They were identified by hand using human’s recognition. However, there was not enough resources to perform an extensive hand-searching, hence, it was used a relative recall. Complementary, it was also collected a set of records to discriminate economic evaluations from other economic publications. This set of records were called comparator records. When the set of testing records was achieved, a word occurrence analysis was performed. The occurrence or absence of the candidate terms on both kind of records was compiled into a table. Later, they were arranged into an extensive analysis with structures called “classification trees”. This structures used a statistical algorithm to identify terms that could be discriminated from the groups of records. This would allow to define differences between records and allow the type to be identified. Each tree then provided different nodes which would be later simplified to deliver the final filter.

However in their results, they failed to deliver a filter with moderate levels of recall (more than 0.95) and precision (more than 0.5). In fact, the experiments delivered uneven precision and recall levels. Some experimentation delivered a better precision, while other delivered a better recall. With the obtained results it was suggested that text words and indexing terms in economic evaluation records do not sufficiently discriminate economic evaluation record from the other records that deal with economic issues in health care. As result, it was concluded that researchers cannot rely on a few precise search terms to identify economic evaluations. And a possible help to identify the health economics evaluations was to encourage authors to define their papers with clear and relevant terminology to identify them more easily. (Glanville, Fleetwood, Yellowlees, Kaunelis, Mensinkai, 2009)

As a matter of fact, it raises the necessity to find an alternative approach to identify more easily information with a certain type of context. It could demand another mechanism, faster and with a lower error rate than human skills could provide. As data could be analyzed by different techniques, a machine learning approach could have the solution to this task. An automatic technique that learns by example could make use of the terms in a set of documents to predict which document is a health economic evaluation, and which is not. More specifically, classification could determine the category, class or type of document by analyzing the term occurrences in a set of records. The performed research in this thesis provides the alternative technique by classifying the records and automatically determining which of them are relevant, it gives a different perspective by using the word terms which the records are composed of. The obtained results are better than the obtained by the HTA and provide knowledge for further development of a “Health Economic Evaluations Recognition System”, as well as more initiatives to recognize other types of medical records.

1.2 Aim and Scope of the Thesis

The present thesis and the development of this project has the objective to contrast the manual strategies for development of queries for information search with a more data-driven approach. This is finding an alternative to the filters developed by HTA as well as finding
with good accuracy relevant records with a specific healthcare topic, this is the “health economic evaluations”.

The scope of this thesis goes from pre-processing the healthcare records to obtaining an efficient classification model with best precision and sensitivity. This implies processing the provided records as well as making experiments with different arrangement of terms and features. Also, an important goal to achieve is to compare the automatic approach versus the manual approach to obtain information.

As for limitations of the project, it has been provided with uneven quantity of positive labeled articles and negative labeled articles. This makes the classification accuracy outmatch whichever category has a major quantity of labeled articles. The usage of only three classification algorithms limits the opportunities to have a better comparison among classifying algorithms. Finally, the usage of a personal computer to make the experimentation, limits the quantity of word tokens that could be used to create better models with more combinations of features and combinations of terms arrangements.

1.3 Outline of the Thesis

This section describes the following chapters of this thesis:

Chapter 2 describes the composition of the provided data to train the classification model, these data comes from medical articles. They are extracted from an extensive database with life sciences and biomedical literature called Medline. Also, the selection of the features will be analyzed on how they provide information to the “health economics evaluations” as well as the reasons to select them as relevant features for the classifier.

Chapter 3 describes in detail how the project was developed. The methods used for extracting the data and the analysis with the computational tool are described. The feature engineering is also described in this chapter.

Chapter 4 shows the results with its different experimentations, followed by the explanation on the different methods used to get them. In this section, some particular examples will be analyzed. Also, a discussion on how different combinations of features and terms arrangement have impact on the consequence of the results. Here, the obtained results will be compared to the results obtained by the filtering approach performed by HTA on Medline and Embase records, and the differences between the classification approach and the filtering approach.

Finally, Chapter 5 describes the conclusion obtained for this thesis, the future improvements for this research, and the experimentations that is missing to provide new results and the possibility of a more efficient way to identify different types of medical and non-medical topics.
Chapter 2

The Dataset: Medline Health Economics Evaluations

2.1 Medline and PubMed

Medline is a bibliographic database that belongs to the National Library of Medicine (NLM) in the United States. It contains references to journals articles in life sciences and biomedicine. Most articles and journals are selected from the recommendation of the Literature Selection Technical Review Committee (LSTRC). Additional journals are selected based on the NLM-initiated reviews in which they provide their distinctive indexing, the Medical Subject Headings (MeSH). Some titles like history of medicine, health services research, health economics, molecular biology, are examples of the journals contained in the database. The performed reviews involve the consultation of the array on the National Institute of Health (NIH) and some external organizations who have a special collaborative arrangement.

Medline is the fundamental part of PubMed, part of the Entrez databases provided by the NLM National Center for Biotechnology Information (NCBI). This databases contain more than 5,600 worldwide journals in about 40 languages. The scope for Medline is biomedicine and health, it covers areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and people engaged in research. Aspects of biology, environmental science, and marine biology are also covered by the Medline databases. It is available from the NLM homepage (http://www.nlm.nih.gov) and have an indication of free electronic full-text availability. Searching do not require registration.

The resource that gives access to Medline is PubMed. It contains 25 million records (articles) in the biomedical literature and some items from the NCBI Books database. Most records are from Medline, however it also contains articles that are currently being processed to be part of the Medline citations. PubMed doesn’t contain full-text records, but the record in mention could contain information about the Publisher and resources like its web site. Anyway more than 33 percent of the records in the last five years contain free full-text. (Nlm.nih.gov, 2015)
2.2 Health Economic Evaluations

Public Health practitioners care about health of the entire population. But in order to do so, they need to make good use of resources as they are not unlimited. All public health programs, policies, interventions have costs or financial investments. For this reason it has to be a decision on which programs to invest. Information is needed about the cost of potential strategies as well as their effectiveness to ensure a maximum value of the resources. It is needed a value return on society’s investments in public health.

An economic evaluation provide a mechanism to identify, measure, value and compare the charges and impacts of the programs, polices, and interventions. (Dunet, 2012). In a health area, a health economic evaluation could be performed for example in a clinical surgery program in which it reflects the “efficacy” and “effectiveness” of such investment. It should also be designed according to a scientific state-of-the-art knowledge. This means, it should be clearly defined, the hypotheses must be derived precisely and the methodology must be presented and justified. Furthermore, the evaluation look at the investment calculations and it is necessary to choose an appropriate perspective for the economic analysis. This translates in using the appropriate cost analysis, a cost effectiveness analysis, a cost benefit analysis and a cost utility analysis.

The health economic analysis can be classified depending of the type of assessment and the price impact. From this perspective, the outcome varies from non-assessment through assessment in non-monetary, naturalistic units to monetary assessment. Besides it will depend on which is the objective of the analysis and it will always have a justification.

The first method is cost-minimization analysis in which the therapeutic alternatives with same effectiveness will be compared in terms of the net cost to find the cheapest alternative. The second method involves the cost-effectiveness analysis. Here, the prices are assessed in monetary units and qualified as non-monetary but utility-adjusted outcome. This method is related to the life expectancy and the quality of life. The last method is the cost – benefit analysis. This method analyses all effects, for example the health effects in monetary units. The analysis should be performed in a methodologically manner so this is a disadvantage of this method, making it difficult to perform.

Another element that should be taken into account for the analysis is the perspective. A perspective involves the point of view from which the costs and benefits are summarized and determined. It must be derived from the research question. Also the societal/economic perspective is important because it represents the most comprehensive approach. Some other perspectives are also taken into account for the analysis. For example, the health system, the social insurance and the service providers like hospitals. Each perspective must be justified to be used in analysis. If there is the need of use of several perspectives, then the results will be presented for each study perspective.

The outcome parameters depend on the indication and the research question. Clinical outcome parameters include physiological or biochemical, morbidity or mortality related
parameters. Other parameter can be the quality of life which uses an appropriate variable to measure the treatment alternatives. As result, this outcomes are used for comparison of different evaluation studies and validates the literature of the health economic evaluation.

Finally, the presentation of results are reproduced transparently. They are presented in the same way as for publication in journals (detailing title, author, sponsoring, etc.). If there are negative results, they are also published. The different points of view of the presentation are shown comparatively. Additionally, it has a clear description of the cost-effective strategy used to obtain those results. (Walter and Zehetmayr, 2006)

2.3 Medline Articles Structure and Selected Sections

As mentioned before, Medline contains bibliographies authorized by the National Library of Medicine in U.S. They are in charge of creating the bibliographic standards for the publications, and also apply the standards to complex biomedical material. This standards are regulated by the National Information Standards Organization (NISO). (Patrias and Wendling, 2007). Every article in the Medline/PubMed database follows the Introduction, Methods, Results, and Discussion (IMRAD) format. This sections are fundamental part in a scientific paper. The included elements that are part of the articles are: Title, Author (s), Keywords, Corresponding Author, Financial & Equipment Support, Abstract, Introduction, Methods, Results, Discussion, Conclusions, Acknowledgements and References.

A reference of what every section of the article is meant for is given below:

- The Title gives an idea to the reader of what to expect in the paper
- The Author gives information about the person who wrote the article. There could be one or two primary authors. The remaining authors are reviewers of the performed work.
- Keywords given according to the journal of the article.
- The Corresponding Author gives information about the principal author of the article which will be answering any question about the research.
- Financial and Equipment Support details the organizations, agencies or companies that supported the research.
- The Abstracts are the synopsis and summary of the paper. It has become and standard for any scientific article.
- The Introduction gives the background and the foundation of the article.
- The Methods describe the processes used in the study. It can specify the design, the strategy, the performed interventions, the data collection and treatment and outlines the analysis of such methods.
- The Results show what was found in the study. It gives a discussion of the gotten results in reference to previous research. Also details the strongest points in the study as well as the limitations.
• The conclusions states a summary of the results gotten in the study and includes key facts to explain them.
• The Acknowledgements introduces people who contributed in the work but in a shallow manner.
• The references states the citations for articles and materials referenced in text of the article. (Health Sciences Libraries, 2015).

Despite the structure used by most medical research papers, the Medline dataset provided for the thesis in question are composed by the sections detailed before and some unique fields created for the use of the Medline records. Some of this elements could not appear in every record. The fields can be searched in the PubMed Web Database. Table 2.1 shows the additional fields and its abbreviation.
Table 2.1: Medline records fields and abbreviations

The provided data to train the classification system is composed of the abbreviations and the information about the article for that field. Figure 2.1 shows an example of an article in one text file that complies the different labeled articles.
To perform the objective of this thesis, librarians have provided two text files which contain manually classified Medline records. There are 1949 records placed in a text file which are labeled as positive, specifically, this articles are nominated as “health economics evaluations”. In contrast, the other text file contains 4134 records which are labeled as negative, in other words it contains non relevant records with non-related medical topics.

The selected fields to support the achievement of the project are the title of the article, the title of the journal, the medical subject heading (MeSH), and the abstract. The title of the article is the main field in the record. With few words it informs what is expected in such article. The title of the journal is selected because it encompasses the complete subject and professional activity of the article in question. The MeSH headings contains medical vocabulary and terms related to the context of the article. Finally, the abstract contains the summary of the whole article. The group of words that conforms the abstract give the idea of its context. This fields contain valuable information to determine if an article is a “health
economic evaluation” or not. Additionally, the PubMed ID of the record is used to give information of what terms belong to a particular record.
Chapter 3
Methodology and Feature Engineering

3.1 Data Extraction

A set of records have been provided to train, develop and test the system. The records cannot be used as they are, they need to be pre-processed to be used with a computational tool. As a machine learning approach, the computational device needs to learn by example which documents contain terms that probe a record as certain type. For this reason, a set of experiments are performed to obtain a model which will be the guideline to determine the complexion of subsequent documents.

To begin the project, the data had to be extracted from the files in which they were delivered. There were two text files regarding positive and negative labeled records. It was processed with a Python Script, in which it was used regular expressions to extract the different features, regarding the different fields that the record is composed. As seen in figure 2.1, the data was in a Medline format, it made easier to introduce a regular expression which compares the terms in every line. Figure 3.1 shows part of the code involved to extract Mesh terms from the record.

```python
if ("MH -" in verify[x] and MH == 1):
  kms = 1
for y in range(x, len(verify)):
  if ("=" not in verify[y]) and ("KDA1" not in verify[y]) and ("KDA2" not in verify[y]):
    process = re.sub("[a-d]%", "attpercentage", verify[y].lower())
    process2 = re.sub("[a-d]%", "attnumber", process)
    process3 = re.sub("[", ", process2)
    process4 = re.sub("[", ", process)
    process5 = re.sub("[", ", process)
    process6 = re.sub("[", ", process)
    process7 = re.sub("[", ", process)
    process8 = re.sub("[", ", process)
    process9 = re.sub("[", ", process)
    process10 = re.sub("[", ", process)
    kdouble = kdouble + process10
  else:
    MH = 0
  break
```

Figure 3.1: Part of the code to extract MeSH terms

The goal of the script was to extract the text without the inclusion of selected symbols or numbers, with words in lowercase and differentiated from any other field. This fields will be later known as “the features”. For example a term of the MeSH field is “MH-pathology”. In this case “MH-” is the distinctive. However, the terms of each field were also extracted in a natural form (without the distinctive) except for the abstract field which was required to
maintain every term in its natural form. Also, some numerical tokens were modified by changing them for the word “attnumber” as well as changing percentages for the word “attpercentage”.

All this terms were saved into different text files, named with a distinctive word and the PubMed identification that it belongs. For example “mesh5332342” refers to mesh terms present in article number 53322342. Every set of files was contained in a folder regarding to the positive or negative labeling that it was given. This arrangement of the files was done for later processing made in the data mining tool, Weka.

### 3.2 File and Feature Pre-processing in Weka

The computational tool that will handle the extracted data is Weka. This is a datamining tool that provides different features for analysis and classification. When the data is ready to be processed, a classification algorithm provided by the computational tool will be used to create a model. In this tool, there are several groups of algorithms that could be used in the project’s classification approach. However, Weka cannot process text files, it has a special unique format with the extension attribute relation format file (ARFF).

To make the data “comprehensible” for Weka, it provides conversion tools as well as merging and appending tools. The tool “TextDirectoryLoader” was used to convert the files contained in a folder into a set of compatible data for Weka. Also, the name of the folder that contains the files will be the associated class for the group of files. Furthermore, the merging and appending tools was used to integrate all the features of the data into one solid arff file to perform the different experimentations.

The following step, after the arff file was obtained, was to select the features and recognize the terms contained within them. For this task, Weka offers different filters for data treatment. The used filter to manage the terms in the different features was StringToWordVector, which is an unsupervised attribute filter. It can remove symbols, remove stop words and perform stemming. As its name says, it transforms the string into a word vector in which a TF-IDF operation could also be performed. Also, it can perform word tokenizing, it means that it can construct, unigrams, bigrams and k-grams with the terms. For the extents of this project, the texts string were processed as raw text, stopped text, stemmed text with unigrams, bigrams and trigrams. Also, TF-IDF was performed in the word vectors. Figure 3.2 shows the StringToWordVector Tool ready to perform a filtering by removing stop terms, making bigrams and performing the TF-IDF operation.
StringToWordVector keeps a certain number of terms on a basis of most frequent terms to be kept. For this reason, it was necessary to perform a proper feature selection. It was done by using the “select attributes (features)” option in the explorer panel of Weka. It was selected the “ranker” method which selects the top most terms in a performed evaluation. Obviously, it has another option to select a kind of evaluation, and the selected evaluation performed was the “FilteredAttributEval” which selects the terms with the best Information Gain in the string.

After the execution of the mentioned internal Weka tools, the data is ready to act as valid data to obtain the classification model, but not enough to perform a valid classifier experimentation. The next step is to select the training, development and testing data as well as selecting the algorithms to perform the classification.

3.3 Defining the Training, Developing and Testing Data

To understand the performance of the classifier, it must be evaluated. It is required to know the proportion of instances whose classes the classifier can correctly predict. To perform this task, the classifier should be built using one dataset, called the training set. This dataset will be evaluated on a different set called the test set, an independent set out of the training set.

In the extents of this thesis two evaluation methods have been chosen. The holdout method and the cross-validation method.
The holdout method is performed by dividing the dataset in three parts, one dataset will be used for training the model, the second will be used to develop the model, and the third one is used to test the final trained model which is a combination between the first training data and the development data. To achieve this goal, Weka has another tool to make sets of data contained in one arff file. This is the unsupervised instance filter “Resample”. This tool allows to divide the dataset in different percentages of data. Also, it provides options to do a normal selection, or an inverted selection. It means that if 60% is chosen as the normal selection, the inverted selection will be the remaining 40% of the data. Also it has the noReplacement option which allows to select different data without overlapping. For the extents of this thesis, the selected data for training was 60%, for developing was 20% and for testing was the remaining 20%. It is important to notice that each set of data should be saved in three different arff files and the three files will correspond to an experimentation, this means a file who has been filtered its stop words, will not be valid for another which has not.

The other evaluation method is cross-validation. In this method the dataset is partitioned into k subsets of equal size. Each subset is used in turn as the test set and it is tested with the remaining sets. The average measure of the accuracies from each “fold” will give the overall accuracy of the model. In the extent of the thesis, this evaluation method was performed by using the whole dataset and letting the program perform the cross-validation with 4 folds. Finally, the experimentation has to be executed with a criteria of which classifications algorithms to use.

### 3.4 Performing the classification

As the machine learning approach was selected to accomplish the objective, specifically the classifying method. A classifier determines the category in which an “observation” belongs. After the data has been pre-processed, the features will contain different terms that will be handled by a classification algorithm. In this context, this terms have different meaning that could affect the performance of the selected algorithm.

To construct a model, the classifier make use of methods which takes steps and the use of rules. The selected methods in this thesis are probabilistic. They use statistical inference to find the best class for the given data. (Aggarwal, n.d.) Furthermore, Naïve Bayes algorithm, J48 algorithm and Support Vector Machine (SVM) algorithm have been selected due to their simplicity, their efficiency and the fact that they are well known algorithms.

The first algorithm is Naïve Bayes. This algorithm has been proved several times as a successful algorithm due to the simplicity that its mathematical model implies. It is one of the most used algorithms for classification and it is known for giving good results. The success of this classifier is that in terms of zero-one loss (classification error) is not necessarily related to the quality of the fit to a probability distribution. The optimal classifier is obtained by an agreement of the most-probable class made by the estimated distributions. (Rish, 2001). Depending on the probability model, this classifier can be trained using a
supervised learning approach or a non-supervised approach. In many applications, the naïve Bayes algorithm uses the method of maximum likelihood to estimate its parameters.

The second algorithm comes from the Decision Trees, this is the J48 algorithm. The J48 algorithm is an implementation of the data mining tool used in this thesis, Weka. It uses the basis of the C4.5 decision tree learner. Also, it uses greedy technique to induce decision trees for classification, this is the reduced-error pruning. A decision tree uses a tree model over the attributes in the data set. It will try to split the data attributes in a manner that it leads to the best split. The quality of the split is computed from the correctly classified number of cases in each of the resulting nodes of one split. The best split will be found when all nodes become leaf nodes or when there cannot be made more splits. (Ozer, 2008). The advantage of using a classification tree is that it expect linear features and they can handle dimensional spaces as well as large number of training examples.

The third algorithm is a rule-based algorithm, this is the Support Vector Machine algorithm. SVMs are based on the Structural Risk Minimization principle. This means that it should find a hypothesis $h$ who can guarantee a lowest true error. This error will be the probability that $h$ will make an unseen error in a selected test data. The true error of a hypothesis $h$, the error of $h$ in a training data and the complexity of $H$ (the hypothesis space containing $H$) can be connected using an upper bound. Support vector machines find the hypothesis $h$ to minimize this bound of true error effectively and efficiently.

SVM works very well for text classification because it has a high dimensional input space. It means that when it learns a large quantity of features it use an overfitting protection. This protection make it able to handle the large feature spaces. The use of a feature selection setting determines these irrelevant features. Finally, the fact that most text categorization problems are linearly separable make SVM a good performer of text categorization. (Joachims, 1998).
Chapter 4

Experimentation, Results and Discussion

This section describes different experiments performed in this thesis, the results obtained by doing the experimentation as well as an explanation with some examples. Also, a comparing discussion between the automatic approach vs the manual approach developed by HTA will be performed.

4.1 Experimentation and Results

4.1.1 First Set of Experiments

The first set of experiments was performed with the full data set after being processed. This set of experiments was called “Experimentation One”. The evaluation method was the cross-validation, by performing 4 folds in the data. This validation was automatically performed by Weka. Also in this set of experiments it was performed different set of combinations among the features and its processing (removing stop words, stemming the terms), classifying algorithms and the combination of terms (unigrams, bigrams and trigrams). With each algorithm, the combination of features was as follows. Abstracts-Journals-Titles, Abstracts-Titles, Journals-Titles-MeshTerms, MeshTerms-Titles, and using all features. Also, the combinations of terms processing was: Raw text without modifications, Stemmed Terms, Stemmed and Stop words removed terms, and Only stop words removed terms. For this set of experiments, all the features are using the terms in a natural form. There are no distinctive between a term from the title and a term from the abstract.

Table 4.1, 4.2, 4.2 show the results obtained for the classification model, when performing experimentation with unigrams.

Table 4.1 shows the results when performing with a Naïve Bayes algorithm, the best results were obtained when performing the classification with all features.
Table 4.1: Results for experimentation one for Naïve Bayes algorithm with unigrams, and unmodified terms

Table 4.2 shows the results when performing with a J48 algorithm, the best results were also obtained when performing the classification with all features.

Table 4.2: Results for experimentation one for J48 algorithm with unigrams, and unmodified terms

Table 4.3 shows the results when performing with a SVM algorithm, the best results were also obtained when performing the classification with all features.
Table 4.3: Results for experimentation one for SVM algorithm with unigrams, and unmodified terms

From this set of experiments it was observed that the best performing algorithm, was SVM. The reason for SVM to perform better with the data set, is that the internal polynomial kernel performs better as the features are linear. Also, the fact that this is a binary classifier, makes it favorable for SVM to take advantage over the other algorithms. In terms of execution time, Naïve Bayes was the fastest performer, however it gives the worst results among the three classifying algorithms. Also, the J48 algorithm is the slowest performer. The reason for the slow performance, is that has to construct the decision tree, It takes time and internal memory to reach a leaf.

Table 4.3 shows the results for SVM when performing with unmodified word terms. Table 4.4, 4.5, 4.6 and 4.7 show the best performance of the SVM for the unmodified terms, also when stemming the terms, stemming and stopping the terms and only stopping the terms, respectively.

The best results for unmodified terms was with bigrams, this is shown in table 4.4.

Table 4.4: Best results for experimentation one for SVM algorithm with bigrams, and unmodified terms
Table 4.5 show the best results for SVM when stemming the terms. The best results for this term modification was with bigrams, and when performing the experiments with all features.

<table>
<thead>
<tr>
<th>SVM</th>
<th>ABS-JOUR-TITL</th>
<th>ABS-TITL</th>
<th>ALL FEAT</th>
<th>JOUR-TITLE</th>
<th>JOUR-TITL-M</th>
<th>MESH-TITL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td><strong>Precision</strong></td>
<td>0,84</td>
<td>0,841</td>
<td>0,856</td>
<td>0,834</td>
<td>0,815</td>
</tr>
<tr>
<td></td>
<td><strong>Recall</strong></td>
<td>0,825</td>
<td>0,825</td>
<td>0,837</td>
<td>0,808</td>
<td>0,819</td>
</tr>
<tr>
<td></td>
<td><strong>F-Measure</strong></td>
<td>0,832</td>
<td>0,833</td>
<td>0,846</td>
<td>0,821</td>
<td>0,817</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td><strong>Precision</strong></td>
<td>0,918</td>
<td>0,918</td>
<td>0,924</td>
<td>0,911</td>
<td>0,915</td>
</tr>
<tr>
<td></td>
<td><strong>Recall</strong></td>
<td>0,926</td>
<td>0,927</td>
<td>0,933</td>
<td>0,924</td>
<td>0,912</td>
</tr>
<tr>
<td></td>
<td><strong>F-Measure</strong></td>
<td>0,922</td>
<td>0,922</td>
<td>0,929</td>
<td>0,917</td>
<td>0,914</td>
</tr>
<tr>
<td></td>
<td><strong>Correctly Classified</strong></td>
<td>89,36</td>
<td>89,41</td>
<td>90,25</td>
<td>88,68</td>
<td>88,26</td>
</tr>
<tr>
<td></td>
<td><strong>Incorrectly Classified</strong></td>
<td>10,63</td>
<td>10,58</td>
<td>9,74</td>
<td>11,31</td>
<td>11,73</td>
</tr>
</tbody>
</table>

Table 4.5: Best results for experimentation one for SVM algorithm with bigrams, and stemmed terms

Table 4.6 show the best results for SVM when stemming the terms and removing stop words. The best results for this term modification was with trigrams, and when performing the experiments with all features.

<table>
<thead>
<tr>
<th>SVM</th>
<th>ABS-JOUR-TITL</th>
<th>ABS-TITL</th>
<th>ALL FEAT</th>
<th>JOUR-TITLE</th>
<th>JOUR-TITL-M</th>
<th>MESH-TITL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td><strong>Precision</strong></td>
<td>0,838</td>
<td>0,839</td>
<td>0,855</td>
<td>0,835</td>
<td>0,822</td>
</tr>
<tr>
<td></td>
<td><strong>Recall</strong></td>
<td>0,821</td>
<td>0,825</td>
<td>0,833</td>
<td>0,805</td>
<td>0,819</td>
</tr>
<tr>
<td></td>
<td><strong>F-Measure</strong></td>
<td>0,83</td>
<td>0,832</td>
<td>0,844</td>
<td>0,82</td>
<td>0,821</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td><strong>Precision</strong></td>
<td>0,917</td>
<td>0,918</td>
<td>0,922</td>
<td>0,91</td>
<td>0,915</td>
</tr>
<tr>
<td></td>
<td><strong>Recall</strong></td>
<td>0,925</td>
<td>0,925</td>
<td>0,933</td>
<td>0,925</td>
<td>0,917</td>
</tr>
<tr>
<td></td>
<td><strong>F-Measure</strong></td>
<td>0,921</td>
<td>0,922</td>
<td>0,928</td>
<td>0,917</td>
<td>0,916</td>
</tr>
<tr>
<td></td>
<td><strong>Correctly Classified</strong></td>
<td>89,19</td>
<td>89,31</td>
<td>90,1</td>
<td>88,65</td>
<td>88,52</td>
</tr>
<tr>
<td></td>
<td><strong>Incorrectly Classified</strong></td>
<td>10,8</td>
<td>10,68</td>
<td>9,8</td>
<td>11,34</td>
<td>11,74</td>
</tr>
</tbody>
</table>

Table 4.6: Best results for experimentation one for SVM algorithm with bigrams, and stemmed terms

Table 4.7 show the best results for SVM when removing stop words only. The best results for this term modification was with bigrams, and when performing the experiments with all features.

<table>
<thead>
<tr>
<th>SVM</th>
<th>ABS-JOUR-TITL</th>
<th>ABS-TITL</th>
<th>ALL FEAT</th>
<th>JOUR-TITLE</th>
<th>JOUR-TITL-M</th>
<th>MESH-TITL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td><strong>Precision</strong></td>
<td>0,84</td>
<td>0,841</td>
<td>0,856</td>
<td>0,834</td>
<td>0,815</td>
</tr>
<tr>
<td></td>
<td><strong>Recall</strong></td>
<td>0,825</td>
<td>0,825</td>
<td>0,837</td>
<td>0,808</td>
<td>0,819</td>
</tr>
<tr>
<td></td>
<td><strong>F-Measure</strong></td>
<td>0,832</td>
<td>0,833</td>
<td>0,846</td>
<td>0,821</td>
<td>0,817</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td><strong>Precision</strong></td>
<td>0,918</td>
<td>0,918</td>
<td>0,924</td>
<td>0,911</td>
<td>0,915</td>
</tr>
<tr>
<td></td>
<td><strong>Recall</strong></td>
<td>0,926</td>
<td>0,927</td>
<td>0,933</td>
<td>0,924</td>
<td>0,912</td>
</tr>
<tr>
<td></td>
<td><strong>F-Measure</strong></td>
<td>0,922</td>
<td>0,922</td>
<td>0,929</td>
<td>0,917</td>
<td>0,914</td>
</tr>
<tr>
<td></td>
<td><strong>Correctly Classified</strong></td>
<td>89,36</td>
<td>89,41</td>
<td>90,25</td>
<td>88,68</td>
<td>88,26</td>
</tr>
<tr>
<td></td>
<td><strong>Incorrectly Classified</strong></td>
<td>10,63</td>
<td>10,58</td>
<td>9,74</td>
<td>11,31</td>
<td>11,73</td>
</tr>
</tbody>
</table>

Table 4.7: Best results for experimentation one for SVM algorithm with bigrams, and stemmed terms
### Table 4.7: Best results for experimentation one for SVM algorithm with bigrams, and removed stop words.

Experimentation one gave a general outlook of the incoming results for the next sets of experiments. Finally, the best results for experimentation one were deducted as follows:

The support vector machine is the best algorithm for the data set, in all the cases the best results are obtained when using all features and the most combination of terms is by performing bigrams and by only removing the stop words. The best correctly classified percentage is 91.22%, with a precision of 0.86 and a recall of 0.85 for positive records. Also a precision of 0.93 and a recall of 0.93 for negative labeled records.

#### 4.1.1 Second Set of Experiments

The second set of experiments was performed with the general outlook obtained in experimentation one. For this experiments the evaluation method was the holdout validation. 60% data was used to train the model, 20% was used for development and 20% was used for testing. Taking into account the results from experiment one, the selected combinations were bigrams and removing stop words, trigrams and removing stop words, and bigrams and not modifying the terms. Also, this terms didn’t have any distinctive from every other features, they were overlapped words from each feature. In this experiments it was added 4 new combinations of features, and the use of the 4 best combinations from experiment one. RawFull, which means numbers were not removed or changed from abstracts and the modified abstracts, journals, MeSH Terms and Titles by themselves were the new set of features. Also, the combination of Abstracts-Journals-Titles, Abstracts-Titles, Journal-Titles were the combinations adopted from experimentation one.

The results were obtained by training the model and testing it first with the development data to have an idea of which one could be the best data to perform with the final testing data. Table 4.8 show the best results obtained in this case.
In this case the combination of the full features, using bigrams and only removing the stop words, gave the best results. The next results were obtained by appending the best development data with the first training data. This was done for all the combinations to determine if any of them could have better result when testing the model with the final test data. Table 4.9 shows the results obtained with the testing data.

<table>
<thead>
<tr>
<th></th>
<th>RAWFULL</th>
<th>ABSTRACT</th>
<th>JOURNAL</th>
<th>MESH TERM</th>
<th>TITLE</th>
<th>ABS-JOUR-TITL</th>
<th>ABS-TITL</th>
<th>JOUR-TITLE</th>
<th>FULL FEAT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.865</td>
<td>0.869</td>
<td>0.625</td>
<td>0.785</td>
<td>0.836</td>
<td>0.859</td>
<td>0.859</td>
<td>0.824</td>
<td>0.904</td>
</tr>
<tr>
<td>Recall</td>
<td>0.851</td>
<td>0.832</td>
<td>0.671</td>
<td>0.785</td>
<td>0.804</td>
<td>0.859</td>
<td>0.861</td>
<td>0.813</td>
<td>0.891</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.858</td>
<td>0.85</td>
<td>0.647</td>
<td>0.785</td>
<td>0.81</td>
<td>0.859</td>
<td>0.86</td>
<td>0.818</td>
<td>0.897</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.936</td>
<td>0.928</td>
<td>0.853</td>
<td>0.907</td>
<td>0.917</td>
<td>0.939</td>
<td>0.939</td>
<td>0.925</td>
<td>0.959</td>
</tr>
<tr>
<td>Recall</td>
<td>0.942</td>
<td>0.946</td>
<td>0.826</td>
<td>0.907</td>
<td>0.932</td>
<td>0.939</td>
<td>0.939</td>
<td>0.922</td>
<td>0.956</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.939</td>
<td>0.937</td>
<td>0.839</td>
<td>0.907</td>
<td>0.924</td>
<td>0.939</td>
<td>0.939</td>
<td>0.922</td>
<td>0.956</td>
</tr>
<tr>
<td>Correct</td>
<td>91.45</td>
<td>91.12</td>
<td>77.89</td>
<td>87.01</td>
<td>89.31</td>
<td>91.45</td>
<td>91.53</td>
<td>89.07</td>
<td>93.83</td>
</tr>
<tr>
<td>Incorrect</td>
<td>8.54</td>
<td>8.87</td>
<td>22.1</td>
<td>12.98</td>
<td>10.68</td>
<td>8.54</td>
<td>8.46</td>
<td>10.92</td>
<td>6.16</td>
</tr>
</tbody>
</table>

Table 4.9: Best results for experimentation two with the second training data and testing with the final testing data.

From experimentation 2, it can be deducted that a combination of more features gives better results than when using a feature by itself. The best results for this experimentation is 93.83% of correctly classified, a precision of 0.904, a recall of 0.89 for positive labeled terms. And a precision of 0.91 and a recall of 0.92 for negative labeled terms. The worst performance was obtained when using only the journals. This is due the poor information that the terms in this feature can give, also it has just few words that doesn’t provide information gain to predict the class of the instance where it belongs.
### 4.1.3 Third Set of Experiments

The third set of experiments was also performed taking into account the combinations from experimentation one and two. For this experiments it was introduced the distinction to the terms in the different features. The combination of trigrams with stop words removal was substituted for the combination of unigram with stop words removal. This change was done due to the expectative of better results when using the newly added distinctive for the terms. It was also performed the holdout validation from experimentation two.

The first training data and testing on development data delivered the best results when using the combination of bigrams with stop words removed. It was determined that the distinction of terms in the features didn’t do a better performance than when using the terms without a distinctive. In this case, the combination of the raw abstracts with the other features, the combination of the Abstract-Journal-Title and the combination of the full features, gave similar results. Table 4.10 shows the best results obtained for this experimentation.

<table>
<thead>
<tr>
<th>Positive</th>
<th>RAWFULL</th>
<th>ABSTRACT</th>
<th>JOURNAL</th>
<th>KEYWORD</th>
<th>TITLE</th>
<th>ABS-JOUR-TITL</th>
<th>ABS-TITL</th>
<th>JOUR-TITLE</th>
<th>FULL FEAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0,92</td>
<td>0,873</td>
<td>0,653</td>
<td>0,83</td>
<td>0,793</td>
<td>0,913</td>
<td>0,918</td>
<td>0,842</td>
<td>0,918</td>
</tr>
<tr>
<td>Recall</td>
<td>0,86</td>
<td>0,842</td>
<td>0,556</td>
<td>0,759</td>
<td>0,732</td>
<td>0,84</td>
<td>0,865</td>
<td>0,709</td>
<td>0,865</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0,89</td>
<td>0,858</td>
<td>0,601</td>
<td>0,793</td>
<td>0,761</td>
<td>0,875</td>
<td>0,89</td>
<td>0,77</td>
<td>0,89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative</th>
<th>RAWFULL</th>
<th>ABSTRACT</th>
<th>JOURNAL</th>
<th>KEYWORD</th>
<th>TITLE</th>
<th>ABS-JOUR-TITL</th>
<th>ABS-TITL</th>
<th>JOUR-TITLE</th>
<th>FULL FEAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0,934</td>
<td>0,933</td>
<td>0,798</td>
<td>0,887</td>
<td>0,874</td>
<td>0,925</td>
<td>0,936</td>
<td>0,868</td>
<td>0,936</td>
</tr>
<tr>
<td>Recall</td>
<td>0,965</td>
<td>0,947</td>
<td>0,856</td>
<td>0,924</td>
<td>0,907</td>
<td>0,961</td>
<td>0,962</td>
<td>0,935</td>
<td>0,962</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0,949</td>
<td>0,94</td>
<td>0,826</td>
<td>0,905</td>
<td>0,89</td>
<td>0,942</td>
<td>0,949</td>
<td>0,901</td>
<td>0,949</td>
</tr>
<tr>
<td>Correct</td>
<td>93,01</td>
<td>91,53</td>
<td>75,76</td>
<td>87,01</td>
<td>84,96</td>
<td>92,11</td>
<td>93,01</td>
<td>86,11</td>
<td>93,01</td>
</tr>
<tr>
<td>Incorrect</td>
<td>6,98</td>
<td>8,46</td>
<td>24,23</td>
<td>12,98</td>
<td>15,03</td>
<td>7,88</td>
<td>6,98</td>
<td>13,88</td>
<td>6,98</td>
</tr>
</tbody>
</table>

Table 4.10: Best results for experimentation three with the first training data and testing on the development data

The combination of the first training data with the development data, and testing on the final testing data gave similar results to the experimentation with the training data and the development data. Also the combination of bigrams and stop words removal, gave the best results in this experimentations. Table 4.11 shows the best results for this experimentation.
In experimentation three, once again the use of full features is ratified as the best combination of features to obtain the classifying model. Also, the introduction of a distinction in the features didn’t do major effect when performing the classification. Furthermore, the bigrams and performing stop words removal, gave the best model for the classification task. This results were taken to determine the best classification model: “Bigram-Stopped-Fullfeatures”.

### 4.1.4 Best Features in the Best Classification Model

The best classification model is deducted from experimentation three. The most predictive features in the model were obtained when doing the feature selection in Weka. Figure 4.1 show the top sixteen terms that gave the major information gain for the best classification model.
As can be seen in Figure 4.1. The word “cost” has the major weight in the word vector. From a logical perspective, cost is an economical term that could determine if one article refers to a “health economic evaluation”. Other words like “attnumber” also will be weighting a lot due to the number presence in every economical article, however this word can be present in any record and is not a fact to determine if a record is relevant or not. Finally, the ranked attributes weight shows that most terms from a “health economic evaluation” will be pointing to an economical topic.

4.4 Discussion

This research brought an alternative to contrast the manual techniques for information search. Although HTA developed various experiments to find filters to reach the information search on “health economic evaluations”, they didn’t find an effective filter that has an important level of precision and sensitivity (recall). In this research their results suggest that “text words and terms in economic evaluation records do not sufficiently discriminate economic evaluation records from the other records” (Glanville, Fleetwood, Yellowlees, Kaunelis, Mensinkai, 2009). On the contrary, the results when using the automatic approach show that terms could be determinant to predict if a record belongs to a certain category.

The manual technique implemented by HTA researchers, used a different approach to classification where terms weight come into account. Although the selection of records, known as “gold standards” for positive records and the negative records known as “comparator records” give an idea of labeled documents, the manual method used in this approach does not involve a mathematical model to predict which term would be more suitable into a category. Also, the classification three approach that was used in the HTA implementation makes use of only one term (unigram), while in the automatic approach made in this research, it has been proven that a combination of terms (bigrams and trigrams) could give more meaning to the context where these terms belongs. This last difference between the two approaches could be critical in terms of effectiveness. However, one advantage over the automatic approach used in this research is that the HTA research delivered results for two record databases, the EMBASE database and the MEDLINE database. The automatic approach could not give good results when performing over the EMBASE database, due to the composition of the terms.

For the best results of this research, the classification model gives a precision of 0.922 and a recall of 0.86. The best result for the HTA filter approach was the filter MEDLINE G with a precision of 0.257 and a recall of 0.72. This is a clear advantage when performing with a computational system. However, this results could not be called as high performance results. A high performance result would be in a sensitivity range of more than 0.95. This could involve the use of a different algorithm, the use of more labeled records or the use of different combinations of terms.
Chapter 5
Conclusions and Future Work

4.1 Conclusions

This research shows that it can be other efficient alternatives for information search. A computational system combined with a classifying algorithm approach can give good results in terms of precision and recall. While using a manual approach could have a good sensitivity and low precision, an automatic approach could give good precision and good sensitivity. A precision of more than 0.9 could be obtained as well as a recall of more than 0.8. Moreover, this research has revealed that machine learning has a great advantage over a manual approach, this is due to the automatic work it can perform. Pointing to the results, a more precise prediction can be made when more terms are used, in this context, a hand-made analysis could take too much time to be done appropriately. In contrast, machine learning analyze the terms very fast and also learns from them.

It could be seen that different treatment of the different terms that belongs to the record could make a difference to deliver a good classification model. In this context the experimentation made with normal terms (words without any distinctive), gave results not different to the terms that had a distinction. In experimentation one and two, every term was used and merged with all the terms in other features to determine the best model. In experimentation three, it was used different terms referencing to the feature where they belong, but it didn’t make any contribution for the precision and recall scores. But, when removing stop words and stemming there was differences in the results. In this perspective, by removing stop words, it could be permitted that other “more relevant” words be weighting more so they can precisely give a category to a record.

The combinations of terms was another important point in this research. This treatment made difference when delivering scores for precision and recall. By making different tokens of terms (unigrams, bigrams, trigrams), many terms which by themselves doesn’t give enough information in an instance, could be more useful when being together with another term. For example, the best model involves bigrams. Also, starting from experimentation one it is observed that the use of more features can give more information of the category of a term in question. A combination of all features can give more terms to be analyzed, then it can predict better the category of a record.

Finally, Support Vector Machine was the ideal classification algorithm for this research. As SVM is a good binary classifier, it delivered the best results. The fact that data was already prepared with feature selection and filtering, made it easier for SVM to make use of the internal polynomial kernel to predict efficiently which terms belong to a category. However,
as the classification in this research is only binary, it could not have the same effect if more classes are added.

4.2 Future Work

The consecution of an automatic system for information search was achieved. However, there can be some modifications that can give a more efficient system, as well as delivering more functions. The following adjustment are considered for future work:

The Medline records were composed of different fields. For the extents of this research, only the title of the record, the title of the journal, the abstract and the MeSH terms were used. However, the extraction of other features from the other fields of the record could give different results in terms of precision and recall. As mentioned before, the combination of the four selected features, gave the best classification model. In this context, more features can give more information to determine the nature of a record. Also by finding patterns in the number can give meaning to the composition of the record.

This classifier was performed for only two classes (categories). It means if records are “health economics evaluations” or not. However, for future work it should be considered the prediction of more categories. Of course it should be included more labeled records with different categories and the use of different classifying algorithms to obtain a good performance and good efficiency.

Finally, the term occurrence analysis could determine which record belong to a certain class. However, by performing text processing techniques it could be determined in what degree a record is more relevant for the human perspective. Records will be ranked on a “human-conceived query” basis.
References


Genkin, Alexander, David D. Lewis, and David Madigan. "*Large-scale Bayesian logistic regression for text categorization*." Technometrics 49.3 (2007): 291-304.