Word-based evidence-level classification of medical publications

Thijs Albers
HAIT Master Thesis series nr. 09-007

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts in Communication and Information Sciences,
Master Track Human Aspects of Information Technology,
at the Faculty of Humanities
of Tilburg University

Thesis committee:

Dr. J.J. Paimans
Prof. Dr. E.O. Postma
Dr. M.M van Zaanen

Tilburg University
Faculty of Humanities
Department of Communication and Information Sciences
Tilburg, The Netherlands
December 2009
Abstract

This thesis presents and investigates a new method, called Classit, for classifying evidence in medical publications. Classit performs a text categorization task by classifying medical publications according to five different publication types which represent different levels of evidence. Classit’s feature selection is based on word frequencies and on the statistical weighting (tf*idf) of words taken from abstracts and full articles. These feature sets are used to classify abstracts as well as full articles.

Five experiments were performed on a data set from Pubmed/Medline consisting of 11,399 medical abstracts, extended with 75 full articles from Pubmed’s database. The experiments showed that abstracts are too sparse to obtain good classification, but abstracts do provide sufficient information for feature selection. The features selected from abstracts achieve good results in classifying full articles whereas the features selected from full articles do not because of the small sample of full articles that was used. This thesis shows that abstracts can be used for feature selection, but that they are too sparse for classification. We conclude that abstracts are most suitable for feature selection, whereas classification is best done on full articles.
Preface

This thesis is the final work of my studies in Human Aspects of Information Technology (HAIT) at the University of Tilburg. During one of our classes in Text Mining taught by Professor Eric Postma we received a guest lecture from Dr. Marius Nap introducing us to the field of medicine and the use of text - and data mining techniques in this field. In following meetings with Dr. Nap and Professor Postma we tried to find a suitable topic at the intersection of medicine and the subject of text classification that we were introduced to in Professor Postma’s class. Although this thesis perhaps did not end where we initially thought it would, I want to thank Dr. Marius Nap for the interesting talks, the feedback he provided me with and all time invested in this thesis.

In particular I would like to thank my supervisor Professor Eric Postma for his guidance and advice in times where his agenda seemed to be over-fitted and teleportation would have been a better subject for research. Furthermore I would like to thank Dr. Menno van Zaanen for his feedback and help with setting up and conducting the last-minute experiments that would not have made it into my thesis if it wasn’t for his help. Last but not least thank you Dr. J.J. Paijmans for participating in the exam committee.

Finally, family, friends; your feedback, help and patience are greatly appreciated.
Table of contents

1 Introduction ............................................................................................................................. 5
   1.1 Research question ............................................................................................................. 7
   1.2 Research set-up ................................................................................................................. 8
   1.3 Thesis outline ................................................................................................................... 8
2 Theoretical background ........................................................................................................... 9
   2.1 Evidence based medicine ................................................................................................. 9
   2.2 Combining relevance and level of evidence .................................................................. 10
   2.3 Levels of evidence ........................................................................................................... 11
   2.4 Recognizing levels of clinical evidence ......................................................................... 14
       2.4.1 Selecting features – the expert approach ................................................................. 15
       2.4.2 Selecting features – the (more) statistical approach ............................................... 17
       2.4.3 Supervised learning methods – classifiers ............................................................... 19
3 Classit .................................................................................................................................... 24
4 Methods .................................................................................................................................. 27
   4.1 Data ................................................................................................................................ 27
   4.2 Data preparation .............................................................................................................. 29
   4.3 Experimental set-up ........................................................................................................ 29
       4.3.1 Experiments ............................................................................................................... 31
5 Experiments and results ........................................................................................................ 33
   5.1 Exploratory experiments ................................................................................................. 33
   5.2 Classit AA ....................................................................................................................... 36
       5.2.1 Bigram features ......................................................................................................... 41
   5.3 Classit FA ....................................................................................................................... 43
       5.3.1 Introducing full articles ........................................................................................... 43
   5.4 Classit FF ....................................................................................................................... 44
   5.5 Classit AF ....................................................................................................................... 49
6 Discussion and conclusions .................................................................................................. 53
   6.1 Further research .............................................................................................................. 55
7 References ............................................................................................................................. 59
1 Introduction

Several years ago the idea existed that when clinicians faced a patient, they would determine the best course of action based on what was called the ‘art of medicine’ or ‘clinical judgement’ (Eddy, 2005). Even though extensive medical research has been carried out, medical decision making based on proven research barely received attention. Only until recently, the extensive research conducted in the medical field and the now widespread availability of research results has caused the growth of a new field of investigations: the field of medical (decision) support systems that make use of proven research made available electronically. The use of proven research for medical decision making has been coined with the more popular term evidence based medicine. A wide variety of technical systems exists to support medical staff in their everyday routines as well as in the processes of making life-changing and saving decisions. More and more technical applications support medical staff in their decision making, frequently based on medical research that is available electronically.

But the ‘technological revolution’ (within the medical field) has a side effect: everyone can now publish their own research and make it available worldwide. This causes the medical libraries and databases to grow rapidly and raises an important issue: how do we separate the wheat from the chaff? Medical staff is short of time (McColl, 1998) and lack the time to search all sources of medical research. Clinicians do not want to waste time deciding if a paper is of value (in terms of validity) to their specific question or information need with respect to a patient’s issue.

There are two important issues that have to be addressed in order to satisfy this information need: relevance of the results returned and the level of evidence in the results returned. Research on the level of evidence of publications returned in response to a search query is scarce. Improving the process of medical research information retrieval by only selecting research that contains a certain level of evidence will save valuable time of the medical expert. For instance, a single case report on an effective treatment may be relevant to a clinician’s information need, but an extensive cohort study is likely to be of greater relevance. A certain level of evidence could be required to ‘guarantee’ effectiveness of treatment. “Level of evidence” is defined as the quality of the study and the conclusions reported on and the
amount of evidence supporting the conclusions. Sackett (2000) and Hadorn (1996) have proposed hierarchies for the different levels of evidence. The different publication types in the medical field of research are divided into these levels of evidence based on the quality and amount of evidence these publications contain.

Working towards an automated method that is able to distinguish between different (semantic) levels of evidence will help users (clinicians) to faster find the research of interest and faster satisfy their information need. This thesis will focus on the task of automatically recognizing the level of evidence of a medical publication, i.e. a text-classification task. The classification task will be performed using machine-learning methods where a collection of pre-categorized documents is used to induce a statistical model (de Bruijn, 2002). The learned model will be used to classify unseen and unlabeled examples with their (most likely) publication type. We will use textual features in the form of whole words or stemmed words as representations of publication types and our method can therefore be considered a word-based method. We adopt a word-based method to preserve domain independence. Often, short text snippets in the form of abstracts are publicly available whereas access to author and citation information is restricted. By only selecting word features from abstracts we will preserve the domain independence of our approach.

Previous research on classifying publication types to find high-quality studies for clinical care only focused on identifying one or two different publication types (systematic reviews and randomized controlled trials) (White, 2000; Aphinyanaphongs, 2005). Swales (1999) and Hampton (1997) state that it is a misperception that only these two publication types are considered as useful evidence. This motivates our choice for five different publication types that are associated with the five levels of evidence.

Classit is a word-based text-classification method developed in this thesis that is able to distinguish between five publication types that differ for evidence-level identification. Classit has four variants that differ in the way that feature-selection and classification are defined. These variants are called Classit AA, Classit FA, Classit FF and Classit AF. The first letter refers to the source for feature-selection and the second letter refers to the target-source for classification (e.g. Classit FA represents the Classit variant using features selected from the full articles (F) to classify abstracts (A)). Classit’s feature selection is based on word frequencies and on the statistical weighting (tf*idf) of words taken from abstracts and from
full articles. Feature selection will be performed on abstracts as well as on full articles. These two feature sets are used for the classification of abstracts as well as for the classification of full articles. The combination of the source for feature selection and the target (abstracts or full articles) for classification results in the four Classit variants.

Table 1.1
Configuration of the four Classit variants

<table>
<thead>
<tr>
<th>Classification on</th>
<th>Features selected from</th>
<th>Classification on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstracts</td>
<td>Classit AA</td>
<td>Abstracts</td>
</tr>
<tr>
<td>Full articles</td>
<td>Classit AF</td>
<td>Full articles</td>
</tr>
</tbody>
</table>

These four Classit variants are evaluated in four experiments. The experiments involve the automatic classification of medical abstracts contained in Pubmed's dataset. Pubmed contains thousands of medical research abstracts. In addition, experiments are performed on a set of 75 full articles. By preserving domain independence and classifying multiple publication types, our work will result in a method that is applicable to different domains of medical research and easy to integrate because only short snippets of text are required as input. This work will therefore be a valuable improvement to current approaches and possibly provide a better alternative ranking of search results based on their levels of evidence.

1.1 Research question

This thesis focuses on a new method of evidence-level classification in medical publications. Classit will do so by selecting (whole) words as features for characterizing the different publication types, both from abstracts as well as from full articles. These two different types of feature sets are used for the classification of abstracts as well as for the classification of full articles.
The research question we address in this thesis by investigating the performance of the four Classit variants is the following:

*To what extent is it possible using a word-based approach to automatically classify medical publications in terms of level of evidence?*

### 1.2 Research set-up

We address the research question by evaluating the performance of the four Classit variants. The four Classit variants are evaluated in four experiments, preceded by a set of exploratory experiments. The exploratory experiments investigate the performance of features selected by other researchers combined with features selected from our own data set. The first experiment of the four Classit experiments investigates the performance of Classit AA which selects word-based features from abstracts and uses this feature set to classify abstracts (chapter 5.2). The second experiment investigates the performance of Classit FA, using word-based features selected from full articles for the classification of abstracts (5.3). The third experiment evaluates the performance of Classit FF, using word-based features selected from full articles for the classification of full articles (5.4). The fourth and final experiment investigates the performance of Classit AF using word-based features selected from abstracts to classify full-articles (5.5). An overview of the four Classit variants is provided in Table 1.1.

### 1.3 Thesis outline

Following the introduction, chapter two will provide the theoretical framework for this thesis that is needed to answer the research question. Chapter three will provide a detailed explanation of Classit. Chapter four describes the methodology and set-up of our experiments. Chapter five reports on the experiments carried out and the results that were achieved. Finally, chapter six will discuss the results from our experiments, followed by the conclusions that can be drawn from these results. We conclude with recommendations for further research.
2 Theoretical background

The aim of this thesis is to provide clinicians with an indication of each report’s level of evidence. It is therefore essential that we have a clear view of what constitutes high-quality clinical evidence and why the recognition of high-quality evidence is important and how this recognition can be performed automatically.

Section 2.1 explains what constitutes high-quality evidence and discusses the use of high-quality evidence in the field of medical research. Section 2.2 will demonstrate the need for high-quality evidence and how this evidence can be used to improve current search processes. In section 2.3 we discuss existing level of evidence hierarchies which are the starting points for the more technical theoretical background of this thesis. Starting from the level of evidence hierarchies, section 2.4 discusses current approaches for automatically recognizing high-quality evidence and leads to the technical procedures required for the task of automatically recognizing evidence in section 2.4.2 and 2.4.3.

2.1 Evidence based medicine

Every day clinicians make decisions on what course of action is best for a patient, based on the symptoms observed, on their experience, on the opinions of colleagues and on the available research combined with the patient’s wishes. These elements for decision making are illustrated in Figure 2.1 taken from Haynes (2002). The process of clinical decision making is an ongoing process of weighing internal evidence and external clinical evidence. Internal evidence is the result of education and the treatments of patients in the past or valuable information learned from colleagues. External evidence is derived from relevant clinical research and replaces existing experiences and treatments with better ones that are more safe, efficacious and more powerful. Good clinicians continuously combine these two sources of evidence.
(internal and external) and try to improve both. The combination of internal and external evidence is what constitutes clinical evidence and combined with the patient’s wishes is the basis for a clinical decision. Medical decision making based on clinical evidence has been coined with the popular term ‘evidence based medicine’.

Evidence based medicine (EBM) is defined as ‘the process of systematically finding, appraising, and using contemporaneous research findings as the basis for clinical decisions’ (Rosenberg et al., 1995). This definition perfectly illustrates the problem statement for this thesis that leads to our research question: To what extent is it possible using a word-based approach to automatically classify medical publications in terms of level of evidence?

Evidence based medicine traditionally consists of five steps (Straus et al., 2000).

1. Converting an information need (as a result from a clinical problem) into an answerable question.
2. The best evidence to answer this question has to be found.
3. The evidence found then must be appraised for its validity and relevance with respect to the clinical problem and information need.
4. This appraisal has to be combined with the clinician’s expertise and the patients’ values in order to determine the overall relevance and usability; can the result be applied in clinical practice?
5. Performance of the applied evidence must be evaluated.

This thesis will focus on the third step by developing a method that classifies medical texts (with their publication type) and indicates the level of evidence of a research report. This level of evidence score improves the quality of search results; relevant returned search results can now be ranked for level of evidence.

2.2 Combining relevance and level of evidence

Having results returned on the subject of interest is great; but the reliability of every single result – the level of evidence – of the study is at least as important. The level of evidence consists of the relevance with respect to the information need and clinical problem at hand, as well as the usefulness of the research results reported on. The publications retrieved should be
relevant, moreover clinicians should have confidence in the research results reported on. Therefore, adding a measure of level of evidence to existing search engines and algorithms would be very valuable. The growing demand and the increased pressure on clinicians impairs their ability to thoroughly analyse (internal or external) evidence (Straus et al., 2000). With the rise of new technologies and discoveries, treatments and medicine, the volume of available medical information and literature has grown exponentially. Therefore, a classified, targeted access to these information resources is necessary (Vanoirbeek et al., 2000). Instantly providing a score of a report’s level of evidence will contribute to the third step in the process of evidence-based medicine and save clinicians valuable time.

But what does that targeted access look like? This thesis focuses on the appraisal of research reports in order to facilitate the process of appraising research reports returned as results.

Appraising medical research means we need a clear understanding of what constitutes evidence and what does not. Several researchers have done extensive research into the different levels of evidence that exist in the medical field and have proposed evidence hierarchies accordingly.

### 2.3 Levels of evidence

Sackett et al. (2000) have developed a hierarchy of evidence that is widely adopted in the field of medical research and evidence-based medicine. This level of evidence hierarchy is used by clinicians for appraising and evaluating medical research. The different levels of the hierarchy can be interpreted as representing the publication types which are the most trustworthy and reliable or the least vulnerable to bias for the top levels, and the least trustworthy or reliable and the most vulnerable to bias for the lower levels (Rich, 2005).
The Expert Opinion rated as level 5 evidence illustrates the ranking; an expert opinion is very vulnerable to bias and less trustworthy than a systematic review of Randomized Controlled Trials. In Table 2.3.1 we can see that the Systematic Review of Randomized Controlled Trials (RCT) is ranked as the publication type containing the best evidence.

Hadorn et al. (1996) have developed a similar hierarchy although they label the different levels with different names. Hadorn et al. (1996) distinguish three main levels of evidence, Level A, Level B and Level C. Again the first level (A) can be interpreted as containing the best and most trustworthy evidence whereas the last level (C) can be interpreted as the level containing the least trustworthy evidence or the publication type most vulnerable to bias. Level A is research that is a well-conducted Randomized Controlled Trial (RCT) with 100 patients or more (including multi-center and meta-analyses) or a well-conducted RCT with fewer than 100 patients (Rich, 2005). Level B consists of case-control studies, observations studies with high potential for bias, case reports, etcetera. Level C consists of expert opinions.

<table>
<thead>
<tr>
<th>Level</th>
<th>Publication types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1A</td>
<td>Systematic Review of Randomized Controlled Trials (RCTs)</td>
</tr>
<tr>
<td>Level 1B</td>
<td>RCTs with Narrow Confidence Interval</td>
</tr>
<tr>
<td>Level 1C</td>
<td>All or None Case Series</td>
</tr>
<tr>
<td>Level 2A</td>
<td>Systematic Review Cohort Studies</td>
</tr>
<tr>
<td>Level 2B</td>
<td>Cohort Study/Low Quality RCT</td>
</tr>
<tr>
<td>Level 2C</td>
<td>Outcomes Research</td>
</tr>
<tr>
<td>Level 3A</td>
<td>Systematic Review of Case-Controlled Studies</td>
</tr>
<tr>
<td>Level 3B</td>
<td>Case-controlled Study</td>
</tr>
<tr>
<td>Level 4</td>
<td>Case Series, Poor Cohort Case Controlled</td>
</tr>
<tr>
<td>Level 5</td>
<td>Expert Opinion</td>
</tr>
</tbody>
</table>

Table 2.3.1 Levels of Evidence – Sackett et al. (2000)
just like Sackett et al.’s fifth level of evidence. Hadorn et al.’s hierarchy of levels of evidence is shown in Table 2.3.2.

Table 2.3.2 Levels of Evidence – Hadorn et al. (1996)

<table>
<thead>
<tr>
<th>Level</th>
<th>Publication types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level A</td>
<td>Well-conducted RCT with 100 patients or more (including multi-center and meta-analyses); well-conducted RCT with fewer than 100 patients (one or institutions and meta-analysis; well-conducted study).</td>
</tr>
<tr>
<td>Level B</td>
<td>Well-conducted case-control study, poorly controlled or uncontrolled (including RCT with one or more major or three or more minor methodological flaws), observations studies with high potential for bias (case series with comparison to historical controls), case series or case reports, conflicting evidence with more support.</td>
</tr>
<tr>
<td>Level C</td>
<td>Expert opinion</td>
</tr>
</tbody>
</table>

Sackett’s and Hadorn’s hierarchies are very similar. Although they named their categories differently they agree on the ranking of the different types of clinical research reports (publication types). Each category contains specific types of publications. Sackett (2000) ranks systematic reviews of Randomized Controlled Trials as the publication type containing the best possible evidence. Hadorn and colleagues (1996) also rank a ‘well-conducted randomized controlled trial with 100 patients or more (including multi-center and meta-analyses) or well-conducted randomized controlled trials with fewer than 100 patients’ as the publication types containing the best evidence namely ‘Level A’ (Hadorn et al., 1996).

The two hierarchies considered illustrate the fact that there is clear agreement among the different researchers concerning which publication types contain the most evidence. This is the link that this thesis builds on: if we are able to predict or reproduce a reports publication type we can score its position in the hierarchy of levels of evidence.


2.4 Recognizing levels of clinical evidence

Our goal is to classify each report with its publication type. If the publication type of a report is known, we have a clear indication of its level of evidence. Classifying each report with its publication type is a text-categorization task. In our experiments, we group our set of documents into different categories/classes (i.e., publication types).

The text-categorization task is essentially a task of finding a set of representational features that allow unknown documents to be labelled with the right label (category) by assigning them the known labels of highly similar documents. To arrive at such a set of representational features each category (i.e., publication type) needs to have a representation that is suitable for the learning algorithm (classifier) and the classification task (Joachims, 1997). The better the representation, the better the accuracy of the classifier will be.

The process of finding the best representations consists of two steps: 1) feature selection and 2) feature-set transformation (De Bruijn et al., 2002).

Feature selection is the selection of features that characterize documents. Research on information retrieval suggests that words are appropriate as representation units (features) and that the ordering of words in a document is of minor importance for many text-categorization tasks (Joachims, 1997). By disregarding the ordering of words in a document the document is represented as a bag of words.

Feature-set transformation is the modification and optimization of selected features and is necessary in order to improve document representations and improve classifier performance. This is done by reducing the feature set and by feature-transformation methods. Common feature-set reduction methods are stemming, stop word removal (meaningless words) and the removal of very rare words that may contain discriminative power but do not occur frequent enough and only burden the classifier (de Bruin, 2002). Stemming is the procedure of reducing words to their root or stem by removing inflections etcetera. For example, with stemming ‘medic’, ‘medical’ and ‘medics’ would all be stemmed to the same root ‘medic’. In this thesis stemming of abstracts and full articles was performed so that our feature selection and extraction procedures could tap from a richer source of information and synonym terms or terms spelled differently because of English/American (i.e. ‘behavioral’ vs. ‘behavioural’).
spelling were accounted for. Feature-set transformation is done by weighing each feature in the feature set to improve the (document) representations (De Bruin, 2002). The feature set can be weighted using all kinds of statistical functions to improve the document representation. A common weighting measure is the tf*idf weighting scheme.

Using words as features (representation units for each category) is the basis for a large part of research concerning text categorization. Past research performed on recognizing levels of evidence through text categorization in the field of medicine can be divided into two main streams:

1. Selecting features based on expert knowledge / input
2. Selecting features based on statistical analysis

We will elaborate on these two main streams in the following paragraphs.

### 2.4.1 Selecting features – the expert approach

Numerous studies have tried to find the best representation set of features by making use of expert knowledge (Haynes et al., 1994; Robinson et al., 2002). Representational terms that they felt would most likely return high-quality articles were selected based on interviews with expert clinicians and librarians (Aphinyanaphongs et al., 2005). These were words that would appear in article titles, abstracts or in an article’s meta-information like publication type or subject keywords.

These selections of features would serve as the basis for their classifier models and would iteratively be tweaked and improved after each test. However, the expert approach to finding a set of features has three disadvantages: manual labour, bias / domain restriction and the approaches are binary approaches. We now briefly elaborate on these three disadvantages:

1. Manual labour disadvantage

As Aphinyanaphongs et al. (2005) mention, the common methodological characteristics of research where features are selected using expert input is selection through interviews with expert librarians and clinicians or through extensive article inspection by experts. Haynes et al. (2004) also seek input from clinicians and librarians through interviews, reviews of search strategies, etcetera. These are approaches that require large amounts of manual labour and man time not to mention the time and resources needed to educate those experts. It is
therefore an approach that is hardly generalizable to other fields of interest because of the investments (in time) needed.

2. Bias / domain restriction disadvantage
A clear consequence of manual labour and expert knowledge is the risk of bias. The implicit selection bias (Aphinyanaphongs et al., 2005) could cause experts to select high-quality articles as a basis for feature selection that reflect their bias or preference (because of domain knowledge) toward a particular subject. Chosen representational units (features) that seem to do well in classification will therefore not classify well because of sound methodology but because of a favourable bias towards the subject. These features would therefore not represent high-quality evidence. Another consequence of (domain) expert knowledge is the applicability in other domains. For instance, Kilicoglu et al. (2009) note that training their model on a different collection than the one used for creating their clinical query filters (their representational units) is a limitation of their study. It seems impossible to avoid subject preference by asking experts for their domain knowledge to select features. Likewise, asking experts to determine a gold standard for experiments seems a self-fulfilling prophecy that leads to detrimental applicability of the approaches investigated in other domains.

3. Binary approach disadvantage
Most research that uses expert input takes a binary classification approach. The goal is either to find systematic reviews or randomized controlled trials since those are the publication types at the top of the hierarchy or to classify and find those articles that are either scientifically rigorous or not (after extensive manual annotation) (Robinson et al., 2002; Wong et al., 2004). However, the assumption that reports are either scientifically rigorous or not seems false since it completely disregards the different levels of evidence that exist. Hampton (1997) and Swales (1997) state that not just systematic reviews and randomized controlled trials can serve as evidence to an information need. Systematic reviews are often considered to be the gold standard, but they might not be the best types of evidence to answer questions about e.g. diagnosis (Straus et al., 2000). The task of recognizing evidence should therefore not be approached as a binary task but as an ‘incremental’ or multiclass task where different reports may contain different levels of evidence. As Hampton (1997) and Swales (1997) state; it is a misperception that only systematic reviews and randomized controlled trials are considered as evidence in the field of evidence-based medicine.
2.4.2 Selecting features – the (more) statistical approach

The other group of researchers, who have not chosen to make use of experts, adopted a (more) statistical approach. ‘More’ statistical because often these researches still involved subjective choices made by the authors/researchers.

White et al. (2001) aimed to improve previously developed methods for identifying systematic reviews (in Medline). They selected features based on frequency analyses. All article titles were analysed and a table ranking all the words from these titles, based on the number of records containing this word, was produced. The latter number is known as the document frequency (df). They excluded stop words (words without any semantical meaning like ‘for’, ‘the’, ‘other’, etc.) and used a cut-off threshold of 25%. This means that any word that would not appear in at least 75% of all articles was removed and thus excluded from further analysis. This approach selecting candidate features from the article titles yielded only one word as a candidate feature namely ‘meta-analysis’. They repeated this process but now for all the abstracts which yielded a list of fifty-five words. This list was combined with the one candidate feature selected from the article titles and resulted in a list of fifty-six candidate features. From this list of fifty-six candidates several were eliminated by the researchers because they felt that these words did not contain sufficient distinctive value. This step deteriorates the generalizability of their research and can be considered a subjective choice within their statistical approach. Next to that, their data set consisted of articles matched on subject so that any specific terminology used would be found in all types of records (systematic and non-systematic reviews) and not emerge as terms identifying a specific publication type (systematic review). The problem is however that through matching on subject they might actually enlarge the effect of potential subject related terms emerging as identifying terms. These words might not necessarily be methodological terms, but synonyms of terms also appearing in other domains/subjects albeit in other forms. For example; American versus British form, different spelling, different abbreviations, etc.

Each record was then scored for this list of fifty-six words measuring each word’s frequency in that record (article) in both title and abstract. These frequency vectors (the collection of word frequencies) served as input for their statistical analysis in order to determine which of the terms would best distinguish between the types of records (non-systematic review, non-review, systematic review).
White et al.’s statistical analysis consisted of a discriminant function (D):

\[ D = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_{55} x_{55} + b_{56} x_{56} \]

**Equation 1**

In this equation the \( b \)'s are parameters or weights associated with the search terms (x). These weights can be seen as the predictive value for the words being a systematic review (or not). These predictive values (\( b \)) are calculated based on their frequencies in the data (the records).

The probability of a record being a systematic review is then calculated with the following logistic function:

\[ P = \frac{1}{1 + e^{-D}} \]

**Equation 2**

Glanville and colleagues (Glanville et al., 2006) adopt a similar approach to find distinctive terms/words for identifying controlled trials in Medline. They also searched for frequently appearing words in titles and abstracts and statistically tested these to find the best distinctive features.

In our research we will use two methods of feature selection.

*Word frequency.* The frequency of a word within the whole set of documents that belong to the same publication type. Words that are not in the top 100 based of most frequent words are not included in feature selection. Essentially the use of word frequencies is also a weighting measure. By weighing each term’s importance - based on its frequency - terms are either selected or left out for feature selection.

*tf*\(^*\)idf *term weighting.* \( Tf \) stands for term frequency and \( idf \) stands for inverse document frequency. Term frequencies are often highly correlated between different documents in the same collection. Therefore, all documents containing that term tend to be retrieved (Salton et al., 1988). By using a document dependent factor such as inverse document frequency, terms that are concentrated in a limited number of documents are favoured. Consequently, making
use of tf*idf for term weighting, terms that appear in virtually all documents will have a lower score than terms that appear in fewer documents. However, to ensure that words that are typical for a collection of documents will not be ranked too low, we will use an intermediate solution by using the sum of all individual tf*idf values. In this way, less distinctive terms with a high prevalence in the entire collection of documents within a publication type will rank higher.

White and colleagues (White et al., 2001) state that by staying within the field of one subject, special terminology found in a specific publication type should be found in all record types within that field and therefore not emerge as discriminate terms for a certain publication type. However, they do expect methodological terms to differ between different publication (report) types, and expect these terms to come up after statistical analysis. This is an important assumption that is adopted in this thesis as well. Because all records in our data set come from the medical domain (and from different subject fields within the medical domain) the expectation is that methodological terms differ between the five chosen publication types and that these terms will become clear after (statistical) analysis and feature set transformation methods. Because of the number of publications (abstracts) in our data set taken from different subject fields we expect the methodological terms to prevail whereas we expect any special terminology to be overruled because of the sample size.

2.4.3 Supervised learning methods – classifiers

The document representations that consist of features – each feature representing for example a frequency of a word – form a vector. A document vector consists of a set of features and a category at the end. Using vectors with categories as input to the classifier is called supervised learning because the category that the document belongs to is already known (because of manual labelling for example). All document vectors including the categories are passed on to the classifier which will then train a statistical model based on all these examples. The resulting statistical model can then be applied to unknown and uncategorized documents to determine the category. The better the feature selection, the better the document vector and the better the statistical model will be in classifying new documents.

Popular (supervised) learning methods (classifiers) are Naïve Bayes, Decision Trees, Nearest Neighbour and Support Vector Machine. All these methods make use of pre-categorized
documents to train a statistical model that can then be applied to a collection of uncategorized documents. The resulting statistical model is the classifier. In this thesis we make use of two supervised learning methods for the classification task: Naïve Bayes and Sequential Minimal Optimization (SMO) which is a Support Vector Machine learning algorithm.

A classifier is a (statistical) function that maps an attribute vector to a confidence that the attribute vector

\[ \hat{x} = (x_1, x_2, x_3, \ldots x_i) \]

Equation 3

belongs to a class (Dumais et al., 1998). This means that given the attributes (which are words or words’ frequencies in the document) a confidence

\[ f(\hat{x}) = \text{confidence}(\text{class}), \]

Equation 4

is calculated that this document or document vector containing these words or word frequencies belongs to a certain class (category). In text categorization, the attributes of the attribute vector often represent words (frequencies) and the class represents the category which in our case is the publication type the document belongs to.

An example of a learned function within a classifier could be the following:

\[ \text{confidence(randomized controlled trial)} = 0.3 \cdot \text{randomized} + 0.4 \cdot \text{placebo} + 0.7 \cdot \text{clinical trial} \]

Equation 5

The confidence of the attribute vector with the attributes ‘randomized’, ‘placebo’ and ‘clinical trial’ as word-features belonging to the class ‘randomized controlled trial’ is the summation of the features’ frequencies multiplied by the parameters. The parameters 0.3, 0.4, 0.7 in the function above (equation 5) are the learned parameters for each feature based on the examples of document vectors supplied to the classifier.
Evaluation metrics

Precision and recall

Precision and recall are evaluation metrics standard to the field of information retrieval. Precision is the proportion of retrieved documents that are labelled with a considered class and actually belong to that class. Recall refers to the set of (relevant) documents retrieved from a considered class from all documents in the data set belonging to the considered class (all relevant documents).

\[
\text{precision} = \frac{\text{number of documents retrieved and relevant}}{\text{all retrieved documents}}
\]

Equation 6

\[
\text{recall} = \frac{\text{number of documents retrieved and relevant}}{\text{number of relevant documents in database}}
\]

Equation 7

In order to be able to calculate these metrics it is essential that we know which documents in our data set are relevant and which are not. In other words; the documents must be labelled/categorized in order to judge the classifiers performance. There is an important trade off between recall and precision: do you either want to retrieve all documents that you were looking for at the risk of having more results and possibly some incorrect ones or do you strive for a smaller but more accurate set of results at the risk of not including some relevant documents? This is a choice with important consequences since we want to save the clinicians as much time as possible in deciding on results’ evidence levels.

An important third evaluation metric that deals with this trade off issue is called the F-Measure. The F-Measure is a weighted average of recall and precision. The F-measure is defined as:

\[
F - \text{measure} = \frac{(1 + \beta^2) pq}{\beta^2 p + q}
\]

Equation 8
\( p \) represents precision and recall is denoted by \( q \). \( \beta \) is a parameter which represents the relative weight assigned to precision and recall. The default F-measure balances the two and assigns equal weights to both precision and recall. However, it can be important to assign more weight to for example recall. I.e. when it is crucial to retrieve all documents that are relevant to the search query – so that all important information will be retrieved - more weight should be assigned to recall. A side effect is that possible irrelevant documents are more likely to be introduced into the set of retrieved documents. When it is important to have as little noise as possible within the set of retrieved documents more emphasis can be placed on precision. The risk here is that some documents that actually are relevant are left out in the set of retrieved documents. Putting more emphasis on either precision or recall can be done by increasing the value of \( \beta \). Values of \( \beta \) higher than 1 assign more weight to recall whereas values of \( \beta \) lower than 1 emphasize precision.

Sensitivity and specificity: The Area Under the Curve

The Area Under the Curve (AUC) Receiver Operating Characteristic (ROC) curves allow visual analysis of the scores of sensitivity (y-axis) and the ‘1 – specificity’ (x-axis) of a classifier test. Sensitivity refers to the true positives; e.g. the documents of class ‘systematic review’ that are correctly classified as ‘systematic review’. Specificity refers to the true negatives; e.g. the documents that are correctly not labelled as ‘systematic reviews’ and that also aren’t ‘systematic reviews’.

Ideally one would want to maximize both sensitivity and specificity by labelling as much documents with the right label as well as leaving as many documents out that are not relevant to the search query. However, it is rare that this 100% score on both measures can be achieved. In some cases one may opt for a higher sensitivity at the cost of lower specificity (Fan et al., 2006). By ensuring that a larger proportion of the relevant documents (true
positives) are retrieved, one risks to also include more true negatives i.e. documents that are not relevant. By visualising the results of a classifier through ROC curves one can easily read and decide on cut-off threshold levels and the assignment of weights but also instantly spot performance of the classifier.

The area under that ROC curve is a measure of a classifier’s discriminative power (Faraggi et al., 2002). An area under the curve of 1.0 would mean a 100% perfect test, whereas an area under the curve of 0.5 would mean that the test has no discriminative power.
3 Classit

This chapter explains the specific choices made for the set-up of Classit, building on the theoretical backgrounds discussed in chapter two. Classit’s goal is to score medical publications for level of evidence by classifying medical publications for their publication type. By classifying for publication type we will have a clear indication of a report’s measure of evidence.

As we argued in chapter one, the four different Classit variants differ in the way that feature selection is defined. Feature selection is either performed on abstracts or on full articles. These two feature sets are then used to classify either abstracts or full articles. Classit’s feature selection is based on word frequencies and on the statistical weighting (tf*idf) of words taken from abstracts and from full articles.

Classit variants

The four Classit variants are explained below. The first letter refers to the source for feature selection (which can either be abstracts or full articles) while the second letter refers to the source that is used for the classification procedures.

Table 3.1

The four Classit variants

<table>
<thead>
<tr>
<th>Variant</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classit AA</td>
<td>Investigates the performance of word-based features selected from abstracts and used to classify abstracts.</td>
</tr>
<tr>
<td>Classit FA</td>
<td>Investigates the performance of word-based features selected from full articles for the classification of abstracts.</td>
</tr>
<tr>
<td>Classit FF</td>
<td>Investigates the performance of word-based features selected from full articles for the classification of full articles.</td>
</tr>
<tr>
<td>Classit AF</td>
<td>Investigates the performance of word-based features selected from abstracts to classify full-articles.</td>
</tr>
</tbody>
</table>
**Word-based**

By adopting a word-based approach we hope to preserve the applicability of the approach in all domains. We do so by performing our experiments using a data set that is not restricted to any medical domain. The data set that is used covers a wide variety of subjects within the field of medicine as described in the medical publications listed in Pubmed. Succeeding in finding such an approach that is able to find and use words as representational features for classifying the level of evidence within a text would yield an approach that is widely usable. Classit makes use of whole-word features only because words are always available, in any domain. Features representing number of authors, citations or journal origin are examples of features that are not present in all domains. By adopting a word-based approach our method is not restricted to a single domain.

The selection of features will be performed using two methods: 1. word frequency and 2. tf*idf score. For each publication type we will generate frequency lists of the words appearing in those publication types, and rank those words on their frequency in descending order. Secondly we will generate lists of the words appearing in each publication type, ranked by their tf*idf score. These feature-selection methods are explained in chapter two.

**Abstracts and full articles**

Initially the choice was made to only analyse abstracts because abstracts are publicly available. This goes for Pubmed / Medline but for various other repositories containing medical research as well. Often abstracts or text snippets are publicly accessible while full-article access is restricted. Analysing abstracts is therefore done because of two reasons: 1) abstracts are publicly available and offer the best and richest form of semantic word content for analysis and 2) this is an approach that can be used within other repositories as well. If this approach proves useful it can be used in all domains (within the medical field) where smaller bits of text are (often) provided. The first experiments however indicated that the abstracts in our dataset were too sparse for classification procedures. For means of comparison as well as the exploration of possibilities using the proposed method with full articles we introduced full articles into the data set.
Five publication types

Five publication types were chosen as target publication types for the classification task: Review, Randomized Controlled Trial, Multicenter Study, Comparative Study and Case Reports. By choosing five different publication types we aim to develop a method that is able to classify more than one level of evidence. The five different publication types are each represented in the hierarchies of level of evidence. If we are able to identify five publication types, we are able to identify five (different) levels of evidence. By classifying five different publication types Classit becomes a method that aims to solve the task of text classification for multiple categories and goes beyond the binary approaches investigated by others who aimed at identifying only one publication type and therefore only one level of evidence (White et al., 2001; Robinson et al., 2002; Glanville et al., 2006; Kilicoglu et al., 2009).

Three publication types (Multicenter Study, Comparative Study and Case Reports) were added to investigate if publication types ranked lower in the hierarchy of levels of evidence can also be classified using a single approach. The three additional publication types were chosen randomly; the only requirement was that they would appear at least 100 times in the data sample. The publication types Review and Randomized Controlled Trial were investigated for reasons of comparison with research done by others (White et al., 2001; Robinson et al., 2002) and because of these publication types’ high positions in the level of evidence hierarchies.

Adopting this method will lead to a set of keywords and features that can be placed on the levels of evidence hierarchy and that is indicative for the different levels of evidence. This would provide a semantic keyword alternative to Sackett’s levels of evidence (hierarchy). This set of keywords can then be used as a feature set to rank relevant returned results (alternatively) on their level of evidence by scoring each abstract or short piece of text for these features.
4 Methods

This chapter will present the way the experiments were structured, the way they were set up and conducted. The goal is to investigate to what extent our word-based approach is able to distinguish between the five publication types in order to get an indication of an article’s level of evidence.

4.1 Data

Pubmed / Medlines sample data\(^1\) have been chosen for all analyses. These data samples were used as the basis for all experiments because they contain an extensive set of research reports which are all made available in xml-format. This format enables easy import for later use and analysis. Another important consideration for using this dataset was the fact that each record was already labelled with a publication type. As argued in chapter two this is a very important requirement for supervised machine learning; the availability of categories/labels in order to train the classifier so that it can use this model to label new / unseen documents. Next to that, similar samples from this Medline data set that have been used by other researchers as well (White et al., 2001; Glanville et al., 2006).

An important limitation is the fact that this gold standard (the availability of labels) might not be the best gold standard one would want to use as the basis for machine learning experiments. Even though all labels (of publication types) might be correct, the records’ contents might be of lower quality and might contain research articles that are published in medical journals not highly ranked by others within the medical research community. These records could therefore appear in our dataset without being representative for true quality or evidence.

Since the goal is to find an approach that is as generic and as widely applicable as possible the variable quality of the records may at the same time have an advantage: by not only selecting ‘true gold standard articles’ a data set is used that represents the composition of medical research everywhere and a dataset is used that helps finding those features that represent all

\(^1\) These data samples are available at http://www.nlm.nih.gov/bsd/sample_records_avail.html
levels of evidence from all fields and areas of medicine. Evidence that can be found in the top journals but also evidence that can be found in less well read journals or in publication types not at the top of the level of evidence hierarchy.

The initial set of abstracts consisted of 7,899 abstracts, distributed across five publication types. After an exploratory test this set of abstracts was extended with another 3,500 abstracts to 11,399 abstracts. The composition of the data set is given in Table 4.1.

Table 4.1
Number of abstracts per publication type.

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Number of abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review</td>
<td>4,208</td>
</tr>
<tr>
<td>Randomized Controlled Trial</td>
<td>919</td>
</tr>
<tr>
<td>Comparative Study</td>
<td>3,952</td>
</tr>
<tr>
<td>Multicenter Study</td>
<td>283</td>
</tr>
<tr>
<td>Case reports</td>
<td>2,037</td>
</tr>
</tbody>
</table>

Abstracts with double publication types

During the experiment of Classit FA we noticed that some publications in our data set of 11,399 abstracts were labelled (by Pubmed) with more than one publication type from our five selected publication types. Further analysis showed that in total 1757 abstracts equally distributed over the five publication types were double labeled. All earlier experiments were repeated using the set of 9642 abstracts (after removal of the double labeled abstracts) but performance scores did not change significantly. It is therefore that the remaining experiments Classit FA and Classit AA using all possible features were performed using this smaller set of 9642 abstracts, while the results of the experiments already performed and reported were left unchanged. For the same reason of publication types double labelled the baseline in the Classit FA experiment and the Classit AA using all possible features experiment is slightly higher.
Full articles

Because of limited resources and time for this thesis and the time needed for manual import of full articles, only fifteen full articles per publication type were imported. Unfortunately Pubmed does not support automatic export/import of full articles as it does support automatic export of abstracts and citation information.

The procedure of importing full articles was as follows: for each publication type all Pubmed id’s were first exported from a MySQL database set-up for the experiments and converted into a query suited for retrieving all the articles from Pubmed’s online search engine\(^2\). This query was expanded by asking Pubmed only to return results with links to the free full articles. The building of this query decreased the time needed to find full articles of which the abstracts already resided in our data set. With a custom PHP script each full article was then inserted into the MySQL database after removal of stop words and stemming all remaining words using Porter’s Stemmer.

4.2 Data preparation

All data was imported into a MySQL database using a custom PHP script. For each record imported from the Pubmed / Medline data sample the title, abstract, publication type and Pubmedid (unique article identifier) were saved in the database.

A non-domain specific list of English stop words\(^3\) was used to eliminate stop words from the dataset, both in the titles and the abstracts. Later on, stop words were also removed from the full articles. The document vectors’ word frequencies were not normalized for document length since document lengths were similar for the whole dataset; all abstracts generally were of the same size.

4.3 Experimental set-up

The availability of the data samples in xml format and consequently the import of those records into a MySQL database provided the advantage and opportunity of running all experiments on a local computer without having to make use of an internet connection. Even


\(^3\) [http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words](http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words)
though Pubmed offers support for sending search queries to Pubmed and retrieving article titles, abstracts and citations online (see Entrez Utilities\textsuperscript{4}) the choice was made to run every experiment local using a XAMPP (Apache, MySQL, PHP) set up because of computational efficiency. After a few initial tests using Pubmeds Entrez utilities it was clear that running experiments offline would save time in the whole process of developing the experiments and running these experiments.

\textit{Weka Data Mining Software}

Weka is an open source software package (Hall et al., 2009) written in Java and is a collection of machine learning algorithms designed for data mining tasks. In this thesis Weka is used for all task classification tasks as well as some data pre-processing tasks. Weka is available for download at http://www.cs.waikato.ac.nz/~ml/weka/ . The version of Weka that was used in this thesis is Weka 3.6.1.

Weka offers easy import of vector data in different forms and allows to run different classifiers on this data within a few clicks. Performance and evaluation metrics are given automatically and different visualization methods (to identify outliers for example) are available. Before importing all vector data into Weka the sets of vectors were first converted to .csv files. In this thesis we use the Naïve Bayes and SMO (Sequential Minimal Optimization) methods.

\textit{Significance testing of recall scores}

One may notice that the recall scores reported in our tables and the recall scores (means) reported in the results of our statistical significance tests differ a little. In particular this is the case when we report on the results of the approaches classifying full articles. These recall scores may differ from the results reported in the tables because all experiments were repeated ten-fold in a different run of Weka so that each fold’s recall score could be used as input to our statistical significance test.

\textit{Feature selection, vector creation}

All procedures of selecting features, creating vectors with feature frequencies were done using custom written PHP scripts. Where others have used software packages like Wordstat, Simstat

and advanced Machine Learning Software there were no resources available to purchase and use any of these packages except for the free and open source Weka Data Mining Software. Moreover, working with custom PHP scripts has the advantage and possibility to easily make small changes to experiments and perform new experiments.

The vectors created for every publication consisted of each feature’s count (frequency) within that publication, followed by the class of the publication at the end of the line. An example of a vector consisting of thirteen features is given below:

0, 3, 1, 5, 0, 2, 1, 1, 1, 0, 0, 1, 3, Randomized Controlled Trial
Example of a feature vector

4.3.1 Experiments

Exploratory experiments

A set of exploratory experiments preceded our main set of experiments evaluating the four Classit variants. These exploratory experiments tested the performance of a combination of features selected by other researchers to classify Reviews and (Randomized) Controlled Trials (White et al., 2001; Robinson et al., 2002) together with features selected by us from our own data set. White et al. (2001) and Robinson et al. (2002) investigated the best search strategies and term combinations for the retrieval of systematic reviews and controlled trials. The approach they adopted was a binary approach, only focused at finding and identifying one specific publication type. From their research the best performing keywords were taken and selected as representational features for the publication types Review and (Randomized) Controlled Trial. That is, the keywords with the most discriminating value with regard to the target publication type. The features selected from our own data set were selected from frequency lists for each publication type that were created without the use of any feature transformation or reduction method.

Classit experiments

The four Classit variants differ in the way that feature selection is defined. Feature selection will be performed on abstracts as well as on full articles, resulting in two feature sets. These two feature sets are used for the classification of abstracts as well as for the classification of
full articles. The combination of the source for feature selection and the target (abstracts or full articles) for classification results in the four Classit variants.

The four variants of Classit were evaluated in four main experiments.

1. Classit AA investigates the performance of word-based features selected from abstracts and used to classify abstracts.
2. Classit FA investigates the performance of word-based features selected from full articles for the classification of abstracts.
3. Classit FF uses word-based features selected from full articles for the classification of full articles.
4. Classit AF uses word-based features selected from abstracts to classify full-articles.

The order of the Classit experiments is motivated in chapter five. To test if abstracts are too sparse, the experiments evaluating Classit FA and Classit FF (using full article features to classify abstracts as well as to classify full articles) were performed after the first experiment evaluating Classit AA.

The different variants of Classit that were experimentally evaluated are depicted by the overview of Classit shown in figure 4.1. Note that the order depicted below is not the actual order in which the experiments were performed.

**Figure 4.1 Overview of Classit variants**
5 Experiments and results

This chapter describes the experiments and the results. We start by presenting the results of the exploratory experiments (5.1), followed by the Classit AA experiments using features selected from the abstracts and tested on abstract vectors (5.2). The next section reports on Classit FA, the experiments performed with features selected from the full articles and used for the classification of abstracts in section 5.3. Section 5.4 describes the experiments performed for the evaluation of Classit FF using features selected from the full articles for the classification of full articles. Section 5.5 reports on the results of Classit AF, the experiments with abstract features used to classify full articles.

5.1 Exploratory experiments

The first exploratory experiment investigated the performance of the features selected from White et al.’s (2001) research and Robinson et al.’s (2002) research combined with features selected from our own data set. The two publication types Randomized Controlled Trial and Review were each represented by six features selected from White et al. (2001) and Robinson et al. (2002).

The keywords for the other publication types Comparative Study, Case Report and Multicenter Study were selected by creating frequency lists of all the words from all abstracts within each publication type after the removal of stop words. These frequency lists were generated using custom written PHP scripts. The six most frequent keywords were selected as representational features for each publication type and these keywords were not stemmed. Because our data set was not restricted to one subject (as White et al.’s data set was), there is a risk of lower performance due to subject related terminology or researcher related abbreviations and writing. We prevented this by manual checks for abbreviations or subject related terminology on the six features selected per publication type.

Keywords that would also appear within the other publication types were discarded as candidates for representational features. Due to this process the final vector for the five publication types consisted of only 24 features not equally divided over all publication types,
followed by the publication type. 7,899 vectors were created, each containing every feature’s frequency in the article scored for, followed by the class (publication type) at the end.

A first test showed that this approach resulted in a poor performance. Outlier analysis showed that especially the first seven keywords had a very low frequency in the data set. Five of the features chosen based on White et al.’s research (2001) and Robinson et al.’s research (2002) occurred very rarely within the data set. This could mean that our data set is too small and different or that our strategy is too different from the strategies used by White et al. (2001) and Robinson et al. (2002). To test if the size of the data set was the cause of this we extended the data set with 3,500 abstracts. Another frequency analysis showed that this set of features still achieved poor results and that the outliers scoring low on frequency had not increased in frequency. Figure 5.1 shows that in particular feature 2, 3, 4 and 7 occur rarely, their frequency in the dataset is too low to be of any distinctive and informative value. An attribute evaluator test with an information gain ranking filter proves this assumption. Table 5.1 show the 10 features with the lowest information gain/value. Feature 2, 3, 4 and 7 are at the absolute bottom.
Table 5.1

Information gain of 10 least distinctive individual features

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Information gain</th>
</tr>
</thead>
<tbody>
<tr>
<td># 21 treatment (pt: case reports)</td>
<td>0.0333</td>
</tr>
<tr>
<td># 22 cases (pt: case reports)</td>
<td>0.0193</td>
</tr>
<tr>
<td># 5 studies (pt: review)</td>
<td>0.0187</td>
</tr>
<tr>
<td># 24 clinical (pt: multicenter study)</td>
<td>0.0177</td>
</tr>
<tr>
<td># 6 design (pt: review)</td>
<td>0.0164</td>
</tr>
<tr>
<td># 1 controlled (pt: review)</td>
<td>0.0079</td>
</tr>
<tr>
<td># 2 extraction (pt: review)</td>
<td>0.0</td>
</tr>
<tr>
<td># 7 randomized controlled trial (pt: randomized controlled trial)</td>
<td>0.0</td>
</tr>
<tr>
<td># 3 randomized controlled trials (pt: review)</td>
<td>0.0</td>
</tr>
<tr>
<td># 4 meta-analysis (pt: review)</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Note*

*pt = publication type*

For reasons of comparison two classifiers were run on this extended dataset of 11,399 instances. The results are shown in Table 5.2

Table 5.2 Performance of exploratory strategy

*11,399 instances, ten-fold cross validation*

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.57</td>
<td>0.59</td>
<td>0.55</td>
<td>0.78</td>
</tr>
<tr>
<td>SMO</td>
<td>0.61</td>
<td>0.62</td>
<td>0.60</td>
<td>0.76</td>
</tr>
</tbody>
</table>

The most frequent class baseline equals .369 (36.9%)
With an average precision of 59% with the Naïve Bayes classifier there is a lot of room for improvement. Further analysis shows that the Multicenter Study class in particular achieves poor results. Due to this class only being represented by one feature and that feature having no distinctive value (Keyword 24, see Table 5.1) this result seems unavoidable.

The features that were selected based on the research done by White et al. (2001) and Robinson et al. (2002) scored poor results. An explanation for this result is that their search strategies also involved searching publication types, titles and subject terms whereas our research only searches abstracts. Next to that, their data set consisted of articles matched on subject whereas we aimed to keep our data set as broad as possible. This strategy can be considered both an advantage as well as a disadvantage. Our strategy might be more generalizable and (eventually) applicable in different domains but at the same time introduce more noise and errors because of terminology and subject related problems when terms do not match between different subjects.

Since the exploratory experiments with the features taken from White et al.’s research (2001) scored very poor we conclude that the difference in data set and strategy applied is too big. As a next step for our own experiments we will therefore use a better and balanced set of features selected from our own data only.

5.2 Classit AA

The procedure of creating frequency lists for every publication type was repeated. For each publication type we generated frequency lists of the top 100 most frequent words in all abstracts after the removal of stop words in descending order, showing the words with the highest frequency at the top of the list. Instead of working with a percentage as cut-off threshold level the cut-off threshold level was set at the top 100 most frequent keywords. Words ranking lower than the 100th position based on their frequency only occurred twice or once within the whole dataset which is too low to be of any informative and distinctive value with regard to publication type.
Again these frequency lists for each publication type were compared and only keywords that seemed to have distinctive value were selected. Note that this could cause the results to be biased because of the researcher’s preferences. Although words like ‘patients’ hardly contain any distinctive value since it appears in nearly every abstract in every publication type, this must still be considered a subjective choice. ‘Randomized’, ‘double-blind’ and ‘placebo’ however seem to be of more distinctive value for the publication type ‘Randomized Controlled Trial’ (RCT). This time each publication type was represented by six keywords as representational features in the frequency vector, making each vector balanced and the features equally distributed across all publication types. The final vector now consisted of 30 features representing 30 keyword frequencies followed by the publication type.

In comparison with the first exploratory strategy the results have improved only marginally (Table 5.3). Precision and recall have only improved by 2 percent at most. A paired-samples t-test was conducted to evaluate the improvement in the recall score of the Naïve Bayes classifier between the exploratory strategy and the Classit AA strategy. There was a statistically significant increase in recall for the Naïve Bayes classifier from the exploratory strategy ($M=57.01, SD=1.57$) to Classit AA [$M=58.58, SD=1.63$, $t(9), p<.05$]. We also conducted a paired-samples t-test to evaluate the improvement in recall of the SMO classifier. There was a statistically significant increase in recall for the SMO classifier from the exploratory strategy ($M=61.01, SD=1.19$) to Classit AA [$M=62.39, SD=1.11$, $t(9), p<.05$]. However, the better balanced feature set selected from this data set did not result in the improvement that was expected.
Table 5.3 Performance of Classit AA compared with exploratory strategy

Six abstract features per publication types. 11,399 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall Expl. strategy</th>
<th>Precision Expl. strategy</th>
<th>F-Measure Classit AA</th>
<th>AUC Classit AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.57</td>
<td>0.59</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>SMO</td>
<td>0.61</td>
<td>0.62</td>
<td>0.64</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Note: superscript indicates significance of difference with column to the left

a: $p < .01$; b: $p < .05$; c: $p < .0001$;

The most frequent class baseline equals .369 (36.9%)

From the above we conclude that the improvements are only marginal and that this strategy needs better feature transformation methods in order to improve feature representations. White and colleagues note that highly relevant terms chosen subjectively do not perform as well as those derived by a statistical approach (White et al., 2001). Therefore we will continue along the road of their findings and follow a more statistical approach of keyword selection.

Feature set transformation

By calculating the tf*idf score for each keyword in each publication type we will have a measure of each keyword’s importance and relevance with respect to that document in the corpus. After calculating the tf*idf score for each individual keyword the individual tf*idf scores of keywords also appearing in other documents were summed. A top 100 of keywords that had the highest tf*idf scores was generated for each publication type. Again the top 100 – this time with the highest tf*idf scores - is the cut-off threshold level for all feature candidates.

Keywords that would also occur in other publication types (doubles) were deleted from these lists by a custom PHP script to remove these feature candidates with less distinctive value (because of these features’ presence in two or more publication types). By removing these doubles we aimed to increase the selection of features with more distinctive value. After removing these doubles the top 100 lists (ranked by tf*idf score in descending order) were
reduced significantly and the number of keywords – feature candidates - remaining in each publication type was at least halved. In Table 5.4 we can see that the Case Reports publication type has the highest number of keywords remaining with 45 candidate features. The publication type Randomized Controlled Trial has the lowest number of candidate features remaining with 13 candidate features left.

**Table 5.4**  
*Number of keywords after feature transformations per publication type*

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Number of keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review</td>
<td>37</td>
</tr>
<tr>
<td>Randomized Controlled Trial</td>
<td>13</td>
</tr>
<tr>
<td>Comparative Study</td>
<td>20</td>
</tr>
<tr>
<td>Multicenter Study</td>
<td>17</td>
</tr>
<tr>
<td>Case reports</td>
<td>45</td>
</tr>
</tbody>
</table>

Since the publication type Randomized Controlled Trial only contained 13 keywords (feature candidates) and equal class representations were to be maintained every publication type was represented by the top 13 keywords according to their tf*idf score calculated earlier. By extending the number of features per publication type but maintaining an equal number of features per class we aimed for a better set of representational features.

Consequently, the final vector for each abstract consisted of 65 features plus the class (publication type) at the end. The set of 11,399 vectors that were generated using this feature set served as input to the classifier. The (most frequent class) baseline (Jurafsky & Martin, 2009) for this experiment to compare the classifier’s performance to is 36.9 %. Our classifier should therefore at least label 36.9% documents correctly to make the results meaningful. From the results shown in Table 5.5 we can see that the performance of both classifiers have degraded. In comparison with the last experiment the Naïve Bayes classifier precision has dropped more than four percent and recall has dropped almost ten percent. A paired-samples t-test was conducted to evaluate the decrease in recall of the Naïve Bayes classifier. There
was a statistically significant decrease in recall for the Naïve Bayes classifier from Classit AA ($M=58.58, SD=1.63$) to Classit AA tf*df (with statistical weighting) [$M=49.58, SD=1.90, t(9), p<.0001$]. We also conducted a paired-samples t-test to evaluate the decrease in recall of the SMO classifier. There was a statistically significant decrease in recall for the SMO classifier from Classit AA ($M=62.39, SD=1.11$) to Classit AA tf*df (with statistical weighting) [$M=58.79, SD=2.17, t(9), p<.01$]. Overall performance deteriorated. This is a result opposite to our expectations. Extending the feature set and adding feature transformation in the form of tf*idf weighting does not improve the performance of the classifier.

Table 5.5 Performance of Classit AA tf*idf

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classit AA</td>
<td>0.59</td>
<td>0.60</td>
<td>0.56</td>
<td>0.49</td>
</tr>
<tr>
<td>Classit AA tf*idf</td>
<td>0.50$^c$</td>
<td>0.64</td>
<td>0.61</td>
<td>0.57</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.62</td>
<td>0.59$^a$</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td>SMO</td>
<td>0.59$^a$</td>
<td>0.64</td>
<td>0.61</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Note: superscript indicates significance of difference with column to the left

a: $p < .01$; b: $p < .05$; c: $p < .0001$;

The most frequent class baseline equals .369 (36.9%)

The approaches used up to this point use frequencies of words for feature selection and tf*idf for feature set transformation (term weighting). We conclude from the performance results of the classifiers that the feature set representations as selected using the frequency and tf*idf approach are not able to represent their classes (publication types) so that the classifiers can learn from these examples and build a model to accurately classify unlabeled instances. Sparsity of the abstracts – the abstracts are only 154.51 words in length on average – might be another explanation. Finally, one of the factors contributing to the poor results may be the amount of noise and subject related terms (due to our data set that is not matched on subject) that are still not accounted for. In our next experiments we will use stemmed abstracts to select feature candidates that do not suffer from the potential side effect of our broad data set,
while also improving the source for vector scoring because the stemmed features will appear more often in our stemmed abstracts and thus provide more information to the classifiers.

The approaches investigated so far are single-word approaches. A potential strategy for improvement of performance is to use combinations of two neighbouring words (bigrams) rather than single words as features. Through making use of bigrams instead of single words we hope to find features with more distinctive value, even though bigram features are less prevalent than single words.

5.2.1 Bigram features

Before starting the process of selecting bigrams as features all abstracts in the data set were stemmed using Porters Stemmer. Secondly the procedures for creating frequency lists for every publication type were modified. Instead of searching and listing single words every set of two neighbouring words (bigrams) was counted and added to the frequency list. Again, only the top 100 occurring bigrams were selected as a cut-off threshold level. Bigrams that would also appear within other classes/publication types were automatically removed and for each class within the dataset thirteen bigrams were selected as features. Although we performed no experiments with other numbers of features per class we chose to maintain the number of thirteen features per publication type.

The selection of these bigram features was based on distinctive value by the researcher’s standards. For the publication type ‘Review’ a feature like ‘treatment patient’ does not seem to contain as much distinctive value as does ‘review literatur’ (stemmed bigram). The final vector for each abstract contained 65 features with the class (publication type) at the end. The 11,399 vectors that were created for all abstracts served as input into the classifier.

Table 5.6 shows that the bigram approach as described above is not an improvement to earlier approaches investigated looking at the recall score which has dropped to 0.38. A paired-samples t-test was conducted to evaluate the impact of the use of bigrams on the recall score of the Naïve Bayes classifier. There was a statistically significant decrease in recall for the Naïve Bayes classifier from Classit AA ($M=58.58$, $SD=1.63$) to Classit AA using bigram

---

5 Available from [http://tartarus.org/~martin/PorterStemmer/](http://tartarus.org/~martin/PorterStemmer/)
features \[ M=37.46, SD=2.10, t(9), p<.0001 \]. Also, a paired-samples t-test was conducted to evaluate the impact of the use of bigrams on the recall score of the SMO classifier. There was a statistically significant decrease in recall for the SMO classifier from \( M=62.39, SD=1.11 \) to Classit AA using bigram features \[ M=39.08, SD=1.36, t(9), p<.0001 \]. Further analysis of the classifier performance results show that especially the Multicenter Study deteriorates the results with a precision score of zero and a recall score that is zero. Similar results were found for the Case Reports publication type. Moreover this approach barely performs better than the calculated most frequent class baseline of 36.9 percent with recall scores of 0.38 and 0.39 for the Naïve Bayes and SMO classifiers.

Table 5.6 Classit AA bigrams compared with Classit AA

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classit AA</td>
<td>Classit AA bigrams</td>
<td>Classit AA</td>
<td>Classit AA bigrams</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.59 (^c)</td>
<td>0.38 (^c)</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td>SMO</td>
<td>0.62 (^c)</td>
<td>0.39 (^c)</td>
<td>0.64</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Note: superscript indicates significance of difference with column to the left
a: \( p < .01 \); b: \( p < .05 \); c: \( p < .0001 \);

The most frequent class baseline equals .369 (36.9%)

These results seem to point out two problems:

1.) Our feature set is not appropriate for classifying abstracts or,

2.) The abstracts are too sparse to provide the classifier with vectors that contain enough distinctive frequency information.

Even though the abstracts contain 154.51 words on average (including stop words), the hypothesis that through statistical analysis the methodological terms will come up as differing between publication types seems false. Similarly, although the selected features do appear in
abstracts - often more than once - these ‘hits’ contain too little distinctive value for the classifier to build a proper statistical model.

The next step is to find a better set of representational features. We will attempt to do so by introducing full articles and repeating our feature selection procedures on these full articles. By extending our methods of classifying by testing the approaches investigated so far on full articles we will test the hypothesis that the abstracts are too sparse and that the features selected from the abstracts are not appropriate.

**5.3 Classit FA**

**5.3.1 Introducing full articles**

For each publication type fifteen full articles were introduced. The procedure of feature selection was the same as on earlier occasions. Stop words were removed and the full articles were stemmed using Porter’s Stemmer before importing the full articles in our database. For each publication type the top 100 frequency lists of bigrams were created with a custom PHP script and bigrams appearing in other publication types were automatically removed. Thirteen bigrams per publication type were selected as features and these features were used to score each abstract. The set of 11,399 abstract vectors each consisting of 65 features plus the publication type at the end served as input to the classifier.

In Table 5.7 the results are shown for the two classifiers using 13 features per publication type, selected from the full articles but tested on the abstracts. Table 5.7 shows that the results have deteriorated significantly compared to the best performing strategy (Classit AA) yet. A paired-samples t-test was conducted to evaluate the impact of the selection and use of full article features on the recall score of the Naïve Bayes classifier. There was a statistically significant decrease in recall for the Naïve Bayes classifier from Classit AA \( M=58.58, SD=1.63 \) to Classit FA \( M=39.01, SD=1.99, t(9), p<.0001 \). A paired-samples t-test was also conducted to evaluate the impact of the selection and use of full article features on the recall score of the SMO classifier. There was a statistically significant decrease in recall for the SMO classifier from Classit AA \( M=62.39, SD=1.11 \) to Classit FA \( M=39.89, SD=2.01, t(9), p<.0001 \).
Further analysis also shows that the Randomized Controlled Trial and Multicenter Study publication types have performance (precision, recall) scores of zero. This again seems to point towards the problem of sparsity within the abstracts and that abstracts therefore are unable to provide the classifier with enough distinctive frequency information on the feature sets used. This seems logical since the single word approach already did not achieve good results and bigrams are even less prevalent than single words. Table 5.7 also shows that the Naïve Bayes classifier’s recall score is only 0.8% higher than our most frequent class baseline of 38.2%.

Table 5.7 Classit FA

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classit AA</td>
<td>Classit FA</td>
<td>Classit AA</td>
<td>Classit FA</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.59</td>
<td>0.39(^c)</td>
<td>0.60</td>
<td>0.32</td>
</tr>
<tr>
<td>SMO</td>
<td>0.62</td>
<td>0.40(^c)</td>
<td>0.64</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Note: superscript indicates significance of difference with column to the left
a: \(p < .01\); b: \(p < .05\); c: \(p < .0001\);

The most frequent class baseline equals .382 (38.2%)

5.4 Classit FF

To test the hypothesis that sparsity of abstracts is responsible for low classifier performance the next experiment was to classify the full articles using the features selected from the full articles.

After discarding bigrams that would also appear in other publication types’ frequency lists we now - different from earlier approaches - automatically selected the top 40 bigrams as features to make sure no bias was introduced as a consequence of researcher’s choices. We chose to extend the number of features per publication type to see if any improvements could be
achieved with a larger feature set. At the same time this does hamper the comparability of the results with earlier experiments.

With forty features per publication type and fifteen vectors (consisting of 5*40 features) per publication type 75 vectors in total were created. These served as input to the classifier. The results of this experiment show a lot of improvement on earlier methods with a precision score of 0.78, a recall score of 0.60 and a AUC score of .87 using a Naïve Bayes classifier. The SMO classifier shows no clear improvement or deterioration of the results.

A paired-samples t-test was conducted to evaluate the impact of the selection and use of full article features on the recall score of the Naïve Bayes classifier. There was a statistically significant increase in recall for the Naïve Bayes classifier from Classit FA ($M=39.01$, $SD=1.99$) to Classit FF [$M=62.50$, $SD=14.11$, $t(9)$, $p<.01$]. A paired-samples t-test was also conducted to evaluate the impact of the selection and use of full article features on the recall score of the SMO classifier. There was no statistically significant increase in recall for the SMO classifier from Classit FA ($M=39.89$, $SD=2.01$) to Classit FF [$M=43.57$, $SD=23.81$, $t(9)$, $p=.63$].

### Table 5.8 Performance of Classit FF compared to Classit FA

*40 full article features per publication type. 75 instances, ten-fold cross validation*

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classit FA</td>
<td>Classit FF</td>
<td>Classit FA</td>
<td>Classit FF</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.39</td>
<td>0.60$^a$</td>
<td>0.32</td>
<td>0.78</td>
</tr>
<tr>
<td>SMO</td>
<td>0.40</td>
<td>0.49</td>
<td>0.47</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Note: superscript indicates significance of difference with column to the left
a: $p < .01$; b: $p < .05$; c: $p < .0001$

*The most frequent class baseline equals .20 (20%)*

Further analysis of the features and outliers reported by the classifier model shows that the feature set contains numerous features (bigrams) which are subject/research related and concern disease names, medicines or disease related bacteria. These features are listed within...
the top 40 frequenting bigrams but hold no distinctive value with regard to publication type and only create noise within the feature set. Moreover, these subject-related bigrams seem to result in the higher performance scores of the Naïve Bayes classifier. These results are therefore not representative and seem to reflect overfitting in this set of full articles. It is unlikely that these results can be achieved using a different data set.

Through creating a higher cut-off threshold level and using fewer features we may be able to filter the noise and subject-related terms from the (candidate) feature set. Therefore we performed an experiment using a set of 20 features per publication type. However, Table 5.9 shows that the results only have deteriorated and still seem to suffer from overfitting since the Naïve Bayes classifier yields much higher results than SMO classifier, possibly because of higher flexibility of this classifier. Precision has only decreased with two percent and the AUC score is still 0.80. The SMO classifier precision score is 13 percent lower and recall has decreased with 9 percent. Together with the analysis of the features selected these results show that noise also resides in the top positions of the frequency lists. We conclude that creating a higher cut-off threshold level of 20 features per publication type is not an appropriate way of filtering and removing noise.

A paired-samples t-test was conducted to evaluate the impact of the size of the feature set on the recall score of the Naïve Bayes classifier. There was no statistically significant difference in recall for the Naïve Bayes classifier between Classit FF using 20 features ($M=54.64, SD=15.98$) and Classit FF using 40 features ($M=62.50, SD=14.11, t(9), p=.32$). A paired-samples t-test was also conducted to evaluate the impact of the size of the feature set on the recall score of the SMO classifier. There was no statistically significant difference in recall for the SMO classifier between Classit FF using 20 features ($M=37.14, SD=9.21$) and Classit FF using 40 features ($M=43.57, SD=23.81, t(9), p=.35$).
The next step was to improve the candidate feature set by making use of feature set transformations through statistical weighting. The goal of this additional feature transformation method was to remove noise in the form of subject related bigrams. As was done on earlier occasions the tf*idf statistical weighting was added for feature selection. Again we selected the top 40 features per publication type after removing bigrams that would also appear in other publication types. These 40 selected features had the highest scores after summing all individual tf*idf values within each publication type.

Table 5.10 shows that the results have degraded using this approach. The results are worse compared to the last strategy without statistical weighting and 20 features per publication type. The Naïve Bayes classifier shows a precision score that is 37 percent lower and a recall score that has decreased with almost 30 percent. The SMO classifier results show that precision has decreased with almost 20 percent and recall has decreased with twelve percent.

A paired-samples t-test was conducted to evaluate the impact of statistical weighting on the recall score of the Naïve Bayes classifier. There was a statistically significant decrease in recall for the Naïve Bayes classifier between Classit FF using 40 features and statistical weighting ($M=41.07$, $SD=21.08$) and Classit FF using 40 features without statistical weighting [$M=62.50$, $SD=14.11$, $t(9)$, $p<.05$]. A paired-samples t-test was also conducted to evaluate the impact of statistical weighting on the recall score of the SMO classifier. There was no statistically significant difference in recall for the SMO classifier between Classit FF using 40 features and statistical weighting.
features and statistical weighting ($M=33.57$, $SD=19.90$) and Classit FF using 40 features without statistical weighting ($M=43.57$, $SD=23.81$, $t(9)$, $p=.23$).

Table 5.10 Classit FF tf*idf compared with Classit FF using 40 features and no statistical weighting

40 full article features with statistical weighting per publication type. 75 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classit FF 40</td>
<td>Classit FF</td>
<td>Classit FF 40</td>
<td>Classit FF tf*idf</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.60</td>
<td>0.33$^b$</td>
<td>0.78</td>
<td>0.41</td>
</tr>
<tr>
<td>SMO</td>
<td>0.49</td>
<td>0.37</td>
<td>0.61</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Note: superscript indicates significance of difference with column to the left

a: $p < .01$; b: $p < .05$; c: $p < .0001$

The most frequent class baseline equals .20 (20%)

Probably this poor result is due to the effect we saw in the previous experiment – subject related terms were ranked higher and favoured. Due to the size of the data set this effect is even stronger with the use of statistical weighting.

Feature analysis shows that different subject related bigrams rank even higher after statistical weighting. These terms are distinctive to a research report, but because of the small (total) data set they are also favoured as distinctive to a publication type which is false. Due to the very small sample (fifteen articles per publication type) these features are ranked as distinctive to that small corpus. When we look at the features that were selected using this approach, features like the following are in the top 40 ranked on their total tf*idf score: “estrogen receptor”, “neoforman var”, “h pylori”.

These features seem very distinctive at first but they are distinctive to research report rather than publication type. They describe random noise – subject related bigrams – instead of bigrams related to publication type. One of the full articles in the data set is titled “Accuracy
of the stool antigen test in the diagnosis of *Helicobacter pylori* infection before treatment and in patients on omeprazole therapy.” We can see the bigram ‘h pylori’ listed in the top three of (what should be) distinctive bigrams.

By selecting features from the set of articles that is also used for classification the high(er) performance scores of the experiments using full-articles features are not representative and the result of overfitting even though ten-fold cross validation was used for classification. By using the same small data set for feature selection and classification procedures the results cannot be generalized. It is unlikely that similar results will be achieved when this model is tested on a different data set.

The experiments using full article features did not yield any promising results. From the experiments performed we can conclude that the size of the full article sample is too small to offer a good base for feature selection and classifying. The experiments using features selected from the full articles demonstrated that using the proposed method it is not possible to select good features and classify full articles or abstracts. These representations contain too much noise and are distinctive to the different research reports but not to the publication types. The expectation that publication type related terms (methodological terms) would come up proves to be false when selecting features from a too small number of full articles.

### 5.5 Classit AF

The final experiment used features selected from abstracts to classify full articles. The set of features that was used in the bigram approach – selected from the abstracts and tested on the abstracts - was now used to create the vectors for 50 full articles. These are the bigrams selected from the set of 11,399 abstracts. To see if any significant improvements could be made by extending the data set the first experiment was performed using ten full articles per category. Fifty vectors (ten per publication type) each consisting of 65 features followed by the class were created and served as input to the classifier.

The classification of full articles using features selected from abstracts seems to cause an improvement compared to the approaches investigated so far. Moreover, this approach is less likely to suffer from overfitting due to the larger sample of abstracts the features were selected
from. The results are shown in Table 5.11. The Area Under the Curve now scores .81 for SMO which is a good result. However, we should keep in mind the precision and recall scores are still rather low: from the retrieved documents only 58% is classified correctly and it is possible that relevant documents (publication types) are not even part of the set of retrieved documents.

Table 5.11 Performance of Classit AF with 50 instances

13 abstract features per publication type. 50 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.56</td>
<td>0.58</td>
<td>0.56</td>
<td>0.78</td>
</tr>
<tr>
<td>SMO</td>
<td>0.58</td>
<td>0.64</td>
<td>0.59</td>
<td>0.81</td>
</tr>
</tbody>
</table>

The most frequent class baseline equals .20 (20%)

To test the influence of sample size the data set was extended with another five full articles per publication type. The results from this experiment show that the Naïve Bayes classifier achieves better results for precision, recall and Area Under the Curve. The SMO classifier only shows a small improvement in the Area Under the Curve (improvement of two percent) when using fifteen full articles per publication type rather than ten.

A paired-samples t-test was conducted to evaluate the impact of the sample size on the recall score of the Naïve Bayes classifier. There was no significant difference in recall for the Naïve Bayes classifier from Classit AF with 50 full articles ($M=60.00, SD=355.56$) to Classit AF with 75 articles [$M=66.78, SD=20.36$, t(9), p=0.41]. We also conducted a paired-samples t-test to evaluate the impact of the sample size on the recall score of the SMO classifier. There was no significant difference in recall for the SMO classifier from Classit AF with 50 full articles ($M=40.00, SD=21.08$) to Classit AF with 75 articles [$M=56.25, SD=19.18$, t(9), p=0.09].
Table 5.12 Performance of Classit AF compared to Classit AF with 50 instances
13 abstract features per publication type. 75 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classit AF 50</td>
<td>0.56</td>
<td>0.67</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td>Classit AF 75</td>
<td>0.58</td>
<td>0.67</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.56</td>
<td>0.67</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td>SMO</td>
<td>0.58</td>
<td>0.67</td>
<td>0.64</td>
<td>0.57</td>
</tr>
</tbody>
</table>

The most frequent class baseline equals .20 (20%)

A paired-samples t-test was conducted to evaluate the impact of source for classification on the recall score of classifiers. There was no significant difference in recall for the Naïve Bayes classifier between Classit AA (M=58.58, SD=1.63) and Classit AF [M=66.78, SD=20.36, t(9), p=.24]. A paired-samples t-test was also conducted to evaluate the impact of source for classification on the recall score of the SMO classifier. There was no significant difference in recall for the SMO classifier between Classit AA (M=62.39, SD=1.11) and Classit AF [M=56.25, SD=19.18, t(9), p=.34].

Table 5.13 Performance of Classit AF compared to Classit AA
13 abstract features per publication type. 75 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classit AA</td>
<td>0.59</td>
<td>0.67</td>
<td>0.60</td>
<td>0.67</td>
</tr>
<tr>
<td>Classit AF</td>
<td>0.62</td>
<td>0.56</td>
<td>0.64</td>
<td>0.57</td>
</tr>
</tbody>
</table>

The most frequent class baseline equals .20 (20%)

The strategy using features selected from abstracts to classify full articles can be considered the most successful strategy. We conclude that the abstracts offer the best source for feature selection but are too sparse to provide distinctive vectors. However, due to the small size of
the full article collection the full articles do not provide enough information for good feature selection. In contrast, the full articles provide enough information to create distinctive vectors for the classification task.
6 Discussion and conclusions

This thesis investigated the possibilities for text classification of medical research reports to provide clinicians with an evidence score that could be used as an alternative ranking of search results. The research question that was formulated was: To what extent is it possible using a word-based approach to automatically classify medical publications in terms of level of evidence?

The aim was to classify abstracts based on their semantical contents since full article access is often restricted and short bits of text are all that is available. In answering this research question four variants of the Classit method were developed and evaluated, defined by their source of feature selection and their target for classification. The first experiment evaluating Classit AA showed that it was not possible to achieve good results classifying abstracts with the proposed approaches. To test the assumption of data sparseness within abstracts two alternative variants of Classit were evaluated namely using features selected from full articles to classify abstracts (Classit FA) as well as using these features to classify full articles (Classit FF). Single words as well as bigrams using stemming and statistical weighting as feature transformation methods functioned as representational features. The experiments using full article features showed that our data set of full articles is too small to yield reliable results. The higher performance scores achieved were biased since they result from training and testing within the same small set. Feature analysis also confirms this conclusion with subject-related terms appearing in the feature sets.

The results from our last experiment classifying full articles using features selected from abstracts (Classit AF) show that it seems possible to classify full articles using a bigram word-based approach. The increase of performance gained by extending the dataset to 15 articles per publication type shows that further extension is promising. With an AUC score of .85 and an increase in precision of nine percent using a Naïve Bayes classifier this approach achieved the best and most reliable results. These results imply that abstracts are appropriate sources for feature selection, but are too sparse for classification procedures whereas full articles are able to provide enough information for the classification procedures. The two experiments classifying abstracts using abstract features in the form of single words and bigrams (Classit
A satisfactory word-based approach for scoring medical publications with evidence is yet to be found. Although four different approaches were explored, none of them led to satisfactory results and all approaches seemed to lead to the same conclusion: abstracts are too sparse for a purely word-based approach that is able to score for evidence. The experiment on Classit AF using features selected from the abstracts to classify both abstracts and full articles supports this conclusion. The approach does achieve good results in classifying full articles. However, finding an approach that is able to classify abstracts or smaller bits of text would be a valuable improvement to the ranking of search results. Whereas full articles are not always available through automated procedures, short snippets of text often are present. Further research should therefore focus on finding a better set of representational features within the abstracts or focus on extending the data set for vector scoring and classification procedures.

We conclude that words or combinations of words are not flexible enough. These features in combination with the sparseness of abstracts and the low frequencies of features therein do not provide enough distinctive value for training a classifier model that is able to achieve satisfactory results.

We expected that methodological terms distinctive to publication type would prevail after statistical analysis but these expectations have proven to be false. The experiments using feature selection from abstracts showed several issues. First, although methodological terms distinctive to publication type did occur in the lists of most frequent words and bigrams for each publication type, these lists also contained a lot of medical terms and combinations of terms (for the bigram approach) which are of no distinctive value since they appear in all publication types. The automatic deletion of these ‘doubles’ led to a reduction of candidate features but at the same time revealed the main issue: sparsity of data within the abstracts. Candidate features were found, but their frequency in abstracts is too low to be of enough informative value to the classifiers. This is well illustrated by the experiments where classification is done both with abstracts and with full articles. Unsatisfactorily, only the
experiments classifying full articles achieved good results, hereby showing that the single word and bigram approaches seem to be unable to yield good results on this classification task. The largest limitation to our research is the size of the data set and the unequal representation of publication types within the data set.

6.1 Further research

To investigate where future research could gain interesting improvements we also performed two additional experiments. Since the experiments performed in this thesis used a relatively small amount of features we wanted to explore the possibilities of using all possible features that were available in our data set. The first experiment using all features investigated the performance of all features selected from full articles to classify full articles while the second experiment investigated the performance of all possible features selected from abstracts to classify the abstracts. All features means that every word (with the exception of stop words) was selected as a feature and no cut-off threshold levels were used.

For the full article – full article strategy (Classit FF) this resulted in a set of 19,558 features from the 75 full articles. These features were used to classify the 75 full articles using ten-fold cross validation. We can see from Table 6.1 that using the approach of a large set of features does not lead to any improvements for the Classit FF approach compared to the experiments performed with smaller feature sets. We should bear in mind that the amount of examples (number of medical publications) is still an important limitation to the reliability of these results. Table 6.1 is shown on the next page.
Table 6.1 Classit FF using all possible features compared to Classit FF using 40 features
19,558 full article features, 75 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classit FF 40</td>
<td>0.60</td>
<td>0.47</td>
<td>0.54</td>
</tr>
<tr>
<td>Classit FF all features</td>
<td>0.64</td>
<td>0.78</td>
<td>0.61</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.59</td>
<td>0.60</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The most frequent class baseline equals .20 (20%)

We also investigated the performance of this approach selecting all features from the abstracts and using this set of 41,935 features to classify abstracts (Classit AA). This approach achieves lower performance than the approach using features selected from full articles to classify full articles (Classit FF). The better results achieved with the Classit FF approach could again be the result of over-fitting, because of the small amount of full articles in our data set.
Unfortunately we are not able to report on the performance of the SMO classifier since the classifier was still running (for more than a week) while the deadline for this thesis was in sight.

Table 6.2 Classit AA using all possible features compared to Classit AA
41,935 abstract features, 9,642 instances, ten-fold cross validation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classit AA</td>
<td>0.59</td>
<td>0.60</td>
<td>0.50</td>
</tr>
<tr>
<td>Classit AA all features</td>
<td>0.56</td>
<td>0.46</td>
<td>0.39</td>
</tr>
</tbody>
</table>

The most frequent class baseline equals .393 (39.3%)
From the two exploratory experiments using all possible features we can conclude that the strategy using a large amount (all) of features does not lead to satisfying results either. It is very well possible that the ideal feature-set size lies somewhere in between the small amounts of features used in this thesis and the other extreme: all possible features. An important side note is that these additional experiments were performed using a small set of medical publications for the FF strategy, and a data set that was not evenly distributed across all publication types for the AA strategy. Further research should therefore first focus on extending the data set to evaluate performance of the current approach on a larger and better balanced data set. The ideal size of the feature-set also depends on the number of examples (medical publications) provided to the classifier procedures. Analysis of performance per individual publication type shows that for example the publication type ‘Review’ (the publication type with the most instances) achieves a precision of 0.72, which is a reasonable result, while publication types represented by very few instances deteriorate the overall precision significantly. I.e. the publication type Multicenter Study scores 0.06 for precision.

The objective of a generic and widely applicable approach seems to have a high cost. The sole use of words and bigrams as features leaves little room for alternative strategies and improvement of performance, although an important limitation to our study is the size of our data set. Because of the size of our data set and because the publication types are unequally represented (for the abstracts) it is difficult to draw definitive conclusions. Individual publication types better represented by more examples seem to achieve reasonable scores looking at the Review publication type scoring 0.72 for precision which is a promising result. A larger and equal sample of abstracts per publication type might result in better feature selection and higher overall classifier performance. Due to limits in resources and time we were not able to perform experiments with a larger data set.

It would therefore be interesting what performance could by achieved if all publication types were represented by larger amounts of examples. Moreover it would be interesting to experiment with a bigger variety of feature-set sizes. Neither the extremely small amount of features nor the extremely large amount of features seem to be the solution. Perhaps the ideal feature-set size lies somewhere in between although – as noted earlier – this ideal size is very likely to also be dependent on the size of the data set.
Apart from extending and creating a better balanced data set further research should also focus on improving the feature representations. This could be along the lines of expanding the feature set by using alternative features. Alternative features such as number of citations - known as citation analysis (Anderson, 2009) - number of authors, journal source combined with publication date could all provide starting points for alternative feature sets. Another potentially useful alternative may be using a metathesaurus such as the UMLS to translate medical concepts named differently in different medical subjects but similar in meaning. The use of additional features that are partly domain or data set specific may present an interesting trade off between input of time and expertise and improvements of performance.

Another starting point for further research could be along the lines of automatic ways for extending and expanding the data set. Phan and colleagues (Phan et al., 2008) have done promising research on the classification of short texts using (unlabeled) background knowledge in the form of longer documents that are available online (Zelikovitz et al., 2000). A promising alternative to the similar principle of query expansion by using large collections of data available through the world wide web could perhaps serve as a solution to the problem of sparsity of abstracts.
7 References


De Bruijn, B., and Martin, J. (2002). Getting to the (C)ore of Knowledge: Mining Biomedical Literature (pp. 7 O 18). International Journal of Medical Informatics. Vol. 67 1-3.


Haynes, R.B. (2002). *What kind of evidence is it that Evidence-Based Medicine advocates want health care providers and consumers to pay attention to?*  
http://www.biomedcentral.com/1472-6963/2/3


[http://www.cebm.net/?o=1025](http://www.cebm.net/?o=1025)


