Finding Patient-Oriented Evidence in PubMed Abstracts

by

David Alexander Robinson

(Under the direction of Michael A. Covington)

Abstract

This project develops a computational method to improve searches of the medical literature by selecting the studies that report reliable evidence of patient-oriented outcomes. These outcomes include morbidity, mortality, symptom severity, and quality of life. Four machine learning methods, Support Vector Machines, naïve bayes, naïve bayes multinomial and logistic regression, achieve over 70% accuracy on the identification of such studies in PubMed abstracts. The accuracy attainable by hand in this task is about 95%. The best machine learning results, just over 80% accurate, were obtained with naïve bayes multinomial on a combination of single words and contiguous pairs of words using WEKA 3.7.5.

Index words: Machine Learning, Evidence-Based Medicine, Text Classification, Logistic Regression, Support Vector Machines, Bayesian Learning

Finding Patient-Oriented Evidence in PubMed Abstracts

by

David Alexander Robinson

B.A., University of Georgia, 1999

M.S., Georgia Institute of Technology, 2004

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2012

Finding Patient-Oriented Evidence in PubMed Abstracts

by

David Alexander Robinson

Approved:

Major Professor: Michael A. Covington

Committee: Mark H. Ebell
Walter D. Potter

Electronic Version Approved:

Dr. Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

Finding Patient-Oriented Evidence in PubMed Abstracts

David Alexander Robinson

July 22, 2012

Acknowledgments

I dedicate this thesis to my parents.

Contents

List of Tables

Chapter 1

Background

1.1 Introduction

This project develops a computational method to improve searches of the medical literature by selecting the studies that report reliable evidence of patient-oriented outcomes. Such a study is called Patient-Oriented Evidence (POE). The goal of this thesis is to automate the identification of POEs. This is one step toward the longer term goal of developing a computational method of selecting Patient-Oriented Evidence that Matters (POEMs) from the medical literature. POEMs are the most important medical articles for general practitioners because they represent studies with the potential to change practice and improve outcomes.

The present study is restricted to articles in PubMed, which contains more than 23 million abstracts. Over 800,000 articles are abstracted and added to PubMed per year. It can be queried by anyone on the Internet based on MeSH terms in certain fields. These include **title**, **author**, **first author**, **publication date**, and **publication type**, **abstract**, or **title and abstract (tiab)**.[1] The MeSH terms include clinical terms that are not contained in

[1]PubMed can be accessed at www.ncbi.nlm.nih.gov/pubmed/.

a small stop words list.[2] There is also a National Institute of Health (NIH) search engine, which can translate common idioms into the appropriate MeSH terms.[3]

Doctors who want the best clinical outcome for their patients may search PubMed. However, PubMed's coverage is wide, so it may return many non-relevant results. My goal was to apply machine learning techniques to the task of identifying reports with patient-oriented outcomes based on reliable evidence, *i.e.*, POEs.

To achieve practical results in a reasonable amount of time, I decided not to apply syntactic analysis to the abstract text as part of my classification algorithms. That is, the methods considered were limited to shallow classification schemes. A number of different methods were tested experimentally, since there is no single approach which works well for all text classification problems. So a best method (among those tested) emerged empirically.

Prior to the experiments I had five expectations for the best classification scheme, as follows.

(i) Its accuracy should be at least 60% in POE identification. This is based on the findings of prior studies on shallow classification of biomedical texts.

(ii) It should enable the user to trade off between reducing false positives and reducing false negatives. This can normally be done in text classification problems. The trade off would enable a doctor to cast a wider net for the answer to a query by generating more abstracts to scan.

(iii) Bayesian methods would be uncompetitive with logistic regression and support vector machines, based on prior experience reported in the text classification literature.

(iv) Feature selection would be important in determining the best classification scheme for POE identification. This is normal for text classification problems.

[2]The complete list of PubMed stop words can be found on the website www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/?report=objectonly.

[3]MeSH is described at www.ncbi.nlm.nih.gov/mesh.

(v) Algorithms based on binary feature data were expected to outperform those based on feature frequencies. This would be normal for classification of short texts, and PubMed abstracts are short by the standards of text classification research.

The experiments turned up several surprises. The best results were obtained with naïve bayes multinomial (NBM), a method which employs frequency data. With suitable feature selection NBM achieved 80.4% accuracy in POE identification. This was significantly better than any of the other methods studied. NBM did allow for trading off between false negatives and false positives over a fairly wide range with no significant loss of accuracy. As expected, feature selection was critical to achieving the best results. The highest accuracy was attained by using words and consecutive word pairs as features, and including only those which occurred in at least 6 documents and afforded an information gain of at least 0.0010.

1.2 Motivation

It is believed that POE identification by computer is novel and significant. This is important because of the focus on the type of outcomes; physicians should be able to restrict their search to a corpus of abstracts reporting Patient-Oriented Evidence. By hand POE Identification takes 2–3 minutes per abstract. Considering that 800,000 or so abstracts are added to PubMed annually, POE identification by computer can save at least a dozen man-years. Thus, POE identification by computer can help doctors find POEMs. This will advance one of the key goals in medicine today, which is to aid medical practitioners in finding the evidence in the literature that they need to treat their patients. Clinical queries in PubMed only focus on study design. The important improvement in this thesis is the new focus on the type of outcome.

Simple searches are likely to provide unsatisfactory results. For instance, PubMed searches for Randomized Controlled Trials (RCTs) return many results based on unreliable evidence,

not based on randomized controlled trials. Even searches for studies on humans return some studies on mice and cows.

The evidence must also measure, as a primary or secondary outcome, things patients care about. Otherwise, the results of the study may not be relevant. There are numerous examples from the literature of drugs and treatments that improve an intermediate marker, but do not help the overall patient-oriented outcome. This includes keeping hemoglobin A1C under 6.5% for diabetics, according to the American Diabetic Association's 2010 guidelines. Keeping hemoglobin A1C under 6.5% in diabetics with intensive insulin therapy does not improve mortality over time. Thus the regime of intensive insulin therapy for diabetic patients is unwarranted. Also, aggressively lowering Hemoglobin A1C in Type II diabetics does not lower the risk of mortality. This is an example of a POEM, since it is a POE which modifies the recommended treatment for diabetics. The earlier studies were not POEMs, and, as it turned out later, were misleading. See the ACCORD, ADVANCE and VA Diabetes trials for better Hemoglobin A1C targets (LaBlanche *et al.* 1997, Kirkman *et al.* 2006, Du *et al.* 2009).

1.3 Problem Description

This thesis's topic is finding valid patient-oriented evidence (POE) in PubMed abstracts, as introduced in section 1.1. POEs consist of papers from the medical literature with reliably measured outcomes that patients care about. These are the papers which satisfy criteria (i) and (ii):

(i) the findings are based on reliable evidence;

(ii) the outcomes are directly relevant to patients;

(iii) the recommendations change existing practice for doctors.

Thus, an abstract is a POEM if and only if it is a POE which satisfies criterion (iii). Conditions (i) and (ii) are discussed below in order.

First, the study must be based on reliable evidence. The most valid evidence comes from well designed RANDOMIZED CONTROLLED TRIALS (RCTs), in which the patients and researchers are both masked, allocations are concealed, and placebos are used as controls. Careful meta-analyses based on several high-quality RCTs can also provide reliable evidence.

Second, the outcomes must be patient-oriented. Patient-oriented results include mortality, morbidity, bleeding, cardiac events such as heart attacks, length of stay in the hospital, and costs. Pain scales and quality of life indicators are patient-oriented. Health related quality of life is patient-oriented, as are numerous quality of life indicators related to a specific body part or disease. Simply taking someone to the hospital for a certain length of time, or performing a procedure, is not patient-oriented. The counts of HIV-1 or other biochemical markers or intermediates in the body, such as triglycerides and cholesterol, are not patient-oriented. Changes in height and weight are not patient-oriented in our study. A delay in getting to a hospital or in obtaining care is not considered patient-oriented if final outcomes are not affected.

Below are examples of outcomes which are patient-oriented followed by examples which are not patient-oriented.

(a) mortality, morbidity, or bleeding — patient-oriented

(b) cardiac events, including adverse cardiac events — patient-oriented

(c) pain scales — patient-oriented

(d) quality of life, including quality of life of a certain body part or disease — patient-oriented

(e) duration of hospital stay — patient-oriented

(f) changes in height or weight — not patient-oriented

(g) delay in treatment, if patient-oriented outcomes are not affected — not patient-oriented

(h) intermediate markers, like blood pressure, HIV-1, or hemoglobin A1C — not patient-oriented

(i) secondary or safety outcomes, even if they are patient-oriented — not patient-oriented

Not all studies reporting patient-oriented outcomes meet condition (i). The acronym POE suggests condition (i) by mention of evidence (as opposed to outcome), as was done in the earlier acronym POEM.

A POEM requires also that the recommendations of the study must change current medical practice. This can not be determined without extensive clinical experience, so condition (iii) is not considered here.

The problem for this study is to automate the accurate classification of PubMed abstracts as POE or non-POE. Inaccuracies may be of two types: a TYPE I ERROR is also known as a FALSE POSITIVE (FP). This is a paper that is a non-POE, but is classified as a POE. A TYPE II ERROR, also known as a FALSE NEGATIVE (FN), is a paper that is a POE, but is classified as a non-POE. A TRUE POSITIVE (TP) is a paper that is correctly classified as a POE. A TRUE NEGATIVE (TN) is a paper that is correctly classified as a non-POE. SENSITIVITY, also known as RECALL, is $TP/(TP+FN)$, so a type II error primarily lowers sensitivity. SPECIFICITY is the ratio $TN/(TN+FP)$, so a type I error lowers specificity. Physicians pay particular attention to sensitivity and specificity, because they do not depend on the prevalence of disease in a given population.

A COST FUNCTION determines the balance of the cost between FP and FN, hence the tradeoff between sensitivity and specificity. Higher cost ratios mean that sensitivity is preferred over specificity, while lower cost ratios mean that specificity is preferred over sensitivity. WEKA implements this preference scheme when an appropriate cost matrix file is supplied to a `CostSensitiveClassifier` class.

The True Positive Rate is $TP/(TP+FN)$, and the False Positive Rate is $FP/(FP+TN)$. Precision, surprisingly, is not the same as specificity. PRECISION is $TP/(TP+FP)$, which is the same as POSITIVE PREDICTIVE VALUE. Authors often report the performance of a classification scheme in terms of ACCURACY, which is simply

$$(TP+TN)/(TP+FP+TN+FN).$$

Another approach to measuring performance is based on likelihood ratios. The POSITIVE LIKELIHOOD RATIO (PLR) is $sensitivity/(1.00-specificity)$. The NEGATIVE LIKELIHOOD RATIO (NLR) is $(1.00-sensitivity)/specificity$. The PLR gives the number of times more likely it is that an abstract represents POE after the algorithm classifies it as such. The NLR gives the number of times more likely it is that an abstract represents non-POE after the algorithm so classifies it. The DIAGNOSTIC DOR RATIO (DOR) is defined to be (*positive likelihood ratio*)/(*negative likelihood ratio*), a measure of the information that a diagnostic test provides.

1.4 Related Work: Text Classification in Medical Abstracts

1.4.1 Determination of Structure

One type of related work involving text classification in medical abstracts is recreating the logical structure of an abstract that has no labels on its various parts. The target structure is the classic four part sequence: introduction/objective/purpose, methods/materials, results/findings, and conclusions/discussion. Increasingly, medical journals are forcing results into this format, a trend which began around the year 2000. Hirohata *et al.* (2008) achieved 95.5% per-sentence accuracy. Per-abstract accuracy is defined as the percentage of abstracts in which every sentence is put in the correct category. Their experiments yielded 68.8% per-abstract accuracy using CONDITIONAL RANDOM FIELDS (CRFs). The authors

created their own corpus of over 7 million medical abstracts, of which just about 102,000 had the prerequisite sections marked and were used in the analysis (51,000 strict and 51,000 more unstructured abstracts). CRFs are a sequencing assignment technique invented by Lafferty *et al.* (2001).

The numbers were so high in Hirohata *et al.* because they only needed to do two things. The first was to determine whether the introduction comes before the methods, or the methods come before the introduction. It is usually the former, but not always. The second was to determine the three break points between the 4 sections of the abstract. There are earlier schemes involving 5 different sections of the abstract, and using Hidden Markov Models, but they will not be discussed here due to the superiority of the CRF method.

There is another, more specifically medical, paradigm of abstract organization for PubMed abstracts; the Problem/Population, Intervention, Comparison, Outcome (PICO) view. Boudin *et al.* (2010a, 2010b) wrote two papers about PICO detection in abstracts by means of SVM-based machine learning. This was done to support evidence based medicine, which requires a careful search of the clinical literature. One rarely finds PICO elements explicitly annotated, but Boudin and colleagues explicitly annotated hundreds of abstracts for one study (2010a), and then used abstracts where PICO was specifically annotated for another (2010b). They were very conservative in the latter study, intent on not dealing with overly noisy information. An increase in P@10 of over 70% was reported. P@10 is defined to be the number of results among the top 10 that are relevant to the query over many runs; P@10 is used in some papers, but it is considered nonstandard even in the information retrieval literature.

1.4.2 Biomedical Sentiment Analysis

Opinion mining of biomedical abstracts was pioneered by Niu *et al.* (2006). The authors defined two tasks. The first was to take a sentence and its position in the abstract and classify the sentence as an **outcome** or **not an outcome**. An outcome sentence is a

sentence that gives the overall results of a medical RCT. Niu *et al.* (2006) achieved 82.5% accuracy for this task using SVMs, improving on a baseline of 65.9% for selection based solely on position in the abstract. The baseline in Niu *et al.* was the same for both tasks, 65.9%. Either this is an unlikely coincidence, or an error in the paper's tables; I do not know which. The second task was four-way classification of sentiment outcome in the abstract; positive, negative, neutral, or no outcome were the four possibilities. The outcome is positive if the results of the research are good overall. Negative results occur if the sentence gives a result that is not what was wanted. Neutral means there was a result, but the result sentence may be mixed, or the algorithm simply could not determine if the outcome was significant. The accuracy was 78.3%, again improving on a baseline of 65.9% provided by regular expression-based determination and position in the abstract.

Sarker *et al.* (2011a) continued the supervised sentiment analysis of biomedical abstracts. Special emphasis was placed on detecting multiple negations. Also, conditions from the medical literature are labeled as **good** or **bad**, and the algorithm determines whether more of a good thing is happening, or whether less of a bad thing is happening. Either of these are positive outcomes. For example, raising the I.Q. of babies is more of a good thing, while lowering mortality is less of a bad thing. The outcome of the polarity analysis is either positive, negative, or neutral, much as in Sarker *et al.*'s other paper from last year (2011b). Sarker *et al.* (2011a) do not attempt to determine whether there is an outcome sentence. The maximum accuracy obtained is 74.9%, but the domain is less restricted than the original 2006 work, the corpus being much larger and more varied. Just because one supervised machine learning paper reports a lower percentage than another does not mean the work is necessarily inferior. The choice of data can significantly affect the difficulty of a given task.

1.4.3 Semantic Work on PubMed Abstracts

The MEDLINE database is the database of medical abstracts behind the PubMed front end on the web, and there is semantic work on MEDLINE abstracts (Abacha and Zweigenbaum, 2010). These authors take a corpus of sentences from MEDLINE abstracts that claim a certain disease is cured by a certain drug class or specific drug. They also determine whether the sentence states that a drug or drug class has a certain condition as a side effect, or if the drug prevents a certain disease. They use the UMLS (Unified Medical Labeling System) to act as a thesaurus, and to identify drug names, which itself is no easy task. There are two approaches used at this point. One is a hand tuned regular expression system to find the verbs in the relations. The other is an SVM-based system, which worked better for the "cure" relationships. The other two relationships ("side effect" and "prevent") did better with the hand-tuned system due to small data sets (Abacha and Zweigenbaum, 2011). The authors combined these two systems to obtain 75.72% precision and 60.46% recall in the extraction of relationships between medications and diseases. The system improved the baseline precision, as determined in a previous paper using naïve bayes on the same data for different purposes, by 19.59%. The baseline precision from the previous paper was around 55.0%.

Abacha and Zweigenbaum enhanced MetaMap for these tests (Aronson, 2001). MetaMap is an existing noun phrase mapping program from the National Institute of Health that achieves over 83% recall and 85% precision, compared to hand mapping of noun phrases into medical concepts. There is also a MetaMap Plus program that extends MetaMap's functionality.[4]

Hansen *et al.* devised a method of automatically determining the number of trial participants from abstracts describing Randomized Controlled Trials with the help of SVMs (2008). They achieved an accuracy of 97.0% and an F1 Score of 0.84. The F1 Score is defined to be the harmonic mean of precision and recall.

[4]More information on MetaMap Plus can be found at www.metamap.com/files/User_Guide.pdf

The ExACT system (Kiritchenko et al., 2010) attempts to automatically find out many things about a study through shallow machine learning and statistical text mining. They want to determine, for instance, the amount by which a drug helped or hindered patients' health directly from the PubMed abstract, or the number of participants in a study, or the participation rate of a study. ExACT was broken down into two parts. The first part found relevant sentences, and the second part extracted 21 entities (details), some but not necessarily all of which are present in any given abstract. ExACT was evaluated using 50 previously unseen abstracts, and 21 details on each one, for a total of 1,050 data items. Most data items were present. ExACT showed five candidate sentences to the user for each of the 1,050 pairs. The system was able to recover 88% of the relevant sentences. Precision was 93% and recall was 91% for the second stage - extracting the relevant details from the relevant sentences. Six hundred ninety-six of the 1,050 data items represented fully correct and complete answers from the system in the very first sentence presented to the system's users.

1.4.4 Corpora for Biomedical Text Mining

Aggarwal and Zhai (2012) contains an overview of biomedical text mining in chapter 14. The authors discuss various biomedical corpora, including the historic and widely used OHSUMED and the highly annotated GENIA MEDLINE corpus. The Collaborative Annotation of a Large Biomedical Corpus (CALBC) aims to create a "silver standard" (one step below a near perfect "gold standard") corpus which has automatically annotated MEDLINE abstracts. Biomedical entities have been added by initiative participants, and this new, large corpus is publicly available. The Colorado Richly Annotated Full Text Corpus is semantically and syntactically annotated, and is designed to be like a "Penn Treebank for biomedicine." The Penn Treebank is a large collection of parsed and tagged sentences, often used to test part of speech taggers and parsers in the natural language processing arena (Marcus et al., 1993).

1.4.5 Non-Medical Prescription Drug Abuse

Daniulaityte *et al.* (2012) studied user opinions of the extra-medical use of three drugs. Sentiment Analysis and Web 2.0 technologies were used to find the drug users. The authors consider their work to be an important advance in drug use epidemiology. This study is another example of how sentiment analysis can be combined with text classification for data in the medical domain.

1.4.6 Automatically Determining Study Design

I am aware of only one prior study on automatically determining study design in the medical literature. Lin *et al.* (2011) developed a computational method for classifying the strength of evidence of cardiology papers. There were 3 levels of evidence for each of 3180 cardiac papers. Level A is the highest and level C the lowest, with Level B in between. Level A was assigned to 1108 papers, level B to 1705, and level C to 367. Finding POEs amounts to finding level A evidence. There are also levels for meta-analysis and broader questions only, with 1 being the highest and 3 the lowest level of evidence consistency and validity. Levels 1 and A are equivalent, but 1 is a level for meta-analyses or groups of studies, while A is a level for an individual research study. Level A evidence has highest priority when determining an evidence level of 1 for a disputed question of the body of medical literature. The authors report over 85% accuracy at determining the strength of evidence, A, B, or C, for the members of a test set of 900 abstracts (300 of each level). They also achieved a K-value (representing Cohen's Kappa) of 0.78. A K-value of 0.40 represents moderate agreement, a K-value of 0.60 represents good agreement, and a K-value of 0.80 represents nearly perfect agreement. So, a K-value of 0.78 is considered quite good.

As in this thesis, Lin *et al.* (2011) used cross validation. However, they employed a C4.5 decision tree algorithm, which is simpler than the naïve bayes multinomial, naïve bayes, support vector machine, and logistic regression methods of this thesis. My situation differs

from theirs in only having two classes of evidence, but also in requiring patient-oriented outcomes.

1.4.7 Identifying Speculative Language and Hedging in Biomedical Abstracts

The amount of work done each year in the area of identifying negation, hedging, and uncertainty in PubMed abstracts is rapidly growing. Hedging, speculation, and uncertainty are not all the same thing, but many of the studies on speculation detection in the medical literature treat them as the same, as will I. The exact definitions of these terms are a matter of some debate. Speculation detection through hand-written rules or machine learning is now a major research area within NLP. The usual classification algorithms for automatic speculation identification are SVMs or CRFs. About 14% of sentences in PubMed abstracts contain speculation, while about 20% of the sentences in the complete articles on PubMed Central contain speculation. Finding hedging, uncertainty, and speculation will all be called speculation identification for the remainder of this discussion.

In subsection 2.4.2, the benefits of negation detection for finding POEs are discussed. Speculation detection can help find open problems, but speculative language can also be a problem for POE identification. For instance, a conclusion sentence might say "drug X may lower mortality in condition Y," even though mortality itself was not measured in the study; then that abstract is more likely than it should be to be labeled a POE because of the speculative conclusion. The last two sentences of PubMed abstracts tend to contain speculation and uncertainty.

Search engines could also benefit from speculation identification. One might want to identify open problems or new areas to study. To find things that scientists suspect to be true but do not know for sure, one can look for speculative sentences. In order to find truths for knowledge bases, one needs to look for facts. One could pair speculative statements from

an earlier time with factual statements from a later time to help in automated scientific discovery through literature mining.

Friedman *et al.* (1994), for instance, found speculative language in clinical radiology reports with hand-crafted rules, assigning one of five levels of certainty. They defined the scope of speculation as narrowly as possible. Szarvas *et al.* (2008) later took a different approach, building a corpus for identifying uncertainty and negation, and their scopes, in PubMed abstracts. The authors tagged specific words reflecting negation and speculative language by hand using the same sort of syntax as HTML tags. They defined the scope of the speculative words and phrases as broadly as possible. This broad choice when the scope of uncertainty is unknown contrasts with previous work, where the scope of uncertainty was defined as narrowly as possible. The corpus they created is freely available on the Internet for other researchers to use. The idea was to create a standard corpus, so that the various approaches to the problem of speculative language detection could be more systematically studied by researchers.

Light *et al.* (2004) claim that there are not just facts and speculations in the biomedical literature. There are also statements of high and low speculation, and they note that one clause in a sentence may be factual while another may be speculative. Thus, the authors break the sentences into phrases with a chunk parser, and classify each phrase or clause as being low speculative, high speculative, or factual. They compared the annotations of two human experts, and concluded that humans can not distinguish high speculative and low speculative language consistently. However, the human experts were able to identify speculative and non-speculative language and their scopes reliably. Humans achieved 68% precision and 78% recall on two-way classification of clauses. Light *et al.* attained 71% precision and 39% recall using SVMs for the same two-way classification task. The SVMs were never applied to the three-way classification task.

Here are two typical examples of speculation from Light *et al.* (2004). "Pdcd4 may thus constitute a useful molecular target for cancer prevention"(pmid 1131400). Here, cancer

is not actually being prevented; there is just a possibility that cancer is being prevented. "Curcumin down-regulates Ki67, PCNA and mutant p53 mRNAs in breast cancer cells, these properties may underlie chemopreventive action" (pmid 14532610). This one is less medical. In this sentence, there may be a gene-protein link, but the matter is far from settled. Light *et al.* have about 20 examples of speculative and factual language from the literature. One could find thousands more, once one knows what to look for.

Chapter 2

The Problem

2.1 Testbed Data

In the course of a previous project Ebell *et al.* (2004) searched PubMed abstracts for POEMs from 1998–2008 in 161 major medical journals. Difficult cases were decided by examining the full texts of the articles. In all just over 3,000 POEMs were identified. In the current search for POEs, articles from the same set of 161 journals were examined. The original data set was taken from October, 2010, through October, 2011. There was a deemphasis on the last few months of 2011 due to slow PubMed updating, limiting availability of some abstracts. Dr. Ebell taught me to classify the PubMed abstracts as POE or non-POE by hand. There is an original dataset and a revised PubMed dataset, based on careful re-examination of some abstracts, as explained in subsection 2.4.2. Both Dr. Ebell and I classified a test set of 200 abstracts, with 95% agreement. I went on to classify 1160 more abstracts with occasional help from Dr. Ebell, for a total of 1356. These constitute our set of TESTBED ABSTRACTS. Of these, 713 were originally classified as POE and 644 as non-POE.

I then converted the testbed abstracts to TESTBED DATA in a format suitable for import into WEKA (Hall *et al.*, 2009). First, PubMed abstracts were downloaded, and everything between tags other than the title and abstract of the paper was deleted. The tags were

also removed, but each PubMed id was preserved for later reference. The HTML codes for double quotes and extra space (`&quot;` and ` `) were also suppressed. The abstracts were written one to a line, with a serial number first, followed by classification, and the journal impact factor. Each line was completed with the entire title and abstract between double quotes, and separated by a space.

A SEPARATOR character was taken to be any numeral, whitespace, or punctuation mark. A WORD, or UNIGRAM, consists of two or more consecutive letters in a TITLE/ABSTRACT string, bounded at each end by a separator. A BIGRAM, for implementation purposes, is two consecutive words, with an x placed between them. A TRIGRAM is three words in a row, with the separators between them replaced by two x's.

The token `beginningtoken` was placed at the beginning of the abstract once for bigrams, and twice for trigrams. This practice captures a paper's first words, and makes special bigrams and trigrams from them. Similarly, the token `endingtoken` was placed at the end of the text the same number of times. A TOKEN consists of a unigram, bigram, trigram, beginningtoken, or endingtoken. All tokens are lowercased after identification. A FEATURE is any token, or the impact factor of the journal the abstract came from.

Initially the token set consists of all unigrams appearing at least once in the testbed data. If bigrams are included in an experiment, then all bigrams derived from the testbed data are added to the initial token set. Similarly, if trigrams are to be included, then all bigrams and trigrams derived from the testbed data are added to the initial token set.

Finally, the testbed abstracts were converted to data files in WEKA 3.7.5's sparse matrix format, in which the features are listed in order, and feature numbers with values are given. In this study, the value is always the number of occurrences for a given word or n-gram in the abstract. Note that naïve bayes always interprets the data as binary. In addition the journal impact factor, a floating point number, is always included in the feature set.

2.2 Feature Selection for Text Classification

A common problem for text classification data, shared by our testbed initial feature set, is that the feature set is too large and needs to be reduced. Selection of features is accomplished by removing those features that fail certain threshold tests. In this work, document frequency and information gain thresholds are used; n–grams below the thresholds are culled from the data. The Document Frequency Minimum (DFM) is the minimum number of documents that an n–gram must appear in to be selected as a feature in the experiments. The Information Gain Minimum (IGM) is the minimum value for the information gain of an n–gram for that n–gram to be included in the experiments. Information gain itself is defined below, while the 50 n–grams with highest information gain are provided in Appendix A.

Information Gain (IG) is a measure of the benefit, for classification purposes, of including a term in the feature set. It is derived from the difference between entropies with and without the terms. The paper of Yang and Pedersen (1997) is seminal in the area of text classification and feature selection. They define the information gain of a term t to be:

$$IG(t) = -\sum_{i=1}^{m} P_r(c_i) \log P_r(c_i) + P_r(t) \sum_{i=1}^{m} P_r(c_i|t) \log P_r(c_i|t) + P_r(\bar{t}) \sum_{i=1}^{m} P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}) \tag{2.1}$$

The above information gain equation gives the change in entropy that would occur if a term was used by itself. It's always between 0.0 and 1.0 in the two-class case. Here, m is the number of categories in the target space, and $\{c_i\}_{i=1}^{m}$ is the set of classes. In our case $m = 2$, and one can take $c_1 =$ POE and $c_2 =$ non-POE.

(Rogati and Yang, 2002) found that Chi-squared feature selection, combined with an Information Gain cutoff or a document frequency cutoff, performed best when combined with either naïve bayes, Rocchio, k-nearest neighbor, or SVM classifier. J. J. Rocchio wrote a simple text classification method, the first based on continuous user feedback (1971). Rogati and Yang's results came from experiments with Reuters-21578 and a small part

of RCV-1, standard benchmarks for classic text classification research (Rogati and Yang, 2002). Rogati and Yang broke down over 100 feature selection methods and found that they preferred χ^2 in combination with certain cutoffs. They noted that cross-correlation could be used, but it generally performed only slightly better than other methods, at massive computational expense.

For a term t and a class c, the χ^2 statistic can be calculated as follows. Let A be the number of times t and c occur together, B the number of times t occurs without c, C the number of times c occurs without t, and D the number of times neither occurs. Also, let N be the total number of documents. Then, according to Yang and Pedersen's seminal paper on text classification and feature selection (1997):

$$\chi^2(t,c) = \frac{N(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{2.2}$$

The χ^2 statistic is not accurate for relatively rare words; it tends to overestimate their performance (Dunning, 1993). The χ^2 distribution is highly correlated with IG (Forman, 2003). Hence in this study information gain and document frequency cutoffs are used.

2.3 Machine Learning Methods Employed in Experiments

There are five machine learning techniques used in this work; two of them differ only slightly in pseudocode. The first and simplest is Naïve Bayes (NB), which is based on assuming that the feature variables conditioned on class membership are independent. The second is Naïve Bayes multinomial (NBM), a variant and generalization of naïve bayes, which is still quite basic. In naïve bayes the prior probabilities of each class are taken. Then the probably that each word or pair of words appears in a given class is computed, and the logarithms of these probabilities are added, in effect multiplying the probabilities together.

The most probable class is chosen. The idea behind naïve bayes classification can be traced back to a book review of *Minds, Machines, and Mathematics* (Good, 1968). In this review, naïve bayes is presented in the context of independent processes or machines with results being combined. Naïve bayes is called "naïve" because it assumes that each feature in the feature space is independent of all other features. This is in contrast to the more complicated bayesian belief networks, which assume something is known about the structure of causality in the problem.

The third type of machine learning employed in this study is SUPPORT VECTOR MACHINES (SVMs), which were invented by Vapnik in the late 1970's. Support vector machines did not become popular until the mid to late 1990's. Vapnik (2000) wrote the definitive book on SVMs, putting the algorithms into polished form. Platt (1998; 1999) went on to invent an optimization for SVM classification on sparse data. Training an SVM entails solving a large quadratic programming problem. Platt's method, SEQUENTIAL MINIMAL OPTIMIZATION (SMO), breaks the large quadratic programming problem into a series of smallest possible problems. In this way, large speedups for sparse data sets can be achieved. This is useful for us because our data is very sparse. SMOs were used in the experiments in this work. The usual way of referring to SMOs and SVMs is "SVM," a standard followed throughout this thesis. The speedup was dramatic. Runs that would have taken hours were reduced to a couple of minutes each.

The fourth type of machine learning used in this study is LOGISTIC REGRESSION (LR), also known as MAXIMUM ENTROPY (ME). This classification algorithm attempts to find the least constrained distribution based on the training data. This constraint that is minimized is the term to which maximum entropy refers. Berger *et al.* (1996) was one of the first papers to use LR in natural language processing.

The fifth type of machine learning employed in this study is the RANDOM FOREST (RF). RFs are examples of averaging methods in machine learning. There are no RF tables in this thesis due to the algorithm's low performance. An RF is a set of complete decision trees

populated by random variables. There is a maximum depth, and the tree is always filled to the maximum. The trees should not be correlated with each other very much. There may be, say, 100 trees and 20 features in each tree. Then, 100 random vectors are chosen, and the trees are aligned with the vectors. In this way a random weight is assigned to each tree. RFs work well when there are hundreds or even thousands of features, no one or small group of which can be used for prediction of the class. RFs are highly parallelizable; they excel in situations involving 1000s of training examples and 1000s of input variables together with powerful, multi-core computers. Breiman *et al.* (2001) introduced random forests, but there are earlier examples of similar methods, for instance Adaboost (Freund and Schapire, 1996), which Breiman considers inferior to his own creations. RFs are rapidly gaining in popularity in a number of fields, but still have not overtaken artificial neural networks in widespread usage.

2.4 Experiments

Ten-fold cross-validation is used for the experiments in this thesis. The data are split into 10 equal parts, and each of the 10 disjoint parts is used as a test set once and as part of a training set the other 9 times. As noted in the previous section, the learning algorithms applied in this study are NB, NBM, SVM, and LR, and RFs. For each of these, one fold of a cross-validation experiment took just a minute or two to complete. ANNs (Artificial Neural Networks) took several hours per fold to complete, so even though they are implemented in WEKA 3.7.5, they were not used in this study due to slowness of implementation. We also do not use GAs (Genetic Algorithms) to optimize the parameters of our system because they are so time consuming.

2.4.1 Experimental Procedure

A VALIDATION SETUP consists of a LEARNING SETUP (learning algorithm + parameter settings), and a FOLD GROUPING (random partition into 10 fold groups of 71 POEs and 64 non-POEs, and one leftover group of 3 POEs and 4 non-POEs). A FOLD consists of one experiment in which a fold group is treated as the test set and the rest of the testbed abstracts form the training set. The algorithm (with any parameters) specified in the validation setup is run in training mode, then in test mode. A VALIDATION EXPERIMENT is 10 folds, one for each fold group treated as the test set. A VALIDATION RUN is a set number k of validation experiments based on a common learning setup; the number k of experiments in the run is called the repetition number. The experiments in a run differ only in the fold groups of the testbed abstracts. These groupings should ideally be formed independently and at random. In actuality, all choices which should be random are only pseudo-random, as they rely on random number generators implemented in Python, Java, and WEKA.

Note that feature selection based on the DFM and IGM values for the learning setup must be carried out for every fold. However, the impact factor for each journal is constant for all folds. It was always included in the feature set. Journal impact factors averaged 10.12 over all POEs, and 7.45 over all non-POEs. The journal impact factor had an information gain of 0.1611 when the average of the POE and non-POE averages was used for a cutoff. The information gain for the journal impact factor was higher than that of any single n-gram ($1 \leq n \leq 4$) by more than a factor of 3. The 50 n-grams with the highest IG values are listed in appendix A.

WEKA 3.7.5 generated *confusion matrices* for each fold of the cross validation, and then these 10 were added together to form one confusion matrix. Each of these is a 2×2 matrix X with the number (TP) of *true POEs* in X_{11}, the number (FN) of *false non-POEs* in X_{12}, the number (FP) of *false POEs* in X_{21}, and the number (TN) of *true non-POEs* in X_{22}. For each learning setup a validation run with 10 repetitions was performed. The measures of classification accuracy are defined in section 1.3. Sensitivity, specificity, accuracy, and

diagnostic odds ratio are computed for each run. For each set of runs each accuracy measure was averaged and a range for it computed at 95% confidence, assuming a normal distribution and based on the standard unbiased estimator for its variance. In general, the runs in each set varied only a little bit in terms of specificity, sensitivity, and accuracy.

Runs were conducted for Support Vector Machines (SVM), Naïve Bayes (NB), and Naïve Bayes Multinomial (NBM), and Logistic Regression (LR). The complexity parameter of the SVM and its kernel were systematically varied.

2.4.2 Analysis of Commonly Misclassified Abstracts

The best results on the original dataset were with naïve bayes multinomial on N=2, IGM=0.001, DFM=4, 5, or 6, and Cost=1:1. The 100 most commonly misclassified abstracts (50 false negatives and 50 false positives) averaged over these three learning setups were studied. To locate articles misclassified in the original dataset, these 100 abstracts were precisely reclassified, taking more time to look at each one. This reclassification found 8 positives that should be negatives, and 8 negatives that should be positives. The classification by hand was about 95% accurate overall, but 16% of the commonly misclassified abstracts had been misclassified by me the first time around. The REVISED DATASET was prepared by correcting the 16 misclassifications found, and formed the basis of all subsequent experimental results.

A number of comparisons were run with the same learning setup between the original and revised datasets. In every case the results from the revised dataset were slightly better than those from the original dataset. For instance, the highest accuracies were attained by the learning setup noted above with DFM=4; these were 79.5% on the original dataset and 80.4% on the revised dataset. For logistic regression Table 2.11 compares results from the original and revised datasets.

Several features appear in abstracts which are commonly misclassified:

(i) a patient-oriented word, but the outcome is a secondary outcome or a safety outcome;

(ii) an implied patient-oriented outcome, such as "clinical outcomes;"

(iii) mixing of patient-oriented outcomes such as mortality with non-patient-oriented outcomes such as height and weight in the same study;

(iv) diseases that are usually measured with surrogates, like AIDS, diabetes, and asthma, but the study actually is patient-oriented;

(v) rarely used phrases in the corpus that are too long for the classification algorithm, like "end stage renal disease;" That's four words, but usually the classification algorithm does not look at 4–grams. Most "renal disease" studies use surrogate markers, but "end stage renal disease" is a dire condition.

Odd study designs also contribute to the difficulty of classifying abstracts. For instance, there are secondary analyses of other studies in the literature, or, worse, studies that describe the design of a trial that has not yet been conducted.

Also, some studies use negation in a way that would seem to require parsing to be interpreted correctly in the classification process. Take for example "neither patients nor doctors were blinded in this study." This statement would cause the classification algorithm to behave as if this were a double blind trial, when in fact it is not even a single blind trial.

Measures that would help some of these cases to be handled correctly in classification include domain-specific semantic analysis of the data, full parsing, and identification of negative and speculative language. Some of these measures have been attempted in other studies, as discussed in subsection 1.4.7.

2.5 Experimental Results

In this section some experimental results are reported which are based on the original testbed data, as well as those based on the revised data. The revision procedure is described in subsection 2.4.2. Since many of the experiments are quite time consuming, only the most important results were rerun on the revised data. In general the results on the revised data followed the same pattern as those for the original data, but with slightly better values for the various accuracy measures.

The first subsection gives an overall comparison of the best results from NBM, LR, SVMs, and NB. Then for each of these algorithms, a subsection is devoted to presenting the results obtained.

Random forests did the worst of all the algorithms; at best, the accuracy was 58%. No tables of results for random forests are included in this thesis. Only the WEKA implementation of random forests was tried, and there was not time to exhaustively check parameters.

Ten runs with 10-fold cross-validation were used to obtain all experimental results in this section. Unless otherwise noted, the IGM for all results in this section is 0.0010, a value that has been found to work well.

2.5.1 Overall Results

From Table 2.1, it is apparent that naïve bayes multinomial has the highest accuracy, and it is apparent from the 95% confidence intervals that the result is statistically significant. The N in the table represents whether 2–grams or 3–grams did best. DFM is the document frequency minimum, the minimum number of documents an n–gram must appear in to be selected in the feature set. Cost is the cost ratio supplied to WEKA. Positive predictive value (PPV) is the probability that something the algorithm selects as a POE is one; the formula, given in section 1.3, is the same as for precision in information retrieval. The table

Algorithm	N	DFM	Cost	Sens	Spec	PPV	Accuracy
Naïve Bayes Multinomial	2	6	1:1	0.826±0.005	0.779±0.004	0.805	0.804±0.003
Logistic Regression	2	4	1:1	0.784±0.007	0.735±0.007	0.766	0.760±0.003
Naïve Bayes -K	3	6	1:1	0.794±0.006	0.701±0.009	0.746	0.750±0.006
SVM	2	4	1:1	0.749±0.002	0.730±0.003	0.754	0.740±0.002

Algorithm	N	DFM	Cost	PLR	NLR	DOR
Naïve Bayes Multinomial	2	6	1:1	3.74	0.223	16.764±0.576
Logistic Regression	2	4	1:1	2.96	0.294	10.083±0.392
Naïve Bayes -K	3	6	1:1	2.66	0.294	9.104±0.576
SVM	2	4	1:1	2.77	0.344	8.135±0.175

Table 2.1: Summary of best overall accuracy on the revised data for each algorithm with 95% confidence intervals for sensitivity, specificity, accuracy, and diagnostic odds ratio

Algorithm	Cost	Accuracy
Naïve Bayes Multinomial	1:15	0.804±0.002
Naïve Bayes Multinomial	1:1	0.804±0.003
Naïve Bayes Multinomial	15:1	0.803±0.003

Table 2.2: Best overall accuracy on revised data for NBM — 2–grams — DFM 6 — Cost set at 1:15, 1:1, and 15:1.

is sorted by performance from highest to lowest. A distant second is logistic regression, followed by naïve bayes in third place. SVMs bring up the bottom of the stack, which is surprising, considering their widespread popularity. SVMs use the polynomial kernel in this case; WEKA's other possibility is the gaussian kernel. The kernel of an SVM is an embedded distance algorithm that an SVM system can swap out, but that is crucial for its functioning. The kernel does not matter much in text classification, but it can be crucial with non-textual data. The naïve bayes family did best with a DFM of 6, while the other algorithms performed better with a DFM of 4. Bigrams were dominant in most of the experiments, but naïve bayes did better with trigrams. The best accuracy achievable in any experiment was roughly 80.4%. This is up from 79.5% with the original data. Hence, the revised dataset improves the maximum accuracy by about 0.9%.

N	DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
1	4	1:15	0.800±0.004	0.797±0.003	0.798±0.003	15.675±0.613
2	4	1:15	0.807±0.004	0.790±0.004	0.799±0.003	15.779±0.729
3	4	1:15	0.802±0.004	0.795±0.005	0.798±0.003	15.702±0.700
1	5	1:15	0.804±0.005	0.798±0.002	0.801±0.003	16.166±0.579
2	5	1:15	0.812±0.004	0.791±0.003	0.802±0.004	16.438±0.745
3	5	1:15	0.807±0.005	0.793±0.002	0.801±0.003	16.103±0.630
1	6	1:15	0.802±0.004	0.803±0.004	0.802±0.004	16.578±0.803
2	6	1:15	0.806±0.004	0.801±0.002	0.804±0.002	16.771±0.531
3	6	1:15	0.796±0.004	0.806±0.002	0.801±0.002	16.208±0.447

Table 2.3: Naïve Bayes Multinomial — revised data — Cost 1:15 — DFM and N vary

N	DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
1	4	1:1	0.824±0.004	0.765±0.005	0.796±0.003	15.316±0.574
2	4	1:1	0.825±0.004	0.772±0.003	0.800±0.003	15.948±0.555
3	4	1:1	0.819±0.004	0.778±0.003	0.799±0.003	15.857±0.588
1	5	1:1	0.825±0.005	0.767±0.004	0.798±0.003	15.634±0.566
2	5	1:1	0.831±0.004	0.771±0.002	0.802±0.003	16.577±0.601
3	5	1:1	0.822±0.004	0.775±0.003	0.800±0.003	15.956±0.612
1	6	1:1	0.823±0.005	0.777±0.004	0.801±0.004	16.287±0.710
2	6	1:1	0.826±0.005	0.779±0.004	0.804±0.003	16.764±0.576
3	6	1:1	0.814±0.004	0.786±0.003	0.801±0.003	16.096±0.592

Table 2.4: Naïve Bayes Multinomial — revised data — Cost 1:1 — DFM and N vary

Table 2.2 shows the highest accuracies found over all learning setups when the cost ratio is set to 1:15, 1:1, and 15:1. These were all produced by NBM with N=2 (bigrams) and DFM=6. The accuracies are indistinguishable given the 95% confidence intervals. In the next subsection there are tables showing sensitivity and specificity. For those performance measures the effect of changing the cost ratio is highly significant.

2.5.2 Results with Naïve Bayes Multinomial

The best overall result, reported on the first line of table 2.1, is expanded on in table 2.4. There the additional performance measures of sensitivity, specificity and diagnostic

N	DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
1	4	15:1	0.847±0.004	0.731±0.005	0.792±0.003	15.085±0.548
2	4	15:1	0.840±0.004	0.751±0.004	0.798±0.003	15.936±0.532
3	4	15:1	0.834±0.004	0.758±0.004	0.798±0.002	15.702±0.392
1	5	15:1	0.849±0.004	0.732±0.005	0.793±0.003	15.401±0.491
2	5	15:1	0.845±0.004	0.749±0.003	0.800±0.002	16.325±0.522
3	5	15:1	0.837±0.003	0.753±0.004	0.797±0.002	15.689±0.455
1	6	15:1	0.847±0.003	0.738±0.007	0.796±0.004	15.719±0.636
2	6	15:1	0.842±0.003	0.760±0.004	0.803±0.002	16.898±0.469
3	6	15:1	0.830±0.005	0.767±0.003	0.800±0.003	16.119±0.570

Table 2.5: Naïve Bayes Multinomial — revised data — Cost 15:1 — DFM and N vary

Cost	Sensitivity	Specificity	Accuracy	DOR
1:20	0.762±0.006	0.778±0.005	0.770±0.004	11.298±0.447
1:15	0.768±0.005	0.774±0.005	0.771±0.003	11.367±0.385
1:10	0.773±0.005	0.770±0.005	0.771±0.003	11.395±0.327
1:5	0.779±0.007	0.759±0.005	0.760±0.003	11.125±0.440
1:1	0.801±0.004	0.736±0.004	0.770±0.002	11.255±0.330
5:1	0.823±0.004	0.715±0.007	0.771±0.004	11.640±0.470
10:1	0.830±0.005	0.704±0.006	0.770±0.003	11.630±0.411
15:1	0.834±0.006	0.698±0.006	0.769±0.003	11.649±0.452
20:1	0.836±0.005	0.693±0.007	0.769±0.004	11.605±0.464

Table 2.6: Naïve Bayes Multinomial — original data — 3–grams — DFM 6 — cost varies.

odds ratio are reported. Also the effect of ranging N over 1, 2, 3 and DFM over 4, 5, 6 is displayed and seen to make only slight differences for NBM. Most of these differences are not statistically significant.

Tables 2.3, 2.4, and 2.5 differ in their parameters only in setting the cost ratio at 1:15, 1:1, and 15:1 respectively. Tables 2.3 and 2.5 exhibit the same near constant behavior as N and DFM are varied as does table 2.4. Significant differences are seen between the tables in that sensitivity increases and specificity decreases as the cost ratio increases. This variation offers the opportunity to trade one for the other, as needed by the particular application. For POEs doctors would probably prefer sensitivity over specificity, as provided by the 15:1 cost ratio.

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	1:15	0.842±0.003	0.760±0.003	0.803±0.002	16.898±0.413
0.002	1:15	0.850±0.003	0.744±0.003	0.800±0.001	16.527±0.302
0.003	1:15	0.838±0.007	0.742±0.008	0.792±0.002	14.933±0.313
0.004	1:15	0.836±0.007	0.730±0.010	0.785±0.003	13.835±0.476
0.005	1:15	0.825±0.009	0.726±0.013	0.778±0.003	12.606±0.392
0.006	1:15	0.824±0.009	0.728±0.011	0.779±0.002	12.656±0.314
0.007	1:15	0.821±0.010	0.720±0.013	0.773±0.003	11.935±0.403
0.008	1:15	0.816±0.012	0.721±0.015	0.771±0.002	11.603±0.368
0.009	1:15	0.807±0.013	0.721±0.016	0.766±0.003	10.951±0.445
0.010	1:15	0.801±0.017	0.720±0.018	0.762±0.002	10.520±0.439
0.011	1:15	0.786±0.020	0.718±0.020	0.754±0.003	9.625±0.470
0.012	1:15	0.778±0.023	0.715±0.025	0.748±0.002	9.095±0.340

Table 2.7: Naïve Bayes Multinomial — revised data — 2–grams — Cost 1:15 — DFM 6 — IGM varies

Table 2.6 shows the dependence of sensitivity and specificity on a range of cost ratios for NBM on the original data with 3–grams and DFM=6. The dependence is smooth over the whole range of cost ratios, from 1:20 to 20:1. The effect seems to flatten out a little at the higher ratios. The diagnostic odds ratio seems to be non-significant in many cases where the other measurements are more significant. The accuracy varies only a little bit, and the differences are mostly non-significant.

Tables 2.7, 2.8, and 2.9 show the effects of changing IGM for 2–grams for cost ratios of 1:15, 1:1, and 15:1 and DFM 6, respectively. All three show that accuracy decreases as IGM increases. Specificity is highest for the lowest values of IGM. This table contains the best accuracy values from the original data. As it stands, table 2.9 does not show the same trends for sensitivity and specificity as tables 2.7 and 2.8. It has flat sensitivity as IGM increases, versus falling for tables 2.7 and 2.8. Also, as IGM increases, specificity goes up (slowly) in table 2.7, is flat in 2.8, and up in 2.9. It is true that for fixed IGM, going from cost ratio 1:15 to 1:1 and 1:1 to 1:15 gives significant increases in sensitivity, significant decreases in specificity, and little change in accuracy.

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	1:1	0.806±0.003	0.801±0.002	0.804±0.002	16.771±0.468
0.002	1:1	0.812±0.003	0.797±0.004	0.805±0.002	16.914±0.356
0.003	1:1	0.845±0.007	0.733±0.015	0.792±0.004	15.122±0.655
0.004	1:1	0.835±0.010	0.724±0.021	0.782±0.005	13.551±0.689
0.005	1:1	0.832±0.012	0.716±0.021	0.777±0.004	12.754±0.506
0.006	1:1	0.828±0.015	0.721±0.022	0.777±0.003	12.714±0.408
0.007	1:1	0.821±0.021	0.713±0.027	0.770±0.003	11.843±0.325
0.008	1:1	0.815±0.027	0.711±0.030	0.766±0.003	11.386±0.418
0.009	1:1	0.811±0.032	0.706±0.032	0.761±0.003	10.946±0.350
0.010	1:1	0.804±0.037	0.702±0.036	0.756±0.004	10.453±0.490
0.011	1:1	0.799±0.043	0.695±0.040	0.749±0.004	9.913±0.411
0.012	1:1	0.790±0.049	0.696±0.043	0.745±0.006	9.622±0.409

Table 2.8: Naïve Bayes Multinomial — revised data — 2–grams — Cost 1:1 — DFM 6 — IGM varies

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	15:1	0.825±0.004	0.778±0.003	0.802±0.002	16.526±0.493
0.002	15:1	0.831±0.004	0.769±0.003	0.802±0.002	16.419±0.401
0.003	15:1	0.817±0.008	0.773±0.008	0.796±0.002	15.317±0.369
0.004	15:1	0.810±0.007	0.766±0.007	0.789±0.003	13.990±0.437
0.005	15:1	0.790±0.011	0.770±0.010	0.780±0.003	12.723±0.458
0.006	15:1	0.785±0.012	0.774±0.011	0.780±0.003	12.622±0.392
0.007	15:1	0.774±0.013	0.773±0.012	0.773±0.003	11.772±0.458
0.008	15:1	0.752±0.016	0.776±0.014	0.764±0.003	10.707±0.370
0.009	15:1	0.736±0.019	0.783±0.014	0.758±0.004	10.188±0.366
0.010	15:1	0.716±0.023	0.786±0.015	0.749±0.006	9.451±0.407
0.011	15:1	0.693±0.025	0.790±0.015	0.739±0.007	8.703±0.428
0.012	15:1	0.674±0.027	0.796±0.015	0.732±0.008	8.269±0.443

Table 2.9: Naïve Bayes Multinomial — revised data — 2–grams — Cost 15:1 — DFM 6 — IGM varies

Cost	Sensitivity	Specificity	Accuracy	DOR
1:20	0.804±0.003	0.804±0.003	0.804±0.002	16.803±0.462
1:15	0.806±0.004	0.801±0.002	0.804±0.002	16.771±0.531
1:10	0.809±0.004	0.798±0.003	0.804±0.003	16.840±0.629
1:5	0.814±0.004	0.793±0.003	0.804±0.003	16.766±0.584
1:1	0.826±0.005	0.780±0.003	0.804±0.002	16.815±0.555
1:5	0.814±0.004	0.793±0.003	0.804±0.003	16.766±0.584
5:1	0.835±0.004	0.766±0.003	0.802±0.002	16.526±0.488
10:1	0.840±0.004	0.762±0.003	0.803±0.002	16.862±0.487
15:1	0.842±0.003	0.760±0.004	0.803±0.002	16.898±0.469
20:1	0.844±0.004	0.758±0.004	0.803±0.002	17.003±0.549

Table 2.10: Naïve Bayes Multinomial — revised data — 2–grams — Cost Varies

D	N	Sensitivity	Specificity	Accuracy	DOR
O	1	0.727±0.008	0.701±0.009	0.715±0.005	6.286±0.331
O	2	0.737±0.005	0.725±0.008	0.731±0.005	7.407±0.363
O	3	0.741±0.007	0.732±0.005	0.737±0.004	7.869±0.320
R	1	0.753±0.008	0.734±0.007	0.744±0.006	8.509±0.545
R	2	0.688±0.003	0.831±0.007	0.756±0.003	10.947±0.537
R	3	0.857±0.004	0.625±0.006	0.747±0.003	10.048±0.406

Table 2.11: LR — Cost 1:1 — DFM 6 — original and revised data — N varies

Table 2.10 shows NBM with 2–grams on revised data, with varying cost. As the cost ratio increases, sensitivity increases, while specificity decreases. The diagnostic odds ratio is as good as for any experiment. Accuracy varies little from condition to condition in this table.

2.5.3 Results with Logistic Regression

In Table 2.11, it can be seen that both the accuracy and diagnostic odds ratios are higher in the revised dataset versus the original dataset. The variation is higher for sensitivity, specificity, and accuracy on the unigrams versus the bigrams and trigrams. The highest sensitivity in the table is on the revised dataset with trigrams, while the revised dataset with bigrams has the highest specificity in the table. The changes in accuracy between the original dataset and the revised dataset, as well as the changes in the diagnostic odds ratios, are statistically significant. The accuracy increases by a percent or more, as in the case of revised bigrams. These two datasets do not always exhibit the same patterns. In the original dataset N equals 3 is best for everything, while in the revised dataset, N equals 2 is best for accuracy and diagnostic odds ratios.

The revised dataset's numbers are given in more detail in tables 2.12, 2.13, and 2.14, which give the revised numbers for 1–3–grams. In Table 2.14, a DFM of 5 and a cost ratio of 1:1 give the highest sensitivity, while in table 2.13, a DFM of 4, again with cost ratio

DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
4	1:15	0.704±0.003	0.793±0.007	0.746±0.004	9.130±0.448
4	1:1	0.773±0.005	0.721±0.006	0.748±0.004	8.816±0.402
4	15:1	0.827±0.004	0.647±0.008	0.742±0.004	8.838±0.417
5	1:15	0.703±0.005	0.787±0.007	0.743±0.003	8.796±0.364
5	1:1	0.765±0.005	0.723±0.008	0.745±0.006	8.551±0.544
5	15:1	0.822±0.006	0.647±0.008	0.739±0.005	8.513±0.446
6	1:15	0.690±0.005	0.794±0.007	0.739±0.005	8.625±0.430
6	1:1	0.753±0.008	0.734±0.007	0.744±0.006	8.509±0.545
6	15:1	0.805±0.007	0.672±0.011	0.742±0.008	8.529±0.656

Table 2.12: LR — revised data — 1–grams — Cost and DFM vary

DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
4	1:15	0.774±0.007	0.743±0.009	0.759±0.006	9.986±0.597
4	1:1	0.784±0.007	0.735±0.007	0.760±0.003	10.083±0.392
4	15:1	0.779±0.007	0.725±0.007	0.753±0.003	9.301±0.307
5	1:15	0.758±0.007	0.749±0.006	0.754±0.005	9.383±0.471
5	1:1	0.776±0.009	0.735±0.008	0.757±0.005	9.696±0.568
5	15:1	0.764±0.008	0.752±0.006	0.758±0.005	9.860±0.573
6	1:15	0.668±0.005	0.843±0.004	0.751±0.003	10.833±0.429
6	1:1	0.688±0.003	0.831±0.007	0.756±0.003	10.947±0.537
6	15:1	0.686±0.008	0.819±0.009	0.749±0.004	9.982±0.534

Table 2.13: LR — revised data — 2–grams — Cost and DFM vary

1:1, gives the highest sensitivity. The highest diagnostic odds ratio in these three tables is in table 2.13, with a cost ratio of 1:1 and a DFM of 6. For 1–grams, the highest sensitivity comes with a cost ratio of 15:1 and a DFM of 4.

2.5.4 Results with Support Vector Machines

Table 2.15 shows our results for SVM on the revised data set. This table shows results on a Polynomial Kernel, with an exponent of 1.01, and 1–3 grams and a DFM of 4–6. The accuracy for most of the experimental conditions is between 73 and 74 percent. Sensitivities, except for trigrams with a DFM of 6, are higher than the specificities. That result is

DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
4	1:15	0.671±0.006	0.829±0.006	0.746±0.004	9.905±0.506
4	1:1	0.678±0.006	0.834±0.006	0.752±0.004	10.636±0.519
4	15:1	0.667±0.008	0.836±0.007	0.747±0.006	10.339±0.765
5	1:15	0.862±0.003	0.623±0.007	0.749±0.004	10.369±0.382
5	1:1	0.863±0.004	0.618±0.007	0.747±0.003	10.224±0.301
5	15:1	0.858±0.006	0.621±0.007	0.745±0.003	9.918±0.453
6	1:15	0.833±0.006	0.655±0.007	0.748±0.004	9.495±0.420
6	1:1	0.857±0.004	0.625±0.006	0.747±0.003	10.048±0.406
6	15:1	0.843±0.006	0.652±0.008	0.752±0.005	10.137±0.573

Table 2.14: LR — revised data — 3–grams — Cost and DFM vary

N	DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
1	4	1:1	0.753±0.002	0.720±0.003	0.738±0.002	7.908±0.151
1	5	1:1	0.755±0.002	0.712±0.002	0.734±0.002	7.631±0.124
1	6	1:1	0.745±0.002	0.728±0.003	0.737±0.002	7.894±0.139
2	4	1:1	0.749±0.002	0.730±0.003	0.740±0.002	8.135±0.175
2	5	1:1	0.746±0.002	0.727±0.002	0.737±0.001	7.860±0.110
2	6	1:1	0.735±0.002	0.733±0.002	0.734±0.002	7.660±0.140
3	4	1:1	0.736±0.002	0.733±0.002	0.735±0.001	7.689±0.113
3	5	1:1	0.741±0.002	0.727±0.002	0.734±0.001	7.645±0.097
3	6	1:1	0.727±0.003	0.739±0.002	0.733±0.002	7.585±0.147

Table 2.15: SVM Polynomial Kernel — Revised Data — Cost 1:1 — DFM and N vary

statistically significant. SVMs have enjoyed wide popularity in recent years, but they did not perform as well with the POE data set as NBM and LR.

2.5.5 Results with Naïve Bayes

Table 2.16 shows the results from the best three experimental conditions for the naïve bayes algorithm on the original dataset. Naïve bayes with bigrams has a sensitivity of 0.840 and a specificity of 0.620, for an accuracy of over 73% in two cases. It was the third most successful algorithm in the tests, followed by SVMs coming in last place.

In the tests discussed below, lower values of the IG cutoff achieved the highest accuracy. This fact was less true for the SVM algorithm than for other algorithms, where the IG cutoff of the three best is 0.05 rather than 0.00 or 0.025 for the others. Note that the minimum value of IG is 0.00, which would include $11,000$ or so features for unigrams, which is far too many.

Table 2.16 shows what happens with 3-grams with the $-K$ option turned on in WEKA. The $-K$ option represents distribution estimation, a process that complicates the original Naïve bayes models by modeling the means of many normals instead of assuming a single normal distribution; the $-D$ option represents classic Naïve bayes. Here, the minimum information gain remains fixed at 0.005, as it does for Table 2.17. If information gain is set reasonably, it does the work of DFM. If information gain is varied, the performance of the system varies more than if DFM is set to different values between 1 and 9. DFM was also varied between 1 and 19 for unigrams and bigrams, but the data reveals no new insights. Interestingly, the balance between sensitivity and specificity varies by option. The sensitivity is higher than the specificity with the $-K$ option, while the specificity is higher than the sensitivity with the $-D$ option.

In Table 2.16, accuracy never increases above 75.0%, at a DFM of 6. The DFM of 1, at 74.8% accuracy, is also quite competitive.

DFM	Sensitivity	Specificity	Accuracy	DOR
1	0.787±0.004	0.705±0.006	0.748±0.004	8.837±0.383
2	0.786±0.007	0.703±0.007	0.746±0.006	8.715±0.538
3	0.786±0.007	0.700±0.007	0.745±0.005	8.600±0.443
4	0.790±0.005	0.699±0.006	0.747±0.006	8.747±0.423
5	0.787±0.009	0.700±0.008	0.746±0.004	8.649±0.360
6	0.794±0.006	0.701±0.009	0.750±0.006	9.104±0.576
7	0.780±0.005	0.703±0.005	0.744±0.004	8.412±0.363
8	0.777±0.010	0.704±0.009	0.742±0.008	8.371±0.669
9	0.779±0.008	0.706±0.007	0.745±0.006	8.529±0.508

Table 2.16: NB — original data — 3–grams –K Option — DFM varies

DFM	Sensitivity	Specificity	Accuracy	DOR
1	0.711±0.004	0.746±0.005	0.728±0.003	7.231±0.250
2	0.709±0.005	0.748±0.006	0.728±0.004	7.253±0.273
3	0.715±0.007	0.744±0.003	0.729±0.004	7.301±0.283
4	0.715±0.007	0.744±0.005	0.729±0.005	7.305±0.357
5	0.714±0.006	0.744±0.006	0.729±0.004	7.299±0.312
6	0.717±0.006	0.750±0.006	0.732±0.004	7.585±0.351
7	0.716±0.007	0.749±0.006	0.732±0.003	7.526±0.201
8	0.712±0.007	0.743±0.006	0.726±0.005	7.162±0.326
9	0.712±0.007	0.750±0.004	0.730±0.005	7.434±0.341

Table 2.17: NB — original data — 3–grams –D Option — DFM varies

Table 2.17 shows naïve bayes with trigrams and the $-D$ option and an information gain of 0.001. There is not much variation in accuracy, sensitivity, or specificity in this table. The accuracy never rises above 73.2%.

The above Table 2.18 shows what happens when only the cost matrix varies in naïve bayes, with both the $-D$ and the $-K$ options. The highest sensitivity in the table is $-D$ with cost ratio 5:1, at 0.819, while the highest specificity is $-D$ with the same cost ratio, for a specificity of 0.821. This is all on the original dataset. The overall trend is that as the cost ratio increases, sensitivity increases. As the cost ratio decreases, specificity increases.

Opt	Cost	Sensitivity	Specificity	Accuracy	DOR
D	1:5	0.588±0.008	0.821±0.006	0.699±0.004	6.585±0.303
D	1:1	0.713±0.005	0.745±0.005	0.728±0.003	7.250±0.246
D	5:1	0.819±0.003	0.662±0.007	0.745±0.003	8.881±0.309
K	1:5	0.760±0.008	0.733±0.008	0.747±0.005	8.689±0.417
K	1:1	0.789±0.009	0.704±0.008	0.748±0.006	8.923±0.616
K	5:1	0.814±0.007	0.677±0.008	0.749±0.006	9.228±0.634

Table 2.18: NB — original data — 3–grams — $-D$ or $-K$ Option and Cost vary

Chapter 3

Conclusions and Future Work

The PubMed abstracts appearing in the selected set of 161 journals between October, 2010 and November, 2011 were divided into POE and non-POE, and a corpus was built. N–grams from the abstracts and the impact factor of each journal were extracted from the corpus. The algorithms NBM, LR, NB, SVM, and RF were tested on the corpus, and the results were averaged over several runs.

The experiments turned up several surprises; for instance, naïve bayes multinomial (NBM) was not expected to do well, nor were methods which employ frequency data. With suitable feature selection NBM achieved 80.4% accuracy in POE identification. This method had significantly higher accuracy than any of the other methods studied. NBM allowed for trading off between false negatives and false positives over a fairly wide range with no significant loss of accuracy. As expected, feature selection was critical to achieving the best results. The highest accuracy was attained by using bigrams as features, and including only those meeting the DFM=6 and IGM=0.0010 thresholds.

A minor objective for future work is to improve on POE classification. Detection of negatives is not too hard, and it should be a significant help. The main objective for future work should be a method for finding POEMs.

Future work should focus on identifying evidence that matters for medical practice, *i.e.*, evidence which satisfies criterion (iii) from the Introduction. Deeper linguistic techniques and/or full text analysis appear to be required for an acceptably accurate computational approach to recognizing criterion (iii) evidence. It is likely to be much harder to find POEMs than POEs.

First, full-parsing and a richer understanding of medical texts should be investigated. There are methods for deeply parsing medical language based on existing English parsers and novel dictionaries, and there are techniques for manually or automatically generating these parser dictionaries. Medical language parsing has its own literature. Parsing the medical literature (Lease and Charniak, 2005) and other deep linguistic methods such as chunk parsing might help significantly to identify POEMs.

PubMed Central contains the full text of some articles a year after they are published. The current study does not use PubMed central, but free, full text access is increasingly common in the medical literature, and it could well be exploited to help find POEMs.

Data from the thesis project will be made publicly available soon after the results are published in a research paper. Addition of POE detection capabilities to the Clinical Query search engine is planned for PubMed.

Bibliography

Abacha, Asma Ben and Pierre Zweigenbaum. 2010. Automatic Extraction of Semantic Relations Between Medical Entities: Application to the Treatment Relation. In *Proceedings of the Fourth International Symposium on Semantic Mining in Biomedicine (SMBM)*, pages 4–11.

Abacha, Asma Ben and Pierre Zweigenbaum. 2011. Automatic Extraction of Semantic Relations Between Medical Entities: A Rule Based Approach. *Journal of Biomedical Semantics*, 2.

Aggarwal, Charu C. and Cheng Xiang Zhai. 2012. *Mining Text Data*. Springer, New York.

Aronson, Alan R. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association*, pages 17–21.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22, 39–71.

Boudin, Florian, Jian-Yun Nie, Joan C. Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. 2010a. Combining Classifiers for Robust PICO Element Detection. *BMC Medical Informatics and Decision Making*, 10.

Boudin, Florian, Jian-Yun Nie, Martin Dawes. 2010b. Clinical Information Retrieval using Document and PICO Structure. In *Proceedings of the Human Language Technologies–North American Association of Computational Linguistics*, pages 822–830.

Breiman, Leo 2001. Random Forests. *Machine Learning*, 45, 5–32.

Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Empirical Methods in Natural Language Processing*, pages 1301–1309.

Daniulaityte, Raminta, Ressel Falck, Linna Li, Ramzi W. Nahhas, and Robert G. Carlson. 2012. Respondent-Driven Sampling to Recruit Young Adult Non-medical Users of Pharmaceutical Opioids: Problems and Solutions. *Drug and Alcohol Dependence*, 121, 23–29.

Du, Xin, Toshiharu Ninomiya, Bastiaan de Galan, Edward Abadir, John Chambers, Avinesh Pillai, Mark Woodward, Mark Cooper, Stephen Harrap, Pavel Hamet, Neil Poulter, Gregory Y. H. Lip, and Anushka Patel. 2009. Risks of Cardiovascular Vents and Effects of Routine Blood Pressure Lowering among Patients with Type 2 Diabetes and Atrial Fibrillation: Results from the ADVANCE Study. *European Heart Journal*, 30, 1128–1135.

Dunning, Ted E. 1993. Accurate Methods for the Statisics of Surprise and Coincidence. *Computational Linguistics*, 19, 61–74.

Ebell, M. H., J. Siwek, B. D. Weiss, S. H. Woolf, J. Susman, B. Ewigman, and M. Bowman. 2004. Strength of Recommendation Taxonomy (SORT): A Patient-Centered Approach to Grading Evidence in the Medical Literature. *Journal of the American Board of Family Medicine*, 17, 59–67.

Forman, George. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3, 1289–1305.

Freund, Yoav and Robert E. Schapire. 1996. Experiments with a New Boosting Algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning, pages 148–156.

Friedman, C., P. Alderson, J. Austin, J. Cimino, and S. Johnson. 1994. A General Natural-Language Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1, 161–174.

Good, Irving John. 1968. Brains, Machines, and Mathematics - Arbib, MA. *Computer Journal*, 8, 88.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The Weka Data Mining Software: An Update. *Special Interest Group on Knowledge Discovery in Databases Explorations.*, 11, 10–18.

Hansen, Marie Juul, Nana O. Rasmussen, and Grace Chung. 2008. A Method of Extracting the Number of Trial Participants from Abstracts Describing Randomized Controlled Trials. *Journal of Telemedicine and Telecare*, 14, 354–358.

Hirohata, Kenji, Naoaki Okazaki, Sophia Ananiadou, Mitsuru Ishizuka. 2008. Identifying Sections in Scientific Abstracts using Conditional Random Fields. Proceedings of the Third International Joint Conference on Natural Language Processing, pages 381-388.

Kiritchenko, Svetlana, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. Exact: Automatic Extraction of Clinical Trial Characteristics from Journal Publications. *BMC Medical Informatics and Decision Making*, 10.

Kirkman, M. S., M. McCarren, J. Shah, W. Duckworth, and C. Abraira. 2006. The Association Between Metabolic Control and Prevalent Macrovascular Disease in Type 2 Diabetes: The VA Cooperative Study in Diabetes. *Journal of Diabetes Complications*, 20, 75–80.

Lablanche, J. M., G. Grollier, J.-R. Lusson, J.-P. Bassand, G. Drobinski, B. Bertrand, S. Battaglia, B. Desveaux, Y. Juilliere, J.-M. Juliard, and *et al.* 1997. Effect of the Direct Nitric Oxide Donors Linsidomine and Molsidomine on Angiographic Restenosis After Coronary Balloon Angioplasty: The ACCORD Study. *Circulation*, 95, 83–89.

Lafferty, John, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabalistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learing*, pages 282–289.

Lease, Matthew and Eugene Charniak. 2005. Parsing Biomedical Literature. *Lecture Notes on Artificial Intelligence*, 3651, 58–69.

Light, Marc, Xin Ying Qiu, and Padmini Srinivasan. 2004. The Language of Bioscience: Facts, Speculations, and Statements in Between. In *Human Language Technologies–North American Association of Computational Linguistics 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24.

Lin, J.-W., C.-H. Chang, M.-W. Lin, Mark H. Ebell, and J.-H. Chiang. 2011. Automating the Process of Critical Appraisal and Assessing the Strength of Evidence with Information Extraction Technology. *Journal of Evaluation in Clinical Practice*, 17, 832–838.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19, 313–330.

Niu, Yun, Xiaodan Zhu, and Graeme Hirst. 2006. Using Outcome Polarity in Sentence Extraction for Medical Question-Answering. In *American Medical Informatics Association Symposium Proceedings*, pages 599–603.

Pang, Bo and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishing, Hanover, MA.

Platt, John C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods–Support Vector Learning*, pages 41–65.

Platt, John C. 1999. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. *Advances in Neural Information Processing Systems*, 11, 557–563.

Rocchio, J. J. 1971. Relevance Feedback in Information Retrieval In *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, 313–323.

Rogati, Monica and Yiming Yang. 2002. High-performing Feature Selection for Text Classification. In *Eleventh ACM International Conference on Information and Knowledge Management*, pages 659–661.

Sarker, Abeed, Diego Mollá-Aliod, and Cécile Paris. 2011a. Outcome Polarity Identification of Medical Papers. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 105–114.

Sarker, Abeed, Diego Mollá-Aliod, and Cécile Paris. 2011b. Towards Automatic Grading of Evidence. In *The Third International Workshop on Health Document Text Mining and Information Analysis*, pages 51–58.

Szarvas, György, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The Bioscope Corpus: Annotation for Negation, Uncertainty, and their Scope in Biomedical Texts. In *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 38–45.

Vapnik, Vladimir Naumovich. 2000. *The Nature of Statistical Learning Theory*. Springer, New York.

Yang, Yiming and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Fourteenth International Conference on Machine Learning*, pages 412–420.

Appendix A

Top 50 Terms for Information Gain

Tables A.1 and A.2 show the highest terms by information gain. Of these Table A.1 shows the top 25 terms and Table A.2 the next 25. Information gain favors common terms. The mention of *pain* and *patients* are strong indicators of a POE, while *glucose* and *insulin* are strong predictors of a non-POE. For the grouping chosen, the information gains for each term were averaged over all 10 folds.

Info Gain	Word	PoeYes	PoeNo	Total
0.047	pain	89	7	96
0.035	patients	476	308	784
0.034	glucose	7	61	68
0.034	primary	264	122	386
0.033	life	98	19	117
0.029	ofxlife	79	13	92
0.029	trial	445	288	733
0.029	thexprimary	178	69	247
0.027	insulin	9	57	66
0.025	qualityxofxlife	68	11	79
0.025	survival	81	17	98
0.025	efficacy	188	82	270
0.023	c	44	108	152
0.023	scale	78	17	95
0.023	patientsxwith	295	167	462
0.022	free	78	18	96
0.022	thexefficacy	111	36	147
0.020	treatment	350	219	569
0.020	fat	4	36	40
0.020	symptoms	100	32	132
0.020	secondary	146	61	207
0.018	death	79	22	101
0.018	score	121	47	168
0.018	fasting	3	31	34
0.017	events	148	66	214

Table A.1: The top 25 terms in 1-4 grams by information gain for a sample grouping of the data.

Info Gain	Word	PoeYes	PoeNo	Total
0.017	freexsurvival	40	5	45
0.017	protein	16	55	71
0.017	thextreatment	82	25	107
0.017	overallxsurvival	38	4	42
0.016	postoperative	44	7	51
0.015	endxpoints	44	8	52
0.015	overall	104	41	145
0.015	thextreatmentxof	42	7	49
0.015	qualityxof	75	23	98
0.015	quality	89	31	120
0.015	randomized	381	259	640
0.015	serum	19	56	75
0.014	diabetes	26	66	92
0.014	scores	95	36	131
0.014	costs	24	1	25
0.014	hazardxratio	64	18	82
0.014	no	302	192	494
0.014	concentrations	15	49	64
0.014	treatmentxof	83	29	112
0.014	improvement	93	35	128
0.014	typexdiabetes	14	47	61
0.014	concentrations	15	49	64
0.014	dietary	3	26	29
0.014	patientsxwere	145	70	215
0.014	levels	66	117	183

Table A.2: Terms 26–50 in 1–4 grams by information gain for a sample grouping of the data.

Appendix B

Additional Naïve Bayes Multinomial Tables

In Table B.1, the highest sensitivity is achieved with an information gain of 0.002. The best specificity, however, comes with an information gain of 0.001. The differences in sensitivity are not statistically significant. Specificity, accuracy, and diagnostic odds ratios decrease while IGM increases.

Table B.2 shows the results for naïve bayes multinomial with 4 grams, a cost ratio of 1:1, and a DFM of 5. Using 4–grams does not produce increases in accuracy. Accuracy falls as IGM rises, and specificity falls as IGM rises.

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	1:15	0.792±0.014	0.787±0.007	0.789±0.010	14.406±2.272
0.002	1:15	0.796±0.010	0.772±0.009	0.784±0.004	13.251±0.679
0.003	1:15	0.789±0.008	0.763±0.013	0.777±0.004	12.123±0.529
0.004	1:15	0.787±0.013	0.750±0.015	0.769±0.003	11.195±0.336

Table B.1: Naïve Bayes Multinomial — original data — 4–grams — Cost 1:15 — DFM 5 — IGM varies

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	1:1	0.808±0.012	0.770±0.007	0.790±0.008	14.334±1.842
0.002	1:1	0.810±0.004	0.759±0.007	0.786±0.003	13.506±0.472
0.003	1:1	0.805±0.006	0.746±0.005	0.777±0.003	12.109±0.420
0.004	1:1	0.810±0.007	0.728±0.008	0.771±0.004	11.473±0.482

Table B.2: Naïve Bayes Multinomial — original data — 4–grams — Cost 1:1 — DFM 5 — IGM varies

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	15:1	0.824±0.010	0.752±0.009	0.790±0.008	14.493±1.639
0.002	15:1	0.827±0.007	0.735±0.004	0.784±0.004	13.336±0.616
0.003	15:1	0.825±0.005	0.718±0.003	0.774±0.003	12.050±0.468
0.004	15:1	0.826±0.013	0.711±0.022	0.771±0.005	11.833±0.524

Table B.3: Naïve Bayes Multinomial — original data — 4–grams — Cost 15:1 — DFM 5 — IGM varies

In Table B.3, which is the same as the last table except that the cost ratio is 15:1, accuracy and specificity decrease as IGM increases.

Table B.10 provides results with 4–grams, information gain of 0.0005, the usual 1:15, 1:1, and 15:1 cost ratios. The accuracies in this table vary from 0.784–0.790. The best accuracies are found with a DFM of 5, but the DFM of 4 at the 15:1 cost ratio also comes in at a hair over 14.0 for the diagnostic odds ratio. These results are some of the most accurate for the original dataset - 79%.

In Table B.11, naïve bayes multinomial with 4–grams, an information gain of 0.0010, and DFMs of 4, 5, 6, and 7 are explored. The cost ratio varies in the usual 1:15, 1:1, and 15:1 pattern. Sensitivity is lower at higher DFM values, and it varies a lot more at a DFM of 4. DOR ratios vary from 13.4–14.1 (roughly) in this table. The best values are for a DFM of 5, and a cost ratio of either 1:1 or 15:1.

In Table B.12, naïve bayes multinomial with 4–grams, information gains of 0.0015, and DFMs of 4, 5, 6, and 7 are explored. The accuracy is highest with a DFM of 6, across the varying cost ratios. All of the accuracies in this table are between 78% and 79%, varying

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.0005	1:15	0.779±0.003	0.792±0.004	0.785±0.003	13.482±0.481
0.0010	1:15	0.790±0.017	0.778±0.010	0.785±0.005	13.467±1.099
0.0015	1:15	0.786±0.005	0.782±0.005	0.784±0.003	13.184±0.433
0.0005	1:1	0.796±0.003	0.781±0.004	0.789±0.003	13.991±0.485
0.0010	1:1	0.807±0.016	0.761±0.013	0.785±0.004	13.575±1.008
0.0015	1:1	0.806±0.004	0.763±0.005	0.786±0.003	13.376±0.495
0.0005	15:1	0.809±0.004	0.767±0.003	0.789±0.002	14.014±0.382
0.0010	15:1	0.823±0.013	0.746±0.011	0.786±0.004	13.850±1.020
0.0015	15:1	0.821±0.005	0.742±0.004	0.783±0.003	13.210±0.437

Table B.4: Naïve Bayes Multinomial — original data — 4–grams — DFM 4 — IGM and Cost vary — Sorted by Cost

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.0005	1:15	0.779±0.003	0.792±0.004	0.785±0.003	13.482±0.481
0.0005	1:1	0.796±0.003	0.781±0.004	0.789±0.003	13.991±0.485
0.0005	15:1	0.809±0.004	0.767±0.003	0.789±0.002	14.014±0.382
0.0010	1:15	0.790±0.017	0.778±0.010	0.785±0.005	13.467±1.099
0.0010	1:1	0.807±0.016	0.761±0.013	0.785±0.004	13.575±1.008
0.0010	15:1	0.823±0.013	0.746±0.011	0.786±0.004	13.850±1.020
0.0015	1:15	0.786±0.005	0.782±0.005	0.784±0.003	13.184±0.433
0.0015	1:1	0.806±0.004	0.763±0.005	0.786±0.003	13.376±0.495
0.0015	15:1	0.821±0.005	0.742±0.004	0.783±0.003	13.210±0.437

Table B.5: Naïve Bayes Multinomial — original data — 4–grams — DFM 4 — IGM and Cost vary — Sorted by IGM

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.0005	1:15	0.785±0.004	0.789±0.006	0.787±0.004	13.677±0.567
0.0010	1:15	0.791±0.005	0.786±0.006	0.789±0.004	13.978±0.659
0.0015	1:15	0.788±0.005	0.780±0.004	0.784±0.003	13.189±0.510
0.0005	1:1	0.805±0.004	0.774±0.006	0.790±0.003	14.153±0.508
0.0010	1:1	0.811±0.007	0.767±0.008	0.790±0.005	14.177±0.869
0.0015	1:1	0.807±0.005	0.759±0.005	0.785±0.004	13.329±0.658
0.0005	15:1	0.819±0.005	0.756±0.006	0.789±0.004	14.065±0.615
0.0010	15:1	0.827±0.008	0.746±0.006	0.789±0.006	14.171±1.012
0.0015	15:1	0.822±0.005	0.740±0.005	0.783±0.003	13.160±0.456

Table B.6: Naïve Bayes Multinomial — original data — 4–grams — DFM 5 — IGM and Cost vary — Sorted by Cost

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.0005	1:15	0.785±0.004	0.789±0.006	0.787±0.004	13.677±0.567
0.0005	1:1	0.805±0.004	0.774±0.006	0.790±0.003	14.153±0.508
0.0005	15:1	0.819±0.005	0.756±0.006	0.789±0.004	14.065±0.615
0.0010	1:15	0.791±0.005	0.786±0.006	0.789±0.004	13.978±0.659
0.0010	1:1	0.811±0.007	0.767±0.008	0.790±0.005	14.177±0.869
0.0010	15:1	0.827±0.008	0.746±0.006	0.789±0.006	14.171±1.012
0.0015	1:15	0.788±0.005	0.780±0.004	0.784±0.003	13.189±0.510
0.0015	1:1	0.807±0.005	0.759±0.005	0.785±0.004	13.329±0.658
0.0015	15:1	0.822±0.005	0.740±0.005	0.783±0.003	13.160±0.456

Table B.7: Naïve Bayes Multinomial — original data — 4–grams — DFM 5 — IGM and Cost vary - Sorted by IGM

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.0005	1:15	0.774±0.004	0.797±0.003	0.785±0.003	13.470±0.502
0.0010	1:15	0.776±0.003	0.796±0.006	0.786±0.003	13.624±0.586
0.0015	1:15	0.778±0.007	0.788±0.008	0.783±0.003	13.111±0.468
0.0005	1:1	0.791±0.004	0.779±0.004	0.785±0.004	13.367±0.593
0.0010	1:1	0.800±0.002	0.772±0.006	0.787±0.003	13.557±0.498
0.0015	1:1	0.801±0.009	0.767±0.011	0.785±0.004	13.338±0.705
0.0005	15:1	0.813±0.004	0.758±0.005	0.787±0.003	13.582±0.506
0.0010	15:1	0.819±0.003	0.748±0.006	0.785±0.004	13.454±0.641
0.0015	15:1	0.824±0.008	0.743±0.009	0.786±0.004	13.642±0.672

Table B.8: Naïve Bayes Multinomial — original data — 4–grams — DFM 7 — IGM and Cost vary — Sorted by Cost

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.0005	1:15	0.774±0.004	0.797±0.003	0.785±0.003	13.470±0.502
0.0005	1:1	0.791±0.004	0.779±0.004	0.785±0.004	13.367±0.593
0.0005	15:1	0.813±0.004	0.758±0.005	0.787±0.003	13.582±0.506
0.0010	1:15	0.776±0.003	0.796±0.006	0.786±0.003	13.624±0.586
0.0010	1:1	0.800±0.002	0.772±0.006	0.787±0.003	13.557±0.498
0.0010	15:1	0.819±0.003	0.748±0.006	0.785±0.004	13.454±0.641
0.0015	1:15	0.778±0.007	0.788±0.008	0.783±0.003	13.111±0.468
0.0015	1:1	0.801±0.009	0.767±0.011	0.785±0.004	13.338±0.705
0.0015	15:1	0.824±0.008	0.743±0.009	0.786±0.004	13.642±0.672

Table B.9: Naïve Bayes Multinomial — original data — 4–grams — DFM 7 — IGM and Cost vary — Sorted by IGM

DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
4	1:15	0.779±0.003	0.792±0.004	0.785±0.003	13.482±0.481
5	1:15	0.785±0.004	0.789±0.006	0.787±0.004	13.677±0.567
6	1:15	0.781±0.005	0.788±0.007	0.784±0.003	13.254±0.431
7	1:15	0.774±0.004	0.797±0.003	0.785±0.003	13.470±0.502
4	1:1	0.796±0.003	0.781±0.004	0.789±0.003	13.991±0.485
5	1:1	0.805±0.004	0.774±0.006	0.790±0.003	14.153±0.508
6	1:1	0.801±0.006	0.772±0.007	0.787±0.004	13.735±0.622
7	1:1	0.791±0.004	0.779±0.004	0.785±0.004	13.367±0.593
4	15:1	0.809±0.004	0.767±0.003	0.789±0.002	14.014±0.382
5	15:1	0.819±0.005	0.756±0.006	0.789±0.004	14.065±0.615
6	15:1	0.819±0.006	0.753±0.005	0.788±0.004	13.860±0.670
7	15:1	0.813±0.004	0.758±0.005	0.787±0.003	13.582±0.506

Table B.10: Naïve Bayes Multinomial — original data — 4–grams — IGM 0.0005 — DFM and Cost vary

DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
4	1:15	0.790±0.017	0.778±0.010	0.785±0.005	13.467±1.099
5	1:15	0.791±0.005	0.786±0.006	0.789±0.004	13.978±0.659
6	1:15	0.773±0.004	0.798±0.003	0.785±0.002	13.453±0.295
7	1:15	0.776±0.003	0.796±0.006	0.786±0.003	13.624±0.586
4	1:1	0.807±0.016	0.761±0.013	0.785±0.004	13.575±1.008
5	1:1	0.811±0.007	0.767±0.008	0.790±0.005	14.177±0.869
6	1:1	0.795±0.003	0.778±0.003	0.787±0.002	13.650±0.344
7	1:1	0.800±0.002	0.772±0.006	0.787±0.003	13.557±0.498
4	15:1	0.823±0.013	0.802±0.005	0.786±0.004	13.850±1.020
5	15:1	0.827±0.008	0.804±0.005	0.789±0.006	14.171±1.012
6	15:1	0.816±0.004	0.803±0.002	0.790±0.002	14.077±0.319
7	15:1	0.819±0.003	0.800±0.004	0.785±0.004	13.454±0.641

Table B.11: Naïve Bayes Multinomial — original data — 4–grams — IGM 0.0010 — DFM and Cost vary

DFM	Cost	Sensitivity	Specificity	Accuracy	DOR
4	1:15	0.786±0.005	0.782±0.005	0.784±0.003	13.184±0.433
5	1:15	0.788±0.005	0.780±0.004	0.784±0.003	13.189±0.510
6	1:15	0.793±0.004	0.784±0.003	0.789±0.001	13.976±0.225
7	1:15	0.778±0.007	0.788±0.008	0.783±0.003	13.111±0.468
4	1:1	0.806±0.004	0.763±0.005	0.786±0.003	13.376±0.495
5	1:1	0.807±0.005	0.759±0.005	0.785±0.004	13.329±0.658
6	1:1	0.812±0.005	0.760±0.003	0.787±0.003	13.747±0.447
7	1:1	0.801±0.009	0.767±0.011	0.785±0.004	13.338±0.705
4	15:1	0.821±0.005	0.742±0.004	0.783±0.003	13.210±0.437
5	15:1	0.822±0.005	0.740±0.005	0.783±0.003	13.160±0.456
6	15:1	0.830±0.005	0.738±0.004	0.787±0.003	13.838±0.514
7	15:1	0.824±0.008	0.743±0.009	0.786±0.004	13.642±0.672

Table B.12: Naïve Bayes Multinomial — original data — 4–grams — IGM 0.0015 — DFM and Cost vary

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	1:15	0.793±0.007	0.783±0.007	0.788±0.004	13.791±0.620
0.002	1:15	0.788±0.008	0.778±0.007	0.783±0.005	13.131±0.824
0.003	1:15	0.788±0.013	0.773±0.016	0.781±0.001	12.763±0.148
0.004	1:15	0.793±0.013	0.747±0.018	0.771±0.004	11.390±0.454
0.005	1:15	0.775±0.015	0.765±0.015	0.770±0.003	11.311±0.476

Table B.13: Naïve Bayes Multinomial — original data — 3–grams — Cost 1:15 — DFM 4 — IGM varies

remarkably little between the various experimental conditions. Diagnostic odds ratios are between 13.1 and 14.0, varying with the accuracy. The highest diagnostic odds ratio is with a DFM of 6 and a cost of 1:15.

Table B.13 has a cost ratio of 1:15, 3–grams, and a DFM of 4. Table B.14 is the same, except that the cost ratio is 1:1 in this table. Accuracy falls as information gain increases. Sensitivity is higher than specificity in this table, and that difference is statistically significant. Table B.14 shows the table for trigrams with a cost ratio of 1:1, a DFM of 4, and only 5 values of information gain. (The table was pruned, based on the earlier results with tables 2.7, 2.8, and 2.9.) Sensitivity is higher than specificity on this table, and accuracy and specificity fall as IGM increases. Table B.15 shows the same situation with a cost ratio

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	1:1	0.809±0.008	0.770±0.008	0.790±0.005	14.250±0.956
0.002	1:1	0.810±0.008	0.761±0.010	0.787±0.005	13.624±0.828
0.003	1:1	0.802±0.006	0.759±0.007	0.782±0.003	12.774±0.473
0.004	1:1	0.811±0.007	0.726±0.005	0.770±0.003	11.344±0.364
0.005	1:1	0.798±0.012	0.743±0.013	0.772±0.004	11.524±0.573

Table B.14: Naïve Bayes Multinomial — original data — 3–grams — Cost 1:1 — DFM 4 — IGM varies

IGM	Cost	Sensitivity	Specificity	Accuracy	DOR
0.001	15:1	0.822±0.009	0.756±0.010	0.791±0.006	14.416±1.098
0.002	15:1	0.827±0.008	0.737±0.008	0.784±0.004	13.432±0.602
0.003	15:1	0.829±0.007	0.725±0.006	0.780±0.004	12.835±0.643
0.004	15:1	0.831±0.012	0.694±0.014	0.766±0.002	11.156±0.320
0.005	15:1	0.837±0.005	0.698±0.010	0.771±0.006	11.933±0.769

Table B.15: Naïve Bayes Multinomial — original data — 3–grams — Cost 15:1 — DFM 4 — IGM varies

of 15:1. In this new situation, accuracy decreases as IGM increases. Sensitivity increases an IGM increases, while specificity decreases.

Appendix C

Additional Logistic Regression Tables

Table C.1 shows an information gain of 0.001, a varying cost ratio, and a DFM of 4 on the original dataset. There is good news and bad news here. The good news is that varying the cost ratio achieves a sensitivity/specificity tradeoff with unigrams under logistic regression. The bad news is the disappointing accuracy, which is less than 73% in all three cases.

The results from Table C.2 only use unigrams, a DFM of 6, and an information gain of 0.0010, on the original dataset. The accuracy varies from 71.2% to 71.5%, with a maximum at a balanced cost ratio. Sensitivity decreases with increasing cost, as expected, but specificity increases. The diagnostic odds ratios are on the low side for this set.

In Table C.3 one can see, for a DFM of 6, a cost ratio of 1:1, and an information gain of 0.0010, that increasing the N–Grams from 1 to 3 improves sensitivity and specificity. The value for bigrams is between these two values. The improvement from unigrams to

Cost	Sensitivity	Specificity	Accuracy	DOR
1:15	0.764±0.017	0.719±0.008	0.720±0.008	6.942±0.642
1:1	0.698±0.016	0.743±0.007	0.727±0.007	7.052±0.517
15:1	0.617±0.009	0.753±0.007	0.720±0.008	7.017±0.611

Table C.1: LR — original data — 1–grams — DFM 4 — Cost varies

Cost	Sensitivity	Specificity	Accuracy	DOR
1:15	0.668±0.006	0.760±0.008	0.712±0.005	6.390±0.326
1:1	0.727±0.008	0.701±0.009	0.715±0.005	6.286±0.331
15:1	0.783±0.006	0.637±0.008	0.713±0.004	6.326±0.253

Table C.2: LR — original data — 1–grams — DFM 6 — Cost varies

N–Grams	Sensitivity	Specificity	Accuracy	DOR
1	0.727±0.008	0.701±0.009	0.715±0.005	6.286±0.331
2	0.737±0.005	0.725±0.008	0.731±0.005	7.407±0.363
3	0.741±0.007	0.732±0.005	0.737±0.004	7.869±0.320

Table C.3: LR — original data — Cost 1:1 — DFM 6 — N varies

bigrams is about 1.5%, which is fairly dramatic for these experiments. The improvement from bigrams to trigrams is only 0.6%, which is fairly low. It seems that logistic regression is not competitive with naïve bayes multinomial, for which the highest accuracy exceeds 80% on the revised dataset. The accuracy for logistic regression is lower by about 5%, which is highly significant. Using trigrams helps, though, by about 2.0%. The diagnostic odds ratios here, 6.3–7.9, are also disappointing compared with NBM.