Outcome Polarity Identification of Medical Papers

Abeed Sarker , Diego Mollá-Aliod
Centre for Language Technology
Macquarie University
Sydney, NSW 2109
{abeed.sarker,diego.molla-aliod}@mq.edu.au

Cécile Paris
CSIRO – ICT Centre
Sydney, NSW 2122
cecile.paris@csiro.au

Abstract

A medical publication may or may not present an outcome. When an outcome is present, its polarity may be positive, negative or neutral. Information about the polarity of an outcome is a vital one, particularly for practitioners who use the outcome information for decision making. We model the problem of automatic outcome polarity identification as a three-way document classification problem and attempt to solve it via supervised machine learning. We combine domain knowledge and linguistic features of medical text, and apply natural language processing to extract features for the chosen classifiers. We introduce two novel features — Relative Average Negation Count and Sentence Signature — and show that they are effective in improving classification accuracy. We also include features, such as n-grams and semantic orientation of terms, that have been used for similar text classification problems in other domains. Using these features, we obtain a maximum accuracy of 74.9% for the classification problem. Our experiments suggest that through careful feature selection, machine learning can be used to solve this problem.

1 Introduction

The phenomenal growth of biomedical literature has presented medical practitioners, particularly those practicing Evidence Based Medicine (EBM), with the problem of information overload. The popular practice of EBM requires practitioners to review medical literature before making clinical decisions (Sackett et al., 1996; Greenhalgh, 2006). When reviewing medical publications, EBM practitioners are mostly interested in identifying the outcomes presented and their polarities. The polarity of an outcome can be positive (e.g. *the study shows that drug X is useful for patients suffering from condition Y*), negative (e.g. *the study suggests that drug X is not recommended for patients suffering from condition Y*) or the publication may present a neutral outcome or may not present an outcome at all (e.g. *the study does not produce conclusive results regarding the efficacy of drug X for condition Y*). Manually assessing the outcomes presented by multiple medical papers on a given topic is a time-consuming task and often cannot be efficiently performed at point of care (Ely et al., 1999). Hence, there is a strong need for automatic outcome polarity identification techniques to aid the decision making process of practitioners.

1.1 Motivation

In order to appease the problem of information overload faced by medical domain experts, research has focused on information retrieval, automatic summarisation and question answering of medical documents (Lin and Demner-Fushman, 2007; Fiszman et al., 2009). Intelligent text processing systems that perform automatic summarisation and question answering for this domain can benefit significantly from techniques that can automatically detect the polarity of outcomes presented in documents. Such techniques will be particularly useful for multidocument summarisation, where the detection of contradictory or consistent outcomes presented in separate documents is vital. Furthermore, recent research on

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

quality assessment of evidence presented in multiple medical documents has also acknowledged the importance of automatic polarity detection techniques for measuring consistency of outcomes in medical articles (Sarker et al., 2011). The research presented in this paper is motivated by these factors.

1.2 Contribution

We present a supervised learning approach to solve the problem of outcome polarity identification of medical publications. We focus particularly on medical publication types that are popularly used in EBM practice and model the problem as a three-way classification problem by separating outcomes presented in medical articles into three classes - *Positive*, *Negative* and *No Outcomes*. Despite the strong motivation behind automatic polarity identification of medical documents, there has not been any concrete research work attempting to solve this problem. We therefore approach this problem by building on and combining previously applied approaches for text classification, sentiment analysis, negation detection and polarity identification. One of the intents of this research work is to explore how the above mentioned approaches can be applied to the medical domain. We also present some novel feature selection ideas and show that some of these features increase classification accuracy.

2 Related Work

Research work related to ours has taken place under various umbrella terms (depending on the domain): sentiment analysis (Pang et al., 2002; Pang and Lee, 2004), semantic orientation (Turney, 2002), opinion mining (Pang and Lee, 2008), evidentiality (Chafe and Nichols, 1986), subjectivity (Lyons, 1981; Langacker, 1985) and many more. All these terms refer to the general method of extracting subjectivity or polarity from text (Taboada et al., 2010). A pioneering work in the area of sentiment analysis was performed by Pang et al. (2002), who attempted to automatically classify movie reviews as positive or negative. The authors applied three machine learning algorithms – Naive Bayes, Maximum Entropy and Support Vector Machines (SVMs) – and using features such as unigrams, bigrams, part-of-speech tags and adjectives, obtained a maximum average accuracy of 82.9% (over three-fold cross-validations). In their work, the best average accuracy was produced by the use of unigrams as features only. Turney's (2002) work was similar and involved the use of an unsupervised learning technique based on the mutual information (semantic orientation) between document phrases and the words 'excellent' and 'poor'. The semantic orientation of phrases were automatically computed using a search engine. His approach classified reviews as positive if they had a positive average semantic orientation and negative otherwise, achieving accuracies between 66% and 84% for different data sets. Following on from these works, research in this area has mostly focused on the binary polarity classification problem from opinionated pieces of text. Similar approaches have been applied for classifying the polarities of product reviews, political speeches and news. Pang and Lee (2008) provide an in-depth survey of approaches in this research area. Although similar in nature, the research work described in this paper differs significantly from approaches applied to sentiment analysis approaches for several reasons. The key reason is the complex nature of text in the medical domain with its domain specific terminologies and semantic relationships between terms (Athenikos and Han, 2010).

Research work closely related to ours in the medical domain is that by Niu et al. (2005; 2006). In their work, they perform polarity classification of sentences, obtained from medical article abstracts, using machine learning. The authors collect the abstracts from MEDLINE[1] and manually annotate each sentence into four classes – positive, negative, no outcome and neutral. Besides using unigrams and bigrams, the authors also use negations and semantic categories of medical concepts, and introduce *Change Phrases* – phrases that indicate the increase or decrease of a *good* or *bad* thing – as features. Precision and recall are shown to be approximately 79% over the four classes, using a data set of 1509 sentences and SVMs for learning. *Change Phrases* indicate the polarity of sentences and the concept is similar to *contextual valence shifters* (Polanyi and Zaenen, 2006; Kennedy and Inkpen, 2006) that have been successfully applied to sentiment classification

[1] http://www.nlm.nih.gov/databases/databases_medline.html

research.

We attempt to classify polarities at the document level, rather than at the sentence level. Our survey of literature in this domain did not reveal any work that attempts to address this specific problem despite its possible usefulness. The task itself is particularly challenging because each document may, and usually does, contain multiple sentences with differing polarities. Additionally, unlike the binary classification problem that sentiment analysis is usually modeled as, our work models the problem as a three-way classification (which, we believe, is the minimum number of classes required in the case of medical documents). Machine learning algorithms have been applied to solve various text classification problems, including those in the medical domain — such as identifying high quality medical articles (Kilicoglu et al., 2009). Among machine learning algorithms, SVMs (Vapnik, 1995) have clearly been the most popular for text classification, particularly because of their ability to robustly handle large feature sets and find globally optimum solutions (Uzuner et al., 2009; Taboada et al., 2010). We apply SVMs in our experiments and compare its performance with some other popular classifiers.

Another important aspect of our work is negation detection. Negated terms in medical text usually indicate the presence or absence of specific medical findings. Additionally, they may also indicate the polarity of the outcome presented in a medical article (e.g., drug X shows *no improvement* for patients suffering from condition Y). Recent research work has shown that information on the polarity of phrase-level assertions does not improve performance in a document level classification task (Goldstein and Uzuner, 2010). However, statistics based on the presence/absence of negations have not been incorporated for text classification in this domain. Negation identification has shown to markedly improve performance of medical information retrieval systems. Therefore, there has been a significant amount of work on automatic negation detection techniques in the medical domain, such as the works of Elkin et al. (2005) and Huang et al. (2007) . Rokach et al. (2008) provides a detailed survey of negation detection techniques for the medical domain. A popular and simple negation detection approach is NegEx (Chapman et al., 2001). It is a powerful, regular-expression-based algorithm and uses a list of phrases which, when present in the same sentence as disease names or findings, are indicative of negation. NegEx has been translated to other languages due to its effectiveness. We use a modified version of NegEx for negation detection in our experiments.

3 Data and Annotation

3.1 Data Collection

When collecting data, our focus was on articles that are commonly used for EBM. NLP research in the domain of EBM has shown that despite the presence of a large number of study types (also referred to as publication types) in the domain, only specific study types are commonly used in the practice[2]. These study types include Systematic Reviews, Meta-analyses, Clinical Trials (mostly Randomised Controlled Trials) and Cohort Studies. Although these are the preferred types of studies, Consensus Guidelines, Expert Opinion and Case Studies are also used in EBM practice when higher quality articles are not available on a specific topic. Sarker et al. (2011) provides an analysis of how publication types are distributed in real-life EBM practice.

To collect our data, we initially identified medical publications, which have been used in EBM practice, from the ‘Clinical Inquiries’ section of the Journal of Family Practice[3] (JFP). This section of JFP contains question-answer type evidence based reviews of specific medical topics that are generated by experts. The reviews also provide references to research articles from which the reviews are generated. We manually obtained a random sample of the abstracts of these references from MEDLINE using the PubMed[4] interface. We wanted to add diversity to our data set by incorporating article abstracts that do not belong to the Family Practice domain but have the potential to be used for EBM. To achieve this, we collected a sample of article abstracts belonging to the study types commonly used in EBM (mentioned above) directly from MEDLINE using the *PublicationType* filter.

[2] A list of publication types used by PubMed can be found at http://www.nlm.nih.gov/mesh/pubtypes.html

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

[4] http://www.ncbi.nlm.nih.gov/pubmed/

3.2 Annotation

We manually classified all the collected abstracts into three classes – *Positive*, *Negative* and *No Outcome*. During annotation, we use the following definitions for the three classes:

Positive: There is a clear indication that a medical process or intervention produces an outcome that is beneficial and/or serves its purpose; or a medical process or intervention is considered to be beneficial overall despite minor adverse effects; or when comparisons are made between two or more interventions or processes and the one that is the focus of the study is mentioned to be better. The following are two examples of positive outcomes:

> *'Depression scores on the Hamilton Rating Scale for Depression and Clinical Global Impressions-Severity scale significantly improved during the bupropion treatment phase.'*

> *'In a group of asymptomatic patients with first episode psychosis and at least one year of previous antipsychotic drug treatment, maintenance treatment with quetiapine compared with placebo resulted in a substantially lower rate of relapse during the following year.'*

Negative: There is a clear indication that an intervention or process produces an outcome that is not beneficial at all and/or is clearly not recommended; or when comparisons are made between two or more interventions or processes, the one that is the focus of the study is not mentioned to be the preferred choice. An example is as follows:

> *'There is a suggestion that routine surgical interference may be harmful by increasing the risk of caesarean section, and this agrees with data from other trials.'*

No Outcome: The outcome is neither positive nor negative or no outcome is specified at all. The latter can happen for systematic reviews or non-systematic reviews that do not present a single polarised answer. Also, when multiple comparisons are made without a final indication that a single process or intervention is preferred. The following is an example:

> *'There is not an important difference in the effects of bed rest compared with exercises in the treatment of acute low back pain, or seven days compared with two to three days of bed rest in patients with low back pain of different duration with and without radiating pain.'*

Our final test data set consists of 520 medical article abstracts, containing 9,221 sentences and 61,579 tokens (6,601 types). Among the 520 documents, 199 are annotated as positive, 161 as negative and 160 as no outcome instances. Approximately one-fourth of our data set consists of articles identified from JFP, while the rest were collected directly from MEDLINE using the approach described above.

The annotation was performed by four annotators (three medical domain experts and one computer scientist). There was about 40% overlap of the data among annotators and we computed Fleiss' Kappa (κ) to measure the extent of agreement among annotators. The formula for this statistic is given by:

$$\kappa = \frac{P_O - P_E}{1 - P_E} \tag{1}$$

where P_O is the observed agreement and P_E is the agreement expected by chance. The κ value we obtained is 70.6%, which falls within the range of values that is usually termed as "good agreement beyond chance".

3.3 Preliminary Analysis

We perform preliminary manual analysis on a small data set (separate from the 520 documents mentioned above) collected and annotated in the same fashion. Our analysis suggests that certain phrases play an important role in polarity determination. For example, *'significantly improves'*, *'no difference'*, *'no result'*, *'side effects'*, *'no improvement'* and similar phrases occur frequently in our data set and provide strong indications regarding the polarity. Similarly, negations also provide cues about the polarity at a document level, e.g., *'not recommended'*. However, a full abstract may, and usually does contain multiple occurrences of such phrases and therefore the presence or absence of these terms in a single sentence may not be indicative of overall polarity. Furthermore, our analysis also suggests that the semantic orientation of words in each abstract may have a correlation with the polarity of the outcome presented. For example, terms such as *'excellent'* tend to occur frequently in positively polarised documents, while terms such as *'unsuccessful'* are more likely to occur in negatively polarised documents. In our experiments, we explore all these possibilities. We attempt to combine various sentence level information to determine overall document polarities. Specific details about our preliminary analysis are provided in the next section, where we provide elaborate details about our feature selection techniques.

4 Methods

We model the problem of document level outcome polarity identification as a three-way classification problem. In this section we describe the features we use for classification, the feature selection techniques and provide justifications behind the choice of the selected features. We also attempt to explain how our feature selection ideas have been influenced by related research work.

4.1 Feature Sets

4.1.1 N-grams.

Word n-grams have been shown to be very important features in text classification problems (Taboada et al., 2010). We therefore use n-grams (n=1,2,3 and 4) from the article abstracts as our first feature set for experimentation. We experiment both with n-gram frequencies and presence. We also experiment with various combinations of n-grams. During preprocessing of the texts, we remove stop words and numbers, stem the individual words using the Porter stemmer and only keep n-grams with frequencies of greater than 4 over the whole data set.

We experiment with two further variations of n-gram feature sets. In the first variation, we only use n-grams from the conclusion sections of the abstracts. Our preliminary analysis suggests that sentences in the conclusion section of documents are most informative regarding the overall outcomes. For abstracts without explicit section headings, we use the last three sentences.

In our last variation, we replace specific medical concepts with a generic *'sem_type'* tag. We use MetaMap[5] to identify domain specific concepts as defined in the UMLS[6] (Unified Medical Language System). The UMLS provides a vast vocabulary of medical concepts and also broad semantic groups into which the concepts can be classified. For example, all disease names fall under the semantic category *Disease or Syndrome (dsyn)*. Replacing each occurrence of a disease or syndrome name with the generic tag ensures that the name does not have an influence on the classifiers used and reduces overfitting. Furthermore, it also enables the identification of specific term patterns in text that can be used for classification (explained later). Generic representations of medical problems have been used in text classification tasks in this domain before, and for our task, we use the same semantic groups as Uzuner et al. (2009): pathological function, disease or syndrome, mental or behavioral disfunction, cell or molecular dysfunction, virus, neoplastic process, anatomic abnormality, acquired abnormality, congenital abnormality and injury or poisoning.

4.1.2 Relative Average Negation Count

As already mentioned, our preliminary analysis suggests that negations provide cues about the overall polarity of an abstract, but the presence of negation in a single sentence may not determine document polarity. Our analysis also suggests that the total number of negations, or the negation count, over the whole document is generally greater for documents presenting a negative outcome than those presenting a positive outcome (the number of negations in documents presenting no outcomes vary significantly). At the same time, the negation count also tends to increase with the length of the abstracts and negations towards the end of the abstracts tend to have greater impact on the final outcome. We therefore use the *Relative Average Negation Count* (RANC) for each document as a feature and define it as follows:

$$RANC_d = \frac{\sum_{i=1}^{l}(n_i \times \frac{i}{l})}{l} \quad (2)$$

where d is a medical abstract containing l sentences in total and n_i is a negation detected in sentence i of the document. The equation shows that each negation is weighted by its relative position and the sum of all the weighted negations is divided by the length

[5] http://metamap.nlm.nih.gov/

[6] http://www.nlm.nih.gov/research/umls/

of the document to give RANC. We experimented with other representations of negations, such as using a vector of negation terms for each document, but found RANC to be the most effective.

To count the number of negations in a document, our algorithm uses a list of negation phrases based on the list used by NegEx (Chapman et al., 2001). In particular, NegEx attempts to identify negations in clinical narratives, and our modifications include adding negation phrases that commonly appear in published papers but are not included in NegEx's original list (e.g. *'not statistically'*). To calculate RANC, our algorithm searches each sentence of an abstract for the presence of any of the terms in our list. All the matches are summed using equation 2 to give the total negation count.

4.1.3 Semantic Orientation

We add a feature set to assess the effect of the semantic orientation of words on overall document polarity. We collect lists of positive and negative words from the General Inquirer dictionary (Stone et al., 1966)[7]. As this list is not specific to the medical domain, we manually modify both lists by removing terms that occur frequently in medical domain texts and whose semantic orientation should not be taken into account when identifying document polarities in this domain. These include terms such as *'disease'*, *'sickness'*, *'intervention'*, *'death'* and *'discharge'*. We have to rely on this time-consuming strategy since there are no such existing lists for the medical domain. For each document, we calculate its average semantic orientation, by counting the number of positive terms and the number of negative terms, subtracting the latter from the former and then dividing by the document length.

4.1.4 Change Phrases and Sentence Signatures

We use an approach similar to Niu et al. (2005; 2006) to identify sentence patterns or *change phrases*. In their work, the authors use a manually created list of *good*, *bad*, *more* and *less* words to identify patterns in sentences. The authors argue that the (sentence level) polarity of an outcome is often determined by how change happens (e.g., a good or bad thing is increased or decreased). For example, consider the following sentence:

> In these three postinfarction trials ACE inhibitor versus placebo significantly *reduced mortality*.

In the sentence, the word *reduce* is a *less* word while the word *mortality* is a *bad* word. Thus the sentence will have the pattern *less-bad* indicating that the sentence has a positive polarity. Similarly a sentence having the pattern *more-good* is likely to have a positive polarity while a sentence with the pattern *more-bad* or *less-good* is likely to have a negative polarity. In our work, we extend the idea of change phrases by including negations and medical semantic types in the patterns. Our intuition is that negations or semantic types can also significantly influence the polarity of sentences. For example, consider the following sentence (modified from the previous one):

> In these three postinfarction trials ACE inhibitor versus placebo *did not reduce mortality*.

The change phrase pattern for this sentence would still be *less-bad* despite the presence of the negation. A more correct pattern for the sentence should be *neg-less-bad* which incorporates the negation. Similarly, the following sentence:

> ... *increased* the probability of *heart failure*.

has a *more-semtype* pattern which may be indicative of negative polarity.

We generate two-term and three-term patterns from each sentence of each abstract and use them as a feature set. We call this feature set *Sentence Signatures* (SS).

4.2 Classification

Using the four feature sets mentioned in this section, we test the accuracy of four machine learning classifiers on our test data set. The four chosen classifiers are — Naive Bayes, Bayes Net, SVMs and C4.5 Decision Tree. Due to the relatively small amount of annotated data available to us, we perform 10-fold cross-validation in our experiments. We use the default implementations of all these classifiers in the software package Weka[8]. For the Bayes Net

[7] Available from http://www.wjh.harvard.edu/~inquirer/.

[8] http://www.cs.waikato.ac.nz/ml/weka/

classifier, we use the K2 search algorithm for local score metrics and the simple estimator for estimating conditional probability tables. For SVMs, we use an RBF kernel and John Platt's sequential minimal optimisation algorithm (Platt, 1999); and solve our multi-class problem using pairwise (1-vs-1) classification. Further details of these classifiers can be found in the documentation provided with the software package.

5 Results and Discussion

Table 1 presents the results of the four classifiers over various combinations of features. The horizontal lines of the table divide the features into groups and the best accuracy obtained for a specific group is shown in bold. The results indicate that the n-grams play an important role in the classification problem, which is consistent with findings in other domains. More specifically, use of uni-, bi- and tri-grams as features show clear improvements in classification but adding longer n-grams does not appear to be beneficial. The results of classification using n-grams only also show that classification accuracies are not significantly different between the use of word frequencies (F) and presence (P). Using n-grams from full abstracts always performs better than using n-grams from conclusion sentences (C) only. Replacing medical terms belonging to specific semantic categories with a generic tag (M) also tends to give better classification accuracies.

Introduction of RANC and SS as features has a positive impact on classification accuracies. The increase in accuracies for our tree-based classifier (C4.5) upon the addition of RANCs as features is particularly significant, which is a clear indication of the importance of this feature set. The highest accuracy we obtain is 74.9% using SVMs for classification and n-grams (n=1,2,3), RANCs and SSs as features. SVMs consistently outperform other classifiers in all the experiments we present, which is what we expected based on the success of SVMs in text classification tasks.

The use of SO as a feature set does not seem to have a positive effect on classification accuracies. This, however, may be due to the absence of a domain-specific dictionary for semantic orientation of terms. Despite our modifications of the list of positive and negative words, this feature set does not play a role in determining polarity. A more in-depth analysis of domain specific terms is required to assess the applicability of this feature set.

Manual analysis of the mis-classified instances reveals a number of key reasons behind the classification errors. Many systematic and non-systematic reviews in our data set present outcomes from multiple trials or studies of both polarities (which is a common feature of this publication type). Manual annotation of these abstracts is easier because the annotators can take the context of the articles into account and identify the overall message represented in the text. When multiple comparisons are presented in a review without a final polarised outcome, we annotated that review as no outcome. However, the n-grams generated by such articles have similarity to articles from the positive and negative classes and are therefore hard to separate automatically.

Furthermore, while RANC plays an important role in identifying negative polarities, introduction of this feature also causes some instances, particularly those with no outcomes, to have large RANCs. This happens when negations occur in multiple places of the abstract text, but none is associated with the final outcome. Negation phrases such as *'no outcome'* and *'no result'* are common in the No Outcome class while various forms of negations are present in articles belonging to the Negative class (e.g, *'not recommend'*). A deeper analysis of negations to see which terms occur more frequently in each of the two classes may reduce this problem.

Finally, the structure and content of the article abstracts vary significantly depending on the type of study. For example, a meta-analysis is considerably different from a randomised controlled trial. A more elaborate approach involving identification of publication types prior to classification and training and testing classifiers on texts belonging to specific study types would perhaps yield better results. Increasing the size of the training set is also likely to result in improved accuracy. However, that will also require significant time contribution for annotation.

6 Conclusion and Future Work

In this work we show that the problem of medical document polarity identification can be treated as a

Features	**Naive Bayes**	**BayesNet**	**SVM**	**C4.5**
Unigrams (P)	65.2	62.3	67.7	55.0
Unigrams (F)	65.2	62.3	68.3	55.0
Unigrams (P, C)	61.3	60.8	62.5	53.7
Unigrams (F, C)	61.3	60.8	62.5	53.7
Unigrams (M, P)	66.3	62.7	67.9	55.2
Unigrams (M, F)	66.3	62.7	**69.4**	55.2
Unigrams (M, P, C)	62.5	60.8	62.9	53.7
Unigrams (M, F, C)	62.5	60.8	62.9	53.7
Unigrams + bigrams (P)	70.4	63.8	72.7	62.3
Unigrams + bigrams (F)	70.4	63.8	72.9	62.3
Unigrams + bigrams (P, C)	66.0	62.7	69.8	60.5
Unigrams + bigrams (F, C)	65.9	62.7	70.0	60.3
Unigrams + bigrams (M, P)	70.1	63.5	**73.9**	63.7
Unigrams + bigrams (M, F)	66.3	62.7	68.3	60.4
Unigrams + bigrams (M, P, C)	63.1	60.8	65.6	59.0
Unigrams + bigrams (M, F, C)	63.0	61.1	66.3	58.1
N-grams (n=1,2 and 3)(P)	70.6	62.7	74.0	60.6
N-grams (n=1,2 and 3)(F)	70.6	62.7	73.9	60.6
N-grams (n=1,2 and 3)(M, P)	70.8	62.7	**74.2**	61.3
N-grams (n=1,2 and 3)(M, F)	70.8	62.6	74.0	61.3
N-grams (n=1,2,3 and 4)(P)	70.8	62.3	**73.0**	61.5
N-grams (n=1,2,3 and 4)(F)	70.8	62.3	72.6	61.5
N-grams (n=1,2,3 and 4)(M, P)	70.8	62.7	72.9	61.3
N-grams (n=1,2,3 and 4)(M, F)	70.6	61.9	72.3	61.3
Unigrams + bigrams + RANC (M, P)	72.1	68.1+	73.3	70.1+
N-grams (n=1,2 and 3) + RANC (M, P)	71.7	67.3	**74.4**	68.6
Unigrams + bigrams + RANC + SO(M, P)	71.5	66.7	73.3	67.5
N-grams (n=1,2 and 3) + RANC + SO (M, P)	71.6	66.5	**74.4**	67.9
Unigrams + bigrams + RANC + SS(M, P)	72.3+	67.3	73.6	66.5
N-grams (n=1,2 and 3) + RANC +SS (M, P)	72.3+	66.9	**74.9***	68.9
N-grams (n=1,2 and 3) + RANC + SO + SS(M, P)	71.7	67.1	**74.7**	68.0

Table 1: Classifier accuracies for various combinations of features. (P) represents word presence, (F) represents word frequencies, (M) indicates medical terms replaced from the text using the generic tag, (C) indicates only conclusion sentences used. RANC – Relative Average Negation Count, SO – Semantic Orientation, SS – Sentence Signatures. Best result produced by a combination of features shown in bold. Best overall accuracy indicated by *. Best accuracy achieved by a specific classifier indicated by +.

classification problem and machine learning algorithms can be used to solve this problem. Our work is the first of its kind in this domain and therefore we incorporate relevant techniques from related research work. Using carefully extracted linguistic features and domain knowledge, we obtain 74.9% accuracy on a data set that contains a variety of medical publication types. Post-classification analysis of our data reveals a number of possible research tasks that can be performed to further improve classification accuracies. Some classification errors can be attributed to subtle weaknesses in our automatic feature generation techniques and also the similarity in content among documents of differing classes.

Incorporating accurate, automatic outcome polarity detection techniques can considerably benefit automatic summarisation and question answering systems in this domain. This will require improving the accuracy of our classifiers and we will address some possibilities in our future work.

One possibility is to automatically identify the context when extracting features such as words, phrases, negations and signatures. Our analysis showed that in EBM practice, the same article may have different polarities depending on the query posed by the practitioner. The context may also be given by the topic of the article.

Our approach of using conclusion sentences can be improved through the use of classifiers that can identify conclusion/outcome sentences from medical abstracts automatically. Such a classifier has recently been presented by Kim et al. (2011) and it has been shown to be highly accurate at identifying sentences presenting medical outcomes. Future work will therefore involve the use of this method of sentence classification and use only sentences classified as ‘outcomes’.

Since the content of publications in this domain vary with the publication types, an approach that automatically detects the publication types followed by the application of customized feature extraction techniques is likely to be more accurate. Careful analysis and ranking of the semantic orientation of words in this domain can also be effective in obtaining higher classification accuracies.

Finally, it is likely that performance can be improved by using a larger data set. This will also make it possible to use separate training and test sets so that the parameters of the classifiers can be optimised based on the training data and then be tested on the test data.

We will attempt to incorporate all the abovementioned ideas in our future work. Considering the strong motivation behind an approach for automatic polarity detection, improvements in classification accuracy will be extremely beneficial for various automatic text processing applications in this domain.

Acknowledgments

This research work is jointly funded by Macquarie University and CSIRO.

References

Sofia J. Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99:1–24, July.

Wallace Chafe and Johanna Nichols. 1986. *Evidentiality: The Linguistic Coding of Epistemology*. Ablex, Norwood, NJ.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp*, pages 105–109.

Peter Elkin, Steven Brown, Brent Bauer, Casey Husser, William Carruth, Larry Bergstrom, and Dietlind W. Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, M. Lee Chambliss, and Eric R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361, August.

Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, and Thomas C. Rindflesch. 2009. Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics*, 42(5):801–813.

Ira Goldstein and Ozlem Uzuner. 2010. Does Negation Really Matter? In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 23–27.

Trisha Greenhalgh. 2006. *How to read a paper: The Basics of Evidence-based Medicine*. Blackwell Publishing, 3 edition.

Yang Huang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *JAMIA*, 14(3):304–311, May.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 33(1):110–125.

Halil Kilicoglu, Dina Demner-Fushman, Thomas C. Rindflesch, Nancy L. Wilczynski, and Brian R. Haynes. 2009. Towards automatic recognition of scientifically rigorous clinical research evidence. *JAMIA*, 16(1):25–31, January.

Su Nam N. Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC bioinformatics*, 12 Suppl 2.

Ronald W Langacker, 1985. *Iconicity in Syntax*, chapter Observations and speculations on subjectivity, pages 109–150. Amsterdam and Philadelphia.

Jimmy J. Lin and Dina Demner-Fushman. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

John Lyons. 1981. *Language, Meaning and Context*. Fontana, London.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *Proceedings of the AMIA Annual Symposium*, pages 570–574.

Yun Niu, Xiaodan Zhu, and Graeme Hirst. 2006. Using outcome polarity in sentence extraction for medical question-answering. In *Proceedings of the AMIA Annual Symposium*, pages 599–603.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proc EMNLP*.

John C. Platt, 1999. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA.

Livia Polanyi and Annie Zaenen, 2006. *Computing Attitude and Affect in Text: Theory and Applications*, chapter Contextual valence shifters, pages 1–10. Springer, Dordrecht.

Lior Rokach, Roni Romano, and Oded Maimon. 2008. Negation recognition in medical narrative reports. *Information Retrieval*, 11:499–538. 10.1007/s10791-008-9061-0.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

Abeed Sarker, Diego Mollá-Aliod, and Cecile Paris. 2011. Towards automatic grading of evidence. In *Proceedings of LOUHI 2011 Third International Workshop on Health Document Text Mining and Information Analysis*, pages 51–58.

Philip J Stone, Dexter C Dunphy, Marshall S Smith, and Daniel M Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2010. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, US. Association for Computational Linguistics.

Ozlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine Learning and Rule-based Approaches to Assertion Classification. *JAMIA*, 16:109–115.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.