

# Medical Language Processing for Patient Diagnosis Using Text Classification and Negation Labelling

Brian Mac Namee, PhD<sup>1</sup>, John D. Kelleher, PhD<sup>1</sup> and Sarah Jane Delany, PhD<sup>2</sup>

<sup>1</sup>School of Computing, Dublin Institute of Technology, Dublin, Ireland  
and <sup>2</sup>Digital Media Centre, Dublin Institute of Technology, Dublin, Ireland

## Abstract

*This paper describes the approach of the DIT AIGroup to the i2b2 Obesity Challenge to build a system to diagnose obesity and related co-morbidities from narrative, unstructured patient records. Based on experimental results a system was developed which used knowledge-light text classification using decision trees, and negation labelling.*

## Introduction

This paper will describe the approach taken by the DIT AIGroup ([www.comp.dit.ie/aigroup](http://www.comp.dit.ie/aigroup)) to the 2008 Second i2b2 Shared-Task in Natural Language Processing for Clinical Data (hereafter referred to as the challenge), organised by the Informatics for Integrating Biology and the Bedside (i2b2) ([www.i2b2.org](http://www.i2b2.org)) centre at Partners HealthCare System in Boston, Massachusetts in the USA. The purpose of the challenge was to build a classification system that could diagnose obesity and a range of co-morbidities, exhibited by patients, based on unstructured, narrative patient records, captured electronically from health professionals.

Our approach is one of knowledge-light text classification. This has the advantage that building the system requires little expert domain-specific (in this case medical) knowledge, and yet the resulting system performs to a high level of accuracy. Additionally, the system labels negations within the texts which helps improve the accuracy of the classifiers created. The classification technique used by the system is a decision tree, which is unusual as decision trees are not typically associated with good performance in text classification problems, due to the large number of features involved.

The paper is structured as follows. The next section describes the challenge, and in particular the data provided by the challenge organisers. Following this, a high-level description of our system is given. Next, the two major components of our system are described in the Negation Labelling and Text Classification sections. The experiments performed to evaluate the system are then described. Finally, thoughts on the system, and suggestions as to how it might be improved are given.

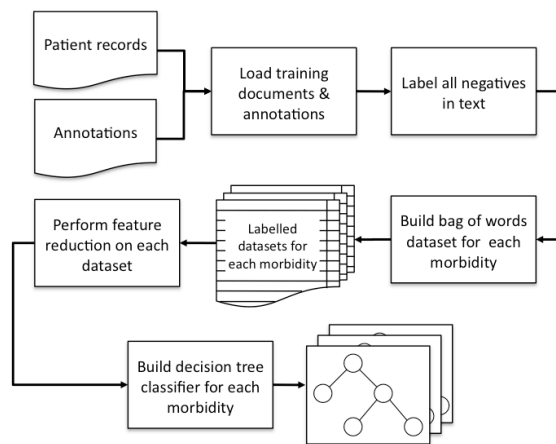
## The Challenge

The challenge asked participants to replicate the judgements of obesity experts in diagnosing obesity and a range of common co-morbidities based on unstructured, narrative patient records. A set of training data was provided which contained 725 fully anonymised patient records, with associated annotations. Patient records varied in length from approximately 600 to 6,000 words and contained various abbreviations, codes, highly technical medical language, and typographical errors.

Each patient record was annotated with diagnoses for obesity and 15 co-morbidities such as diabetes, asthma, and gout. Two kinds of annotations were provided for each disease: *textual* and *intuitive*. In providing textual annotations the obesity experts were asked to use only information that was explicitly given in the record. Textual judgements could be one of four classes: (Y) the patient had the disease, (N) the patient did not have the disease, (Q) based on only the information in the text the patient may have the disease, but the diagnosis is questionable, and (U) from the text alone the diagnosis is unknown. For intuitive judgements the experts could rely on information implicitly provided in the patient records. Intuitive judgements could be one of three classes: (Y), (N), and (Q) as described above. The distribution of classes for each disease were heavily skewed, with some classes (particularly N and Q for textual judgements, and Q for intuitive judgements) barely represented at all.

## System Overview

Figure 1 shows a high-level schematic diagram of our system illustrating its core components. Beginning from the supplied patient records and annotations (the process for textual and intuitive annotations is the same so no differentiation between the two is made) the system first loads the records into a data structure that associates the supplied annotations with each record. Negation labeling is then performed on each record.



**Figure 1.** Total allergy alerts, overridden alerts, or drug order cancelled

When negation labelling is complete the records are converted into a *bag-of-words* representation, a standard machine learning representation suitable for text classification. A separate bag-of-words dataset is produced for each disease under consideration, which leads to individual classification systems for diagnosing each disease. Feature reduction is then performed. The details of these processes will be discussed in the Text Classification section. There are a range of classification techniques that are suitable for text classification, and the selection of the most appropriate one for this problem will be discussed in the Experimental Results section. The next section will describe why and how negation labeling is carried out in our system.

### Negation Labelling

In linguistics, negation is the process that turns an affirmative statement (e.g. *it is a sunny day*) into its opposite denial (e.g. *it is not a sunny day*). Negation is an important and frequent phenomenon in clinical texts, as Chapman et al<sup>2</sup> note “*in clinical reports the presence of a term does not necessarily indicate the presence of the clinical condition represented by that term. In fact, many of the most frequently described findings and diseases in discharge summaries, radiology reports, history and physical exams, and other transcribed reports are denied in the patient*”. Furthermore, the denial of a condition is qualitatively different from the absence of any mention of the condition, as it indicates that an investigation for that condition has been carried out.

Huang et al<sup>7</sup> decompose the structure of negation into three components: a *negation signal*, a *negated phrase* containing one or more concept(s), and optionally some *supporting feature* (pattern), which helps us locate the negated phrase. They illustrate this

decomposition using the following example: *There is no evidence of cervical lymph node enlargement*. In this example, *no* is the negation signal used to denote that the following concept is negated; *cervical lymph node enlargement* is the negated phrase; while *evidence of* is the supporting feature.

Handling negation is difficult for several reasons, including: (1) identifying the *scope* of a negation is often difficult, for example a negation can extend across sentence boundaries; (2) negation is a complex linguistic phenomenon with many different *types*<sup>7,10</sup> (3) it is difficult to handle multiple occurrences of a term in a sentence some of which are negated<sup>2</sup>.

Previous prominent work on negation labelling includes the following systems. The *NegFinder* system<sup>10</sup> uses a lexical scanner with regular expressions and a parser that uses a restricted context-free grammar to identify negated phrases in discharge summaries and surgical notes. The *NegEx* system<sup>7</sup> uses regular expressions to identify negation signals in the text and negates all UMLS<sup>12</sup> terms within a window of 5 words following each negation signal. The system developed by Elkin et al<sup>6</sup> uses a negation ontology containing operators and associated rules. Operators consisted of two sets of terms with one set starting negations and another set stopping the propagation of negations. In the *ChartIndex* system<sup>7</sup> negations are classified based upon the syntactical categories of negation signals, and negation patterns, using regular expression matching. Negated terms are then located in parse trees using negation grammar rules.

Inspired by the NegEx system our system uses regular expressions to find negation signals within the text and negates all terms near to the signal. Two kinds of negation signals are identified – preceding negations and following negations. Preceding negations precede the words they negate. Examples include *absence of*, *no complaints of*, *no radiographic evidence of*, and *sufficient to rule the patient out for*. Following negations follow the words they negate and examples include *unlikely*, *free*, and *has been ruled out*. Our implementation uses 125 preceding negation signals and 7 following negation signals. Words within a window of 5 words to the appropriate side of a negation signal are labeled as negated, however this window can be truncated by the end or beginning of a sentence or the presence of another negation signal.

It is important to note that our implementation of negation labelling left out some of the details of the NegEx algorithm (the use of UMLS, conjunctions and pseudo-negatives), and does not address some issues outside the scope of the NegEx algorithm, for

example negation across multiple sentences. However, as will be seen in the Experimental Results section it proved useful for the challenge.

## Text Classification

Text classification<sup>11</sup> has been researched heavily in recent times in both the machine learning and the information retrieval communities. It has been used for such diverse tasks as spam detection<sup>4</sup>, document routing<sup>8</sup>, categorising web pages<sup>5</sup>, as well as in medical language processing<sup>14</sup>.

The traditional representation for documents in a text classification problem is the vector-space model<sup>1</sup>. Each document is represented as a vector in  $n$ -dimensional space, where each dimension represents a word in the combined vocabulary of all the training documents. The value of each attribute (aka feature) in the document vector represents the frequency of occurrence of the word in the document. This representation is commonly known as the *bag-of-words* representation.

The values of the features within a bag-of-words representation can be binary, i.e. they indicate only the presence or absence of a word, or numeric, i.e. they indicate the frequency of a word in a document. In our system a numeric representation using raw frequencies (i.e. the number of times a particular word appears in a document) is used.

One of the unique characteristics of textual data when converted to a vector space model is the very large number of features involved. The training corpus provided for the challenge contains a vocabulary of approximately 25,000 words that results in feature vectors of the same length. It is advantageous, and common practice, to reduce this number of features as much as possible before attempting to build a classifier, as many classification techniques do not cope well with high dimensional data.

Zipf's Law<sup>15</sup> captures the distribution of the frequency of occurrence of words in natural language texts and states that the frequency of any word is inversely proportional to its rank in the frequency table. Zipf's law suggests that high frequency words occur in too many documents to be useful in prediction, and low frequency words are too rare to be of value. We utilise common feature reduction techniques; *stop-word removal* to remove the high frequency words, and *document frequency* to remove the low frequency words.

Stop-word removal removes stop-words such as *the*, *a*, and *is* which do not serve as good predictors of the category of a document as they are so prevalent. We remove them from the vocabulary using a list of

common stop-words, many of which are publicly available. It is important to note that stop-word removal can only be performed after negation labelling as certain stop-words are important in many of the negation signals used within our system.

Document frequency removes terms that occur in at most  $n$  documents in the training set (popular values for  $n$  range from 1 to 3, we have used 3). Such infrequent terms (which often include spelling mistakes) will not have significant predictive power and so are not likely to help in classification. The combination of stop-word removal and document frequency together reduce the vocabulary arising from the obesity challenge training corpus from approximately 25,000 to 6,500 words.

Once the data is suitably prepared any classification technique can be used in order to perform the actual diagnoses.  $k$ -nearest neighbour ( $k$ NN), Naïve Bayes, and Support Vector Machines (SVM) have all been shown to be particularly well suited to text classification problems<sup>3, 9</sup>, but as will be seen in the following section a decision tree proved to be the most accurate classifier in this case. This is interesting as it goes against the perceived wisdom in text classification.

## Experimental Results

There were two main stages to the experimental work that was performed in order to determine the best system for this challenge. The first step was to determine which classification technique would be most suited to the problem. The second step was to determine if adding negation labelling had any impact on the performance of the system. A range of other investigations were also performed in order to determine other factors, such as: the amount of feature reduction to employ, the effect of stop word removal, and to confirm our results on other morbidities. However, these will not be discussed here due to space limitations.

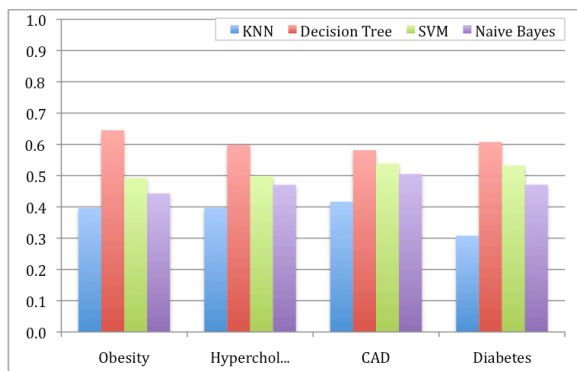
The first experiment attempted to determine which classification technique would be most suited to working with the obesity challenge data. For this experiment the Obesity, Hypercholesterolemia, CAD and Diabetes morbidities were selected, as these had the most even distributions in their annotations. Furthermore, for this experiment only intuitive annotations were considered and no negation labelling was employed. A 10-fold cross validation was performed to determine the diagnosis accuracy achievable for each of the selected morbidities for a selection of possible classification techniques.

For consideration we selected the  $k$ NN, SVM, Naïve Bayes and decision tree classification techniques – the first three due to their popularity for text classification and the last as an interesting counterpoint. The specific implementations of these techniques used were the IBK, SMO, Naive Bayes and J48 implementations available in the widely used Weka machine learning toolkit<sup>13</sup>. The implementations from Weka were used with minimal adjustment to their default parameter sets.

The F1 measure (using macro-averaging)<sup>11</sup> resulting from each of these experiments is shown in Table 1. A comparison of these measures is also shown in Figure 2. It is worth noting that the scores appear low due to macro averaging the precision and recall scores of the low frequency classes for each morbidity. It is clear from these experiments that the decision tree classifier significantly out performs all of the others, with SVM a distant second. Our suggestions as to why this is the case will be discussed in the Conclusions & Future Work section, but the decision tree classifier was selected for all of the remaining experiments.

	Obesity	Hyperchol...	CAD	Diabetes
<b><math>k</math>NN</b>	0.397	0.397	0.417	0.308
<b>Decision Tree</b>	0.645	0.599	0.581	0.607
<b>SVM</b>	0.493	0.499	0.539	0.533
<b>Naïve Bayes</b>	0.443	0.471	0.505	0.471

**Table 1.** Comparison of F1 measures resulting from 10-fold cross validations on 4 morbidities using 4 different classification algorithms.



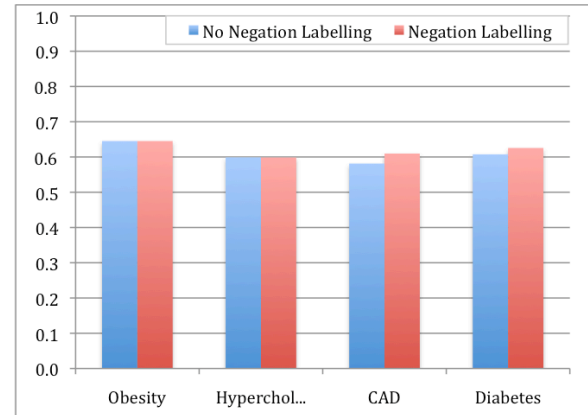
**Figure 2.** Comparison of F1 measures resulting from 10-fold cross validations on 4 morbidities using 4 different classification algorithms.

The second experiment was performed in order to determine if negation labelling had a notable impact on diagnosis accuracy. This experiment used data for the same morbidities as the first experiment. This

time 10-fold cross validations were performed using a data set generated without using negation labelling, and a dataset generated using negation labelling. Only the decision tree classifier was used in this case. The F1 measure results for this experiment are shown in Table 2, and illustrated in Figure 3. Although the impact of negation labelling is marginal it either improves or does not seriously negatively affect the results. Based on these results negation labelling was included in the final system.

	Obesity	Hyperchol...	CAD	Diabetes
<b>No Neg Labelling</b>	0.645	0.599	0.581	0.607
<b>Neg Labelling</b>	0.645	0.598	0.610	0.625

**Table 2.** Comparison of F1 measures for 4 morbidities with negation labelling turned on and off.



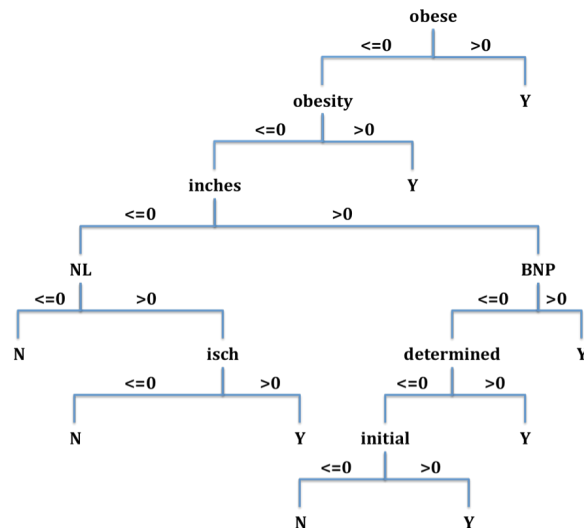
**Figure 3.** Comparison of F1 measures for 4 morbidities with negation labelling turned on and off.

## Conclusions & Future Work

Our experimental results point to two interesting conclusions. The first is that, contrary to expectations, the most suitable classification algorithm for the task was the decision tree. The second is that negation labelling, even if our implementation is crude, can positively impact diagnosis accuracy.

Regarding the first point, we believe that the reason for this is that the challenge corpus contains a small number of highly predictive words, the presence or absence of which can be captured easily in relatively uncomplicated decision trees. To illustrate this point a sample decision tree for intuitive obesity diagnoses is shown in Figure 4. This simple tree contains variants of obesity at its root and a very small number of other branches. The impact of the terms included in the tree is very high, an impact that could be

swamped in other algorithms with better generalization capability such as *k*NN and SVM.



**Figure 4.** A sample decision tree for the obesity morbidity showing the small number of terms required to make good predictions.

Our explanation for the impact of negation is along similar lines. The fact that one of these highly predictive terms is negated is of great information value, and so including this fact within the decision tree (which experiments have shown is a relatively rare occurrence) leads to greater accuracy.

While the work described in this paper is promising there are many ways in which it could be improved. Directions that we have identified as being particularly promising are introducing more sophisticated negation labelling, using an ontology such as UMLS<sup>12</sup> to label medical terms, handling the skewed distributions in the dataset, and finally using ensembling techniques that would allow a two stage classification combine all of the individual morbidity classifications to generate a more accurate result.

## References

1. Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval, Addison-Wesley, Wokingham, UK, 1999.
2. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301-310, 2001.
3. Colas F, Brazdil P. Comparison of SVM and some older classification algorithms in text classification tasks, In Proc. of the Conference on Artificial Intelligence in Theory and Practice, International Federation for Information Processing, 217:169-178, 2006.
4. Delany SJ, Cunningham P, Tsybmal A, Coyle L. A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems* 18 (2005b) 187–195
5. Dumais S, Chen H. Hierarchical classification of web content. In: *Procs of SI-GIR'00*. (2000) 256–263.
6. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, Wahner-Roedler DL. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making* 2005, 5:13, 2005.
7. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14:304-311, 2007.
8. Iyer RD, Lewis DD, Schapire RE, Singer Y, & Singhal A. Boosting for document routing. In *Procs of CIKM-00*, 9th ACM International Conference on Information and Knowledge Management, pp70–77, 2000.
9. Joachims T. Text categorization with support vector machines: learning with many relevant features. In Nédellec, C., Rouveirol, C., eds.: *Proceedings of ECML-98*, 10th European Conference on Machine Learning. Number 1398 in LNCS, Springer Verlag, Heidelberg, DE (1998) 137–142, 1998.
10. Mutalik PG, Deshpande A, Nadkarni P. Use of general purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the American Medical Informatics Association*, 8:589-609, 2001.
11. Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys* 34:1–47, 2002.
12. Universal Medical Language System. URL: [www.nlm.nih.gov/research/umls](http://www.nlm.nih.gov/research/umls)
13. The Weka Machine Learning Toolkit. URL: [www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka).
14. Wilcox A, Hripcsak G. The role of domain knowledge in automating medical text report classification. *Journal of the American Medical Informatics Association* 10 (2003) 330–338
15. Zipf G. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley (1949)