Text Classification for Medical Informatics: A Comparison of Models for Data Mining Radiological Medical Records

William B. Claster[1], Subana Shanmuganathan[2], Nader Ghotbi[3], Philip J. Sallis[4]

[1]Information Communication Technology, Ritsumeikan Asia Pacific University, Beppu, Oita, Japan

[2]Geoinformatics Research Center, Auckland University of Technology, Auckland, New Zealand

Geoinformatics Research Centre School of Computing and Mathematical Sciences Auckland University of Technology Level 7, WZ building 350 Queen Street Private Bag 92006 Auckland 1142 New Zealand Phone: +64 9 921 9999 x5803 Fax: +64 9 921 9807. subana.shanmuganathan@aut.ac.nz

[3]Health and Environment, Ritsumeikan Asia Pacific University, Beppu, Oita, Japan

[4]Geoinformatics Research Center, Auckland University of Technology, Auckland, New Zealand

Geoinformatics Research Centre School of Computing and Mathematical Sciences Auckland University of Technology Level 7, WZ building 350 Queen Street Private Bag 92006 Auckland 1142 New Zealand Phone: +64 9 921 9999 x5803 Fax: +64 9 921 9807. philip.sallis@aut.ac.nz.

Word Count:5638

Number of Tables:4

Number of Figures:6

Abstract

In this study we analyze 1024 free text digital records from pediatric patients who underwent CT scanning. The free text reports are from the digital records of patients who underwent CT scanning in a one-year period in 2004 at the Nagasaki University Medical Hospital in Japan. We use text mining algorithms to model the records. Each scan was evaluated by an expert in the field and classified as to whether the CT scan was necessary or not. A model was built that predicts this classification. The results show that models developed on raw text could contribute significantly to the physician's decision to order a CT scan. Practically this is important because radiation at levels ordinarily used for CT scanning may pose significant health risks especially to children and thus the modeling of unnecessary scanning may lead to less exposure to radiation.

Keywords: Text Mining, Radiology, Bag-of-words, Vectorization, Vector Space Model.

Introduction

In this study we analyse free text medical records using the "bag-of-words" model (also known as a vector space model). Free text poses numerous obstacles to computer analysis but also has great potential for knowledge discovery. The medical records we obtained for this study are clinicians' records from CT scans. We submitted them to text mining algorithms in order to see if a model could be built to distinguish those scans which were necessary from those which may have been unnecessary. These results are illustrative of a methodology that shows that computational text classification can be compared favourably with analysis by a human expert and specifically that models developed on raw text may contribute to the physician's decision to order a CT scan. Practically this is important because radiation at levels ordinarily used for CT scanning may pose significant health risks especially to children and modelling of unnecessary scanning may lead to less use of medical radiation. It is hoped that such a model may be used to develop stepwise procedures to curtail unwanted scans and exposure to radiation especially in children.

Outline of Paper

The paper is organized as follows. Section 2 –a review of related work. Section 3 – a discussion of the objectives of the research. Section 4 – a description the data structure, processing, and modelling as well as the evaluation procedures. Section 5 – a discussion of the results. Section 6 discusses limitations and extensions of the research and conclusions.

Literature Review

Text-mining is applied in various fields to extract useful and previously unknown information contained in databases and text. As early as the late 1950s, (Luhn 1957), and early 1960s (Maron and Kuhns 1960), studied document indexing. There are a variety of approaches when faced with a large corpus of free text. These range from viewing the text as a collection of words (the bag-of-words approach or vectorization approach) to models that incorporate the natural language structure of the text. One of the main problems of the bag-of-words representation is its loss of semantic relations; the meaning of word combinations is lost (Bekkerman and Alan 2003). Although the bag-of-words approach will strip text of any grammatical information the approach has met with significant success as will be outlined below specifically in the field of Medical Informatics. Table 1 is a presentation of work related to this paper that use the 'bag-of-words' approach to text mining. A discussion of this work follows.

Table 1: Listing of recent work in medical text mining where a bag of words approached was used.

[Insert Table 1 here]

In a pioneering study in 2005 (Pakhomov et al. 2005) attempted to identify patients with congestive heart failure using, as input, dictated clinical notes from the Mayo Clinic. The training of the classifier relied on notes that were manually categorized by human experts. The

researchers found that the Naïve Bayes classifier outperformed the Neural Network classifier on recall but not on accuracy.

Two years later (Pakhomov et al. 2007) reported on a comparison between an NLP approach and a vectorization approach (bag-of-words) where the goal was to identify heart failure through the examination of language contained in the electronic medical record (EMR). The NLP approach provided better sensitivity (81.6% versus 56% for the bag-of-words method) and nearly equivalent specificity (97.8% versus 96.0% for bag-of-words method). However the scores for positive predictive value (PPV) were higher for bag-of-words method (82.2% versus an NLP score of 49.3%). The model used was a Naïve Bayes algorithm.

In 2008 (Cohen 2008) participated in the i2b2 smoking status classification challenge task. The goal was to identify patient smoking status [smoker/non-smoker] from the free-text portion of a hospital discharge summary. The method identified "hot spots" within the free-text and then focused the bag-of-words approach on these hot spots. They report a micro-F of 97%.

In 2008, (Pakhomov et al.2008) used a bag-of-words approach "to process the text of physical examination sections of in-patient and out-patient clinical notes in order to identify whether the findings of structural, neurological, and vascular components of a foot examination revealed normal or abnormal findings or if they were not assessed". A support vector machine classifier obtained accuracy of 88% for the vascular component.

Also in 2008 (Pakhomov et al. 2008) compared a bag-of-words approach to a "bag-of-concepts" approach, where the bag-of-concepts were developed through Metamap. The objective was to predict patient responses on a standardized HRQOL assessment. The input into the model was physician reports contained in electronic medical records. The original feature set contained more than 10,000 elements and thus various feature selection procedures were compared. A

support vector machine (SVM) model was tested. The bag-of-concepts reportedly performed better than the bag-of-words model (for example negative agreement was 0.72 for bag-of-words and 0.78 for bag-of-concepts).

In 2009 (Kilicoglu et al. 2009) used a combination of words extracted from the title and abstract of a MEDLINE citation and metadata from the same citation to construct a vector of 2000 features in order to recognize rigorous applicable studies in the context of evidence-based medicine. In this classification problem they achieved a recall of 97.5% using a Naïve Bayes classifier. The precision for this classifier was 13.8%. Other classifiers (Polynomial SVM, Boosting, Stacking) did better on precision but not as well on recall.

These studies show that a bag-of-words approach to free-text has considerable potential to assist in decision making either in a clinical research or in a medical setting.

Methods

Introduction

The increase in the use of medical radiation, especially in diagnostic CT scanning has raised many concerns over the possible adverse effects of procedures conducted in the absence of any serious risk/benefit analysis, especially where these procedures are carried out on children. Overuse can lead to unnecessary risk of exposure to radiation and may also contribute to rising health care costs (Brenner 2001) (Roebuck 1999) (Frush 2003).

In a study done on the use of diagnostic imaging in emergency departments of hospitals in the USA between 1998 and 2007 (Korley, Pham, Kirsch 2010), the prevalence of CT or MRI scanning increased from only 6% in 1998 to 15% in visits to the emergency room with little or no corresponding change in the percentage of patients admitted to the hospital or the intensive

care unit. It was also noted that the chances of a CT or MRI scan occurring increased 3-fold from 1998 to 2007 for injury related conditions. Even so, the number of diagnoses of life-threatening conditions showed only a modest rise.

Originally, prior to our investigation, researchers at Nagasaki Hospital, using conventional methods, attempted to re-evaluate the efficiency of CT scanning in the diagnosis of acute appendicitis and for possible injuries after acute head trauma (Ghotbi 2005). As a result of that study a recommendation was made to the two departments studied. The recommendation was to employ guidelines which present a stepwise set of clinical diagnostic methods and tools. The intention of this recommendation was that CT scans be reserved for patients that may be expected to benefit from them. However, in other departments, due to the lack of such a stepwise approach to diagnosis, many unnecessary CT scans have been and continue to be undertaken (Ghotbi 2005), and sound clinical judgment has been postponed until there is a confirmation by a CT scan. This was the initial impetus for our current work.

The standard procedure adopted for requesting a scan at the Nagasaki Medical University Hospital as well as the domain expert classification is outlined in figure 1.

[Insert Figure 1 here]

Figure 1. Schematic diagram showing the standard procedure followed at Nagasaki Medical University Hospital and the expert classification on the necessity of a CT scan. We intend to develop tools to identify the features/ words that relate to unnecessary scans to curtail its overuse and thereby to overexposure to radiation, especially in children.

Outline of data handling and analysis

Figure 2 shows the procedures by which we processed and analysed the data we received from Nagasaki Medical University Hospital.

[Insert Figure 2 here]

Figure 2: The study design -including data sources, processing components, data flow, and evaluations.

Our hypothesis was that the free-text portions of clinical records would include some factors

that clinicians regarded as grounds for the CT requests. We translated the free-text portion of the records into English and pre-processed them, emerging with a dictionary of 922 features, which were then used as input to 7 different predictive models. The ultimate goal being prediction as to the necessity of the CT scan, we thus employed an expert in the medical field (a physician) to classify each scan in our dataset as necessary or unnecessary. This was our classification variable. The domain expert used two criteria to decide whether taking the CT scan was necessary or not: i) whether the clinical condition of the patient justified ordering a CT scan, according to standard clinical benchmarks (Hagendorf 2004) (Dunning 2004) (Stiell 2001) (Committee on Quality Improvement and American Academy of Pediatrics, 1999) ii) whether the result of the CT scan changed the workup/ management plan of the patient.

Using the classification variable we trained models for prediction and reserved 33% of the data for testing the models. The 7 models had overall accuracy of more than 72% and the accuracy extended as high as 96% for one of the models.

Processing of data in radiological records

Translation

We employed a native Japanese speaker to translate the records into English. We instructed the translator that the sentences produced in English need not adhere to strict grammatical rules but that no words should be dropped.

Data set and classification variable.

Our dataset was from the Nagasaki University Hospital Radiology Department's CT scanning database. It consisted of 1024 patient records. The individuals were children who received CT scans to aid in their clinical diagnosis. The following are the data extracted from the main Nagasaki Hospital database:

1. Exam Title (an anatomical description of the CT scan exposure area, such as head, chest, abdomen, etc)
2. ID number (a unique code for each patient)
3. Age
4. Sex
5. Department
6. In/Out patient status
7. Clinical information (as the reason to request a CT)
8. Findings (as part of CT report by radiologists)
9. Impression 1, 2, and 3 (as part of CT report by radiologists)
10. Result (as part of CT report by radiologists)

The analysis we subsequently performed seeks links between the free text in the physicians' notes (#7 from the list above) and a positive/negative outcome. The positive or negative outcome was determined according to the independent analysis by the physician of items 7 ~ 10, who referred to the standard clinical criteria for such decision-making. In this paper, a positive outcome indicated that the requested CT scan was deemed useful by the domain expert (the physician) in reaching a diagnosis/management of the patient. A negative outcome meant that the CT scan was considered not to have been useful in making a diagnosis by the domain expert.

Bag-of-words and Term-Document Matrix: Covariate Extraction.

"Since texts cannot be directly interpreted by a classifier or by a classifier-building algorithm, it is necessary to uniformly apply a transformation procedure to the text corpora in order to map

a text into a compact representation of its content" (Sebastiani 2002). The translated medical records were examined by breaking each patient record into its constituent words. We then removed the standard 124 stop words (i.e. 'a', 'able', 'about', 'above' etc.,) as well as common medical terms identified by a physician (i.e., 'abduction', 'advance', 'vessel') from the clinicians' notes. We further employed the well known Porter Stemming Algorithm (Van Rijsbergen 1980) which is a process for removing the more common morphological and inflexional endings from words. We used the resulting vocabulary of 922 covariates without any further restrictions. The 42 records that came out blank through this process were removed altogether from this analysis. Consequently, a 982x922 matrix of feature vs. record no. was constructed. The outcome/classification is dichotomous (positive vs. negative) and the covariates are derivatives of the words found in the clinical notes. For example, the column headers for columns 35 through 39 were: 'aseptic', 'asphyxiation', 'aspiration', 'asthma'. We used a standard method of weighting the words which gives consideration to the frequency at which a word occurs in a record and also the overall frequency that the word occurs within the entire corpus. This method is well known as the tf-idf formula (Manning and Shutze 1999) (Salton and McGill 1983). This allowed us to recode text data as numerical data and thus made it amenable to analysis with a variety of modelling procedures. Such a matrix is referred to as a term-document matrix.

Models Tested and Evaluation Criteria

Models Tested

Logistic regression is a statistical model that is used when the outcome is binary in nature. It relates the log odds of Pr(event) to a linear combination of predictor variables. It has been shown to be both fast and accurate for classification tasks (Lim et al. 2000).

Support Vector Machine (SVM) classifiers "use "kernel" functions to map the input space to a higher dimensional space where a maximal separating hyperplane is constructed" (Kilicoglu et

al. 2009). Linear and non-linear SVM classifiers have been successful for text classification. They are well suited for data with a very large number of input fields. Support Vector Machines have shown to have good performance on many types of classification problems including text categorization (Kwok 1998) (Thorsten 1998).

A neural network or artificial neural network (also referred to as a multilayer perceptron) is a model used to predict outcomes based on inputs. It consists of an interconnected group of artificial neurons. Neural networks are non-linear statistical modelling tools and can be used for supervised or non-supervised learning (Cohen and Hersh 2005) (Hastie 2001) . In our work we use them for supervised learning.

The acronym CHAID stands for Chi-squared Automatic Interaction Detector. It is one of the oldest tree classification methods and was originally proposed by (Kass 1980) (Rokach and Maimon 2008). CHAID will "build" non-binary trees (i.e., trees where more than two branches can attach to a single root or node). It may therefore create wider trees (Mckenzie 1993) (see figure 4). It supports both categorical and numeric output fields and input fields.

The CART algorithm stands for ‘classification and regression trees’ (Hong and Weiss 2001) (Breiman et al. 1984). It is a non-parametric technique that can handle categorical or numeric dependent variables and this distinguishes them from C4.5 trees. It produces binary splits at each node.

Quest stands for ‘Quick, Unbiased, Efficient Statistical Tree’ (Loh and Shih 1997). It is a binary split classification method. A major motivation in its development was to reduce processing time as compared with the CART algorithm when a large number of variables is involved. Quest trees along with C4.5 have shown to have some of the best error rate and speeds among decision trees. Quest only allows symbolic output fields.

C4.5 constructs a decision tree using the concept of information entropy developed by (Quinlan 1993). At each node C4.5 chooses one covariate that most effectively separates the records into those within one of the binary classification and those within the other binary classification. C4.5 only allows symbolic output fields. The C4.5 algorithm can support splits at each node that result in more than 2 subgroups for symbolic predictor fields (Kotsiantis 2007).

For all models we used default parameters. These are listed in Table 2.

Table 2: List of models used and parameters for these models

[Insert Table 2 here]

Evaluation Criteria

As noted above, we denoted as positive those scans determined to be necessary by the domain expert and as negative, those scans determined to be unnecessary. In the following, we shall refer to scans that are recommended by the model (correctly or incorrectly) as 'predicted-necessary' and to the scans that are not recommended by the model (correctly or incorrectly) as 'predicted-unnecessary'. Of the 982 records retained for evaluation, 622 (63%) were labelled as a "necessary scan" by the domain expert (a physician) and 360 (37%) were labelled as "unnecessary scans". Models were evaluated using the following criteria.

Let R = Recall/Specificity. Then R = TP/ (TP + FN), where: TP is the number of true positives and FN is the number of false negatives. Thus recall is the percentage of correctly predicted-necessary scans compared to the total number of actual necessary scans. Let P = precision. Then P =TP/ (TP + FP), where: TP is as above and FP is the number of false positives. Thus precision is the number of correctly predicted-necessary scans compared to the total number of predicted-necessary scans. Let S = specificity. Then S = TN/ (TN + FP), where: TN is the number of true negatives and FP is as above. Thus specificity is the number of correctly predicted-unnecessary scans compared to the actual number of unnecessary scans. Let NPV = negative predictive value.

Then NPV = TN/ (TN + FN). Thus negative predictive value is the number of correctly predicted-unnecessary scans compared to the total number of predicted-unnecessary scans. Let Acc = accuracy. Then Acc = C/T, where, C is the overall number of correct predictions and T is the total number of cases. Let F = F-score. Then F = 2*P*R/(P +R).

Results

All Models

As is seen in table 3, the neural network was the most successful on all measures except for recall (and here the difference is negligible). One reason for this may be that the number of nodes in the hidden layer was kept small and this reduces the likelihood of the hazard of overfit of the model to the data. Furthermore three of the decision trees (CART, Quest, Chaid) performed identically across all measures and that the C4.5 model performed slightly worse on specificity and NPV. Furthermore specificity was relatively low except for the neural network.

The fact that the recall rate (sensitivity) was generally high means that most of the models were sensitive enough to be able to identify most of the necessary scans (not to miss a necessary scan). The average recall across models excluding the poorly performing logistic regression model was 97% meaning that a necessary scan will be misidentified (False Negative Rate) as unnecessary in 3% of the cases. For comparison, (Pakhomov et al. 2005) reported recall of 86% using a neural network on positive samples but 95% using a Naïve Bayes classifier on positive samples. In later work (Pakhomov et al. 2007) reported sensitivity of 56%. (Kilicoglu 2009) reported a maximum sensitivity of 84.3%

The precision (or positive predictive value) of the models varied from 68%, for the C4.5 model to 98% for the neural network. Thus when a scan is predicted-necessary by the model it will

actually be necessary between 68 and 98 percent of the time depending on the model used. In particular, with a neural network, the chance of recommending an unnecessary scan is 2%. (Pakhomov et al. 2007) reported 82%. (Kilicoglu 2009) reported a maximum precision of 82.5%.

The negative predictive value (NPV) ranges from 43%, for logistic regression, to 93% for the neural network. NPV measures the percent of scans that are correctly predicted-unnecessary compared with the total number of predicted-unnecessary. In our study, NPV is 93% for the neural network model. Thus for the neural network, the chance of missing a necessary scan is 7% (1-NPV).

The specificity of the models are relatively low with the exception, again, of the neural network. The neural network does not suffer from this drawback as it has a specificity of 96%. (Pakhomov et al. 2007) reported 96% specificity.

The overall accuracy of the models ranged from 60% -for logistic regression, to 96% -for the neural network. The average accuracy for the decision trees is 71%. The F-score ranged from 69% to 97%. The neural network performed slightly better than what was reported by (Cohen 2008) using a weighted support vector machine and slightly worse than that of (Farkas et al. 2009) .(Pakhomov et al. 2007) reported accuracy of 65% for a Perceptron.

Table 3. Comparison of models based on various evaluation criteria.

[Insert Table 3 here]

Table 4. Comparison of models tested over evaluation measures. Shows best performing model, second best, and worst performing model.

[Insert table 4 here]

In table 4 the models are ranked according to their performance on the test partition. The neural

network was by far the most successful model. Support Vector Machines, and Naïve Bayes Classifiers have been reportedly strong in bag-of-words analysis but Pakhomov reported some success with neural networks in earlier work (see table 1). Three of the four decision trees performed almost identically on the data.

The Decision Trees

Although the decision trees (see figures 5 and 6) were not as successful in their predictive capacity, three of them were uniform in their results and also pointed to the same "key" words. These words were 'headache' and 'convulsion' as can be seen in figure 3. The Chaid tree (figure 4) is larger and includes more key words.

[Insert Figure 3]

Figure 3. Quest, C4.5, and CART Tree diagrams. '1' denotes a necessary scan, '2' denotes an unnecessary scan.

[Insert Figure 4]

Figure 4. Chaid Tree Diagram. '1' denotes a necessary scan and '2' denotes an unnecessary scan.

Discussion

Limitations and Extensions

Limitations

A limitation in this study was the fact that the free-text was translated from Japanese to English. However, this should not be a significant concern because machine translation, in the context of a bag-of-words approach, is a fast and acceptable substitute. A second limitation here was the number of records included in the study. As our feature set (the total number of word stems in the "dictionary") was 922 it would be desirable to have more records to analyse. A third weakness was the fact that only one domain expert annotated the training data (as necessary or unnecessary). It may be advisable in a future study to have 2 or more experts who are required to

obtain a consensus or near consensus opinion. Finally, as a validation set was not used and no replication was used the conclusions of the study may be considered corpus-specific.

Extensions

The negative predictive value (NPV) was 93% for the neural network. The higher this value, the less likely the model is to miss a necessary scan and this may arguably be the most important evaluation criteria. It may therefore be appropriate to employ a cost function within the model to further increase this value even at the expense of other evaluation criteria. Another important extension of this work would be to include the other data fields beyond the free text data field that was a part of the patient records (fields 3 thru 7 above). Additionally, as stated above, more records included in the study would also be valuable. In terms of the algorithms and methodology, it will be beneficial to test both boosting and bagging on a larger data set. Furthermore, recently there has been work on modifications of the tf-idf weighting methodology and it may be valuable to recode the weightings according to these algorithms (Reed et al. 2006). (Lan et al. 2005) compare 10 different term weighting systems and suggest that a modification of tf-idf called tf-rf may perform better. (Forman 2008) report that by replacing IDF with Bi-Normal Separation (BNS) both accuracy and F-measure are improved for Support Vector Machine.

Conclusion

This paper addresses a problem of computational text classification in medical informatics, that of the question of whether machine learning algorithms can successfully replicate the manual human annotation of radiological records. In this context we have conducted experiments for the

provision of information that contributes uniquely to the diagnostic process in this domain. Our results show that using the analytical approach we describe, reliable medical assessments can be made, which leads us to propose adoption of this methodology for other areas of diagnosis

In a series of text mining experiments on radiological records for CT scanning of children, we have shown that neural networks and to a lesser extent decision trees can achieve a reasonably high level of accuracy, precision, specificity, and sensitivity in classification of free text medical records. In particular, the neural network method achieved 96% accuracy, 97.6% precision, 95.5% specificity, and 96.1% sensitivity. This shows that medical physicians may be able to consider some pre-test factors such as the predicted benefit of the test, before requesting a CT scan for children and thus they may avoid unnecessary CT scans. This not only reduces the dose of radiation received by children from unnecessary CT scans, but also reduces high costs of diagnostic procedures. Furthermore, our work suggests that similar methodologies could be applied by physicians to support decision making in other diagnostic procedures.

Acknowledgments

The authors wish to express their sincere gratitude to Professor Monte Cassim who has been a tremendous inspiration for all of us here at Discovery Laboratory, especially, into text mining clinical data. Nagasaki University Hospital clinical staff members are acknowledged for permission to use their data in this study.

References

Bekkerman, R., and J. Allen, 2003. "Using Bigrams in Text Categorization". Department of Computer Science, University of Massachusetts, Amherst: CIIR Technical Report IR-408.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. "Classification and Regression Trees." Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brenner D. J, C. D. Ellison, E. J. Hall, W. E. Berdon. 2001. "Estimated Risks of Radiation Induced Fatal Cancer from Pediatric CT." American Journal of Roentgenology, 176: 289-296.

Cohen, A. M. 2008. "Five-way Smoking Status Classification Using Text Hot-Spot Identification and Error-correcting Output Codes" 2008. Journal of the American Medical Informatics Association, Volume 15, Issue 1, January-February, Pages 32-35, ISSN 1067-5027, DOI: 10.1197/jamia.M2434. Available at: http://www.sciencedirect.com/science/article/B7CPS-4RB0P7D-8/2/c16c9727d6d18ed88bc0aec2e7e9d2a4

Cohen A. and W. Hersh, 2005. "A survey of current work in biomedical text mining." Briefings in Bioinformatics,6(1):57-71.

Committee on Quality Improvement and American Academy of Pediatrics, Commission on Clinical Policies and Research, and American Academy of Family Physicians. American Academy of Pediatrics: "The management of minor closed head injury in children." 1999. Pediatrics ;104:1407-1415.

Dunning J, J. P. Daly, R. Malhotra, P. Stratford-Smith1, J-P Lomas, F. Lecky, J. Batchelor, K. Mackway-Jones. 2004. "The Implications of NICE Guidelines on the Management of Children Presenting with Head Injury." Arch Dis Child. 2004;89:763-767.

Farkas R., G. Szarvas, I. Hegedus, A. Almasi, V. Vincze, R. Ormandi, R. Busa-Fekete. 2009. "Semi-automated Construction of Decision Rules to Predict Morbidities from Clinical Texts" Journal of the American Medical Informatics Association, Volume 16, Issue 4, July-August 2009, Pages 601-605, ISSN 1067-5027, DOI: 10.1197/jamia.M3097. Available at: http://www.sciencedirect.com/science/article/B7CPS-4WNB0TY-X/2/b021a9152bea08e651597870aef1dfc0

Forman G., 2008. "BNS Feature Scaling: An Improved Representation Over tf-idf for SVM Text Classification". In Proceeding of the 17th ACM Conference on information and Knowledge Management (Napa Valley, California, USA, October 26 - 30, 2008). CIKM '08. ACM, New York, NY, 263-270. DOI= http://doi.acm.org/10.1145/1458082.1458119

Frush D. P., L. F. Donnelly, N. S. Rosen. 2003. "Computed Tomography and Radiation Risks: What Pediatric Health Care Providers Should Know." Pediatrics, 112: 951-957,.

Ghotbi N., M. Morishita, A. Ohtsuru and S. Yamashita, 2005. "Evidence-based Guidelines Needed on the Use of CT Scanning in Japan." Japan Medical Association Journal (JMAJ) Vol. 48, No. 9 September 2005.

Hagendorf B. A, J. R. Clarke, R. S. Burd, 2004. "The Optimal Initial Management of Children with Suspected Appendicitis: a Decision Analysis." J Ped Surg. 2004;39:880-885.

Hastie T., R. Tibshirani, and J. Friedman. 2001. "The Elements of Statistical Learning." Springer series in statistics. Springer, New York.

Hong S. J., S. M. Weiss, 2001. "Advances in predictive models for data mining." Pattern Recognition Letters, Volume 22, Issue 1, Machine Learning and Data Mining in Pattern Recognition, January 2001, Pages 55-61, ISSN 0167-8655, DOI: 10.1016/S0167-8655(00)00099-4. Available at: http://www.sciencedirect.com/science/article/B6V15-41PP7KX-W/2/6f1147aebf964237dab9bad9c9b2d74e

Kass G. V.. 1980. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." Journal of Applied Statistics, Vol. 29, No. 2, pp. 119-127.

Kilicoglu, Halil, Dina Demner-Fushman, Thomas C. Rindflesch Nancy L. Wilczynski and R. Brian Haynes, 2009. "Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence", Journal of the American Medical Informatics Association, Volume 16, Issue 1, January-February, Pages 25-31, ISSN 1067-5027, DOI: 10.1197/jamia.M2996. Available at: http://www.sciencedirect.com/science/article/B7CPS-4V42CR0-7/2/26d690586d0163237d59f114a7fe49ac

Korley F. K., J. C. Pham, T. D. Kirsch, 2010. "Use of Advanced Radiology During Visits to US Emergency Departments for Injury-Related Conditions, 1998-2007." JAMA, October 6; 304: 1465 - 1471.

Kotsiantis S. B., 2007. "Supervised Machine Learning: A Review of Classification Techniques." Informatica 31: 2007 249-268.

Kwok J. T., 1998. "Automatic Text Categorization Using Support Vector Machines." Proceeding of International Conference On Neural Information Processing, October, pp. 347-351

Lan M., C. Tan, H. Low, and S. Sung, 2005. "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines." In Special interest Tracks and Posters of the 14th international Conference on World Wide Web, Chiba, Japan, May 10 - 14, 2005. DOI= http://doi.acm.org/10.1145/1062745.1062854

Lim T., W. Loh, and Y. Shih, 2000. "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms." Mach. Learn. 40, 3 (Sep. 2000), 203-228. DOI= http://dx.doi.org/10.1023/A:1007608224229

Loh W. Y. and Y. S. Shih, 1997. "Split Selection Methods for Classification Trees." Statistica Sinica, vol. 7, 815-840.

Luhn H. P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." IBM Journal of Research and Development 1(4):390.

Manning C, H. Shutze. 1999. "Foundations of Statistical Natural Language Processing." Cambridge, Mass: MIT Press.

Maron, M., J. Kuhns, 1960. "On Relevance, Probabilistic Indexing, and Information Retrieval." Journal of the ACM. 7.

McKenzie D. P., P. D. McGorry, C. S. Wallace, L. H. Low, D. L. Copolov & B. S. Singh. 1993. "Constructing a Minimal Diagnostic Decision Tree. Methods of Information in Medicine." Vol. 32, pp. 161-166.

Pakhomov S, N. Shah, P. Hanson, S. Balasubramaniam, S. Smith. 2008. "Automatic Quality of Life Prediction Using Electronic Medical Records." AMIA Annual Symposium Proc.; 545–549.

Pakhomov S. V., J. Buntrock, C. G. Chute. 2005. "Prospective Recruitment of Patients with Congestive Heart Failure Using an Ad-hoc Binary Classifier." Journal of Biomedical Informatics, Volume 38, Issue 2, April, Pages 145-153, ISSN 1532-0464, DOI: 10.1016/j.jbi.2004.11.016. Available at: http://www.sciencedirect.com/science/article/B6WHD-4F3FBGS-1/2/edbeea99b4978c3f5f90314e8fc703da

Pakhomov S., P. L. Hanson, S. S. Bjornsen, S. A. Smith. 2008. "Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning." Journal of American Medical Informatics Association. Mar–Apr; 15(2): 198–202. doi: 10.1197/jamia.M2585.

Pakhomov S. Weston, S. Jacobson, C. Chute, R. Meverden. 2007. "Electronic Medical Records for Clinical Research: Application to the Identification of Heart failure." American Journal of Managed Care, 2007-06-vol13-n6-Pt1, Jun07-2488p281-288

Quinlan J. R., 1993. "C4.5: Programs for Machine Learning." Morgan Kaufmann Publishers.

Reed J. W., Y. Jiao, T.E. Potok, B. A. Klump, M. T. Elmore, A. R. Hurson. 2006. "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams." Machine Learning and Applications, 2006. ICMLA apos;06. 5th International Conference on Volume , Issue , Dec. 2006 Page(s):258 – 263

Roebuck D. J. 1999. "Risk and Benefit in Pediatric Radiology." Pediatr Radiol, 29: 637-640.

Rokach L. and O. Maimon, 2008. Data Mining with Decision Trees: Theory and Applications, World Scientific, New York.

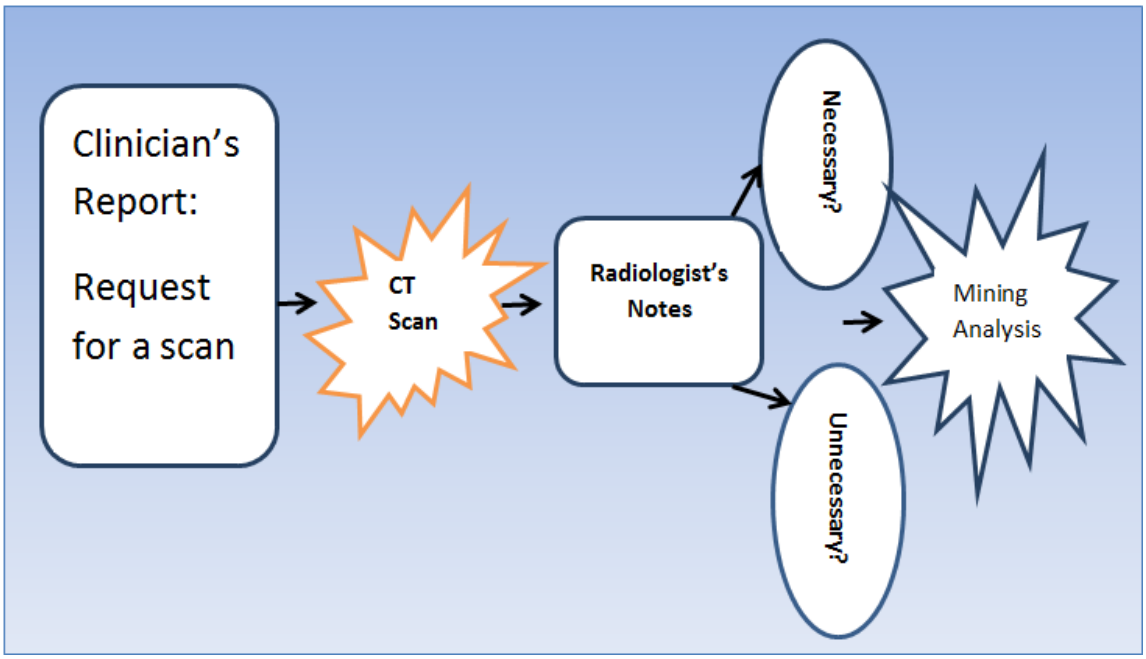
Salton G. and M. J. McGill, 1983. "Introduction to Modern Information Retrieval." McGraw-Hill. ISBN 0070544840.

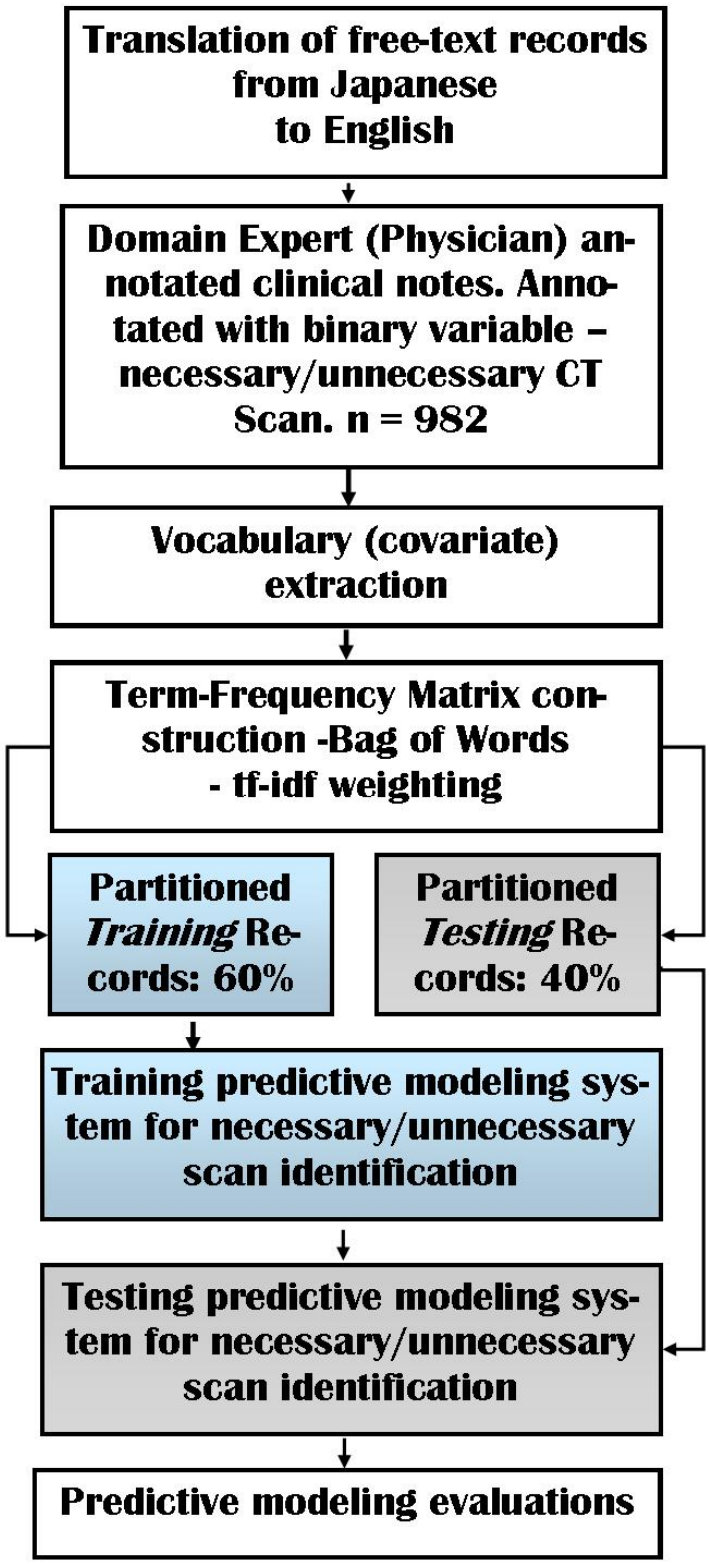
Sebastiani F., K. Seymore, A. McCallum, & R. Rosenfeld,. 2002. "Machine Learning in Automated Text Categorization." ACM Computing Surveys, 34(1), 1-47. doi:10.1145/505282.505283.

Stiell Ian G., George A Wells, Katherine Vandemheen, Catherine Clement, Howard Lesiuk, Andreas Laupacis, R Douglas McKnight, Richard Verbeek, Robert Brison, Daniel Cass, Mary A Eisenhauer, Gary H Greenberg, James Worthington, 2001. "The Canadian CT Head Rule for patients with minor head injury." Lancet. 2001;357:1391-1396.

Thorsten J. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." Proceedings of the 10th European Conference on Machine Learning, p.137-142, April 21-23.

Van Rijsbergen C.J., S.E. Robertson, M.F. Porter, 1980. "New Models in Probabilistic Information Retrieval." London: British Library. British Library Research and Development Report, no. 5587


Clinician's Report:
Request for a scan
CT Scan
Radiologist's Notes
Necessary?
Unnecessary?
Mining Analysis


Translation of free-text records from Japanese to English
Domain Expert (Physician) annotated clinical notes. Annotated with binary variable – necessary/unnecessary CT Scan. n = 982
Vocabulary (covariate) extraction
Term-Frequency Matrix construction -Bag of Words - tf-idf weighting
Partitioned *Training* Records: 60%
Partitioned *Testing* Records: 40%
Training predictive modeling system for necessary/unnecessary scan identification
Testing predictive modeling system for necessary/unnecessary scan identification
Predictive modeling evaluations

Quest Tree	
headache<= 0.385 [Mode: 1]	
	convulsion <= 0.295 [Mode: 1] => 1
	convulsion> 0.295 [Mode: 2] => 2
headache > 0.385 [Mode: 2] => 2	

C4.5 Tree	
convulsion <= 0 [Mode: 1]	
	headache <= 0.600 [Mode: 1] => 1
	headache > 0.600 [Mode: 2] => 2
convulsion > 0 [Mode: 2] => 2	

CART Tree	
headache <= 0.300 [Mode: 1]	
	convulsion <= 0.750 [Mode: 1] => 1
	convulsion > 0.750 [Mode: 2] => 2
headache > 0.300 [Mode: 2] => 2	

Chaid Tree
headache-374 <= 0 [Mode: 1] (719)
convulsion-182 <= 0 [Mode: 1] (696)
sweling-826 <= 0 [Mode: 1] (688)
tympanitistube-874 <= 0 [Mode: 1] (653)
appendictis-49 <= 0 [Mode: 1] => 1 (639; 0.653)
appendictis-49 > 0 [Mode: 1] => 1 (14; 1.0)
tympanitistube-874 > 0 [Mode: 1] => 1 (35; 0.914)
sweling-826 > 0 [Mode: 2] => 2 (8; 1.0)
convulsion-182 > 0 [Mode: 2] => 2 (23; 0.739)
headache-374 > 0 [Mode: 2] => 2 (63; 0.667)

Year	Work	Classification Variable	Primary Model	Other models tested	Maximum Performance Evaluation Reported
2004	Pakhomov et al.[4]	Heart Failure or Not	Naïve Bayes	Perceptron (Neural Network)	recall: 95%
2007	Pakhomov et al.[5]	Heart Failure or Not	Naïve Bayes	Natural Language Processing	specificity: 96%
2007	Cohen[6]	patient smoking status	Weighted support vector machines		"micro" F: 90.00%
2008	Hui Cao et al.[7]	Records similar or not	Similarity metric (Metric-DFMP)	Metric-D, Metric-F, Metric-M, and Metric-P.	Correlation between 60% and 69%
2008	Pakhomov et al.[9]	normal/abnormal/non-assessed	SVM		accuracy 88%
2008	Pakhomov et al.[10]	Health related quality of life (HRQOL)	NLP and SVM		"positive agreement": 78%
2009	Kilicoglu et al.[11]	Topic relevance	Ensemble (Naïve Bayes, support vector machine (SVM), and boosting)		recall: 91.4%

Classifier Parameters

Classifier	Parameters used
Logistic Regression	Multinomial Regression, constant was included in the equation. Stepwise enter method was used where all variables are included in logistic regression, whether they significant or insignificant. Model type: main effects only.
SVM	Stopping criteria: .001, Regularization parameter: 10, Regression precision: 0.1. Kernel type: Radial Basis Function. RBF gamma: 0.1
CART Tree	5 levels below root node. Minimum change in impurity: 0.0001. Tree pruning was used.
C4.5	Pruning severity: 75%, Minimum records per child branch: 2, global pruning.
Quest	Maximum surrogates: 5, Alpha for splitting: 0.05, tree pruning: yes.
Neural Network	Input layer: 921 neurons, Hidden layer: 1 with 15 neurons, Output layer: 1.

Model	accuracy	F-score	Recall/Sensitivity	Precision	Specificity/True Negative Rate	NPV
CART Tree	0.719	0.82	0.969	0.71	0.239	0.8
Quest	0.719	0.82	0.969	0.71	0.239	0.8
Chaid	0.719	0.82	0.969	0.71	0.239	0.8
C4.5	0.679	0.799	0.969	0.679	0.119	0.667
Neural Network	0.959	0.969	0.961	0.976	0.955	0.928
Logistic Regression	0.602	0.685	0.659	0.714	0.493	0.429
Support Vector Machine	0.704	0.782	0.806	0.759	0.507	0.576

	Model Rankings Per Evaluation Measure		
Evaluation Measures	Best	Second	Worst
Accuracy	neural network	CART, Quest, Chaid	logistic regression
F-Score	neural network	CART, Quest, Chaid	logistic regression
Recall/Specificity	all decision trees	neural network	logistic regression
Precision	neural network	logistic regression	C4.5
Specificity	neural network	SVM	C4.5
NPV	neural network	CART, Quest, Chaid	logistic regression