Classification of Clinical Conditions: A Case Study on Prediction of Obesity and Its Co-morbidities

Archana Bhattarai, Vasile Rus and Dipankar Dasgupta

Department of Computer Science, The University of Memphis,
209 Dunn Hall
Memphis, TN 38152-3240, USA
{abhattar, vrus, dasgupta}@memphis.edu

Abstract. We investigate a multiclass, multilabel classification problem in medical domain in the context of prediction of obesity and its co-morbidities. Challenges of the problem not only lie in the issues of statistical learning such as high dimensionality, interdependence between multiple classes but also in the characteristics of the data itself. In particular, narrative medical reports are predominantly written in free text natural language which confronts the problem of predominant synonymy, hyponymy, negation and temporality. Our work explores the comparative evaluation of both traditional statistical learning based approach and information extraction based approach for the development of predictive computational models. In addition, we propose a scalable framework which combines both the statistical and extraction based methods with appropriate feature representation/selection strategy. The framework leads to reliable results in making correct classification. The framework was designed to participate in the second i2b2 Obesity Challenge.

Keywords: Text classification, Information Extraction, natural language processing

1 Introduction

Medical informatics is chiefly inductive and information intensive science where observation and analysis of comprehensive clinical data can lead to complex and powerful evidence based decision support systems. One of the primary goals of these automated systems is to make information more accessible, representative and meticulous in a quick span[4]. Furthermore, they have gained increased importance in the recent years as it can even outperform a human expert in some cases in diagnosing diseases as the process is highly subjective and fundamentally depends on the experiences of the assessor and his/her interpretation on the information[4]. Conversely, most medical institutions are still keeping a large amount of medical data in narrative form resulting in huge volume of potential information with limited or no utility and accessibility. An effort to exploit this data poses multiple challenges as it involves processing free text data with the presence of acronyms, synonyms, negation

Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 183-194

Received 11/11/08
Accepted 11/12/08
Final version 06/02/09

and dependence on temporality. Thus, in this work we try to explore different challenges that arise while trying to identify obese patients and the co-morbidities exhibited by them based on the narrative patient record.

The work was done specifically to participate in the "Obesity Challenge (A Shared-Task on Obesity): Who's obese and what co-morbidities do they (definitely/likely) have?" under Second i2b2 Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data organized by "Informatics for Integrating Biology and the Bedside, i2b2, a National Center for Biomedical Computing"[1].

The problem of identifying obese patients and the co-morbidities exhibited by them can be intuitively modeled as a document classification problem. Traditional approach of document classification task uses keyword based document representation (for example: bag of words representation) along with some statistical techniques to classify relevant and irrelevant documents[4]. These statistical techniques automatically identify useful keywords based on large training corpus. These methods have gained popularity because of ease and full automation. However, these methods exhibit serious limitations on deep semantics representation[5]. Some of the limitations of traditional classification methods in the context of medical reports are as follows:

Synonymy/Hyponymy/Acronym Consideration: It is a common practice to represent the same concept in different forms or words in medical domain. The use of acronyms is also highly predominant in the area. This factor of presence of synonyms/hyponyms and acronyms cannot be captured with word based classification system as traditional classification systems primarily performs keyword based classification without consideration of the context.

Negation handling: Medical documents contain a prominent amount of negative qualifiers. These qualifiers exhibit their significance only in local contexts. For example, in the sentence "the patient does not have diabetes", the negative qualifier "not" should be associated with diabetes. This characteristic cannot be captured fully in word or even n-gram feature based models.

Temporality: Statistical classification algorithms cannot capture the time varying components significantly present in medical reports. For example, a co-morbidity may be present in the past, but not in the present. Identification of such aspect demands complex semantic analysis.

Experiencer: The reference of a keyword, a phrase, sentence or the paragraph may be subjective to a different person and not the patient. For example, the sentence "FAMILY HISTORY: No family history of kidney disease or heart disease" talks about the patients family and not specifically about the patient. Such semantic nformation is not captured in statistical classification algorithms.

Thus we propose to apply the concept of Information Extraction based classification to address some of the problems explained above. Our work mainly focuses on the automated identification of synonymous/hyponymous words and acronyms. The method also learns the interdependence of the co-morbidities to exploit the property of high interdependence of co-morbidities. Finally, we apply existing negation handling algorithm, NegEx[2] in our work. We have not included

the temporality and experiencer problem in our work. However, the results are promising enough without their consideration.

2 Related Work

The task of medical document classification has been successfully applied to several real world medical diagnosis problems [4]. Most of the medical classification task apply statistical machine learning algorithms such as Naïve Bayes' classification, kernel based algorithms (Support vector machines (SVM), rules developing algorithms (decision trees) for different problem diagnosis [4]. Applications of pattern recognition in medical domain have also used Artificial Neural Network (ANN) [6]. However, very little has been done related to deep semantic analysis. Negation handling and word sense disambiguation is relatively more explored area compared to automated synonym/acronym extraction. Several works have been done in the recent years in negation handling. Chapman et al [2] developed a regular expression based negation determination algorithm. It defines a wide range of negation phrases that either appears before or after a finding in medical domain. It also defines a window size 'n' within which if the negation word occurs, the finding is considered to be negated. Another algorithm developed by Aronow et al [3] exploits syntactic processing techniques to identify noun phrases or conjunctive phrases to determine negation scope. Machine learning based algorithms have also found their application in identifying negative patterns in the text. Averbuch et al [7] developed an information gain based selection algorithm to automatically learn negative patterns from the text. Goryachev et al[8] did a comparative analysis on negative algorithms and developed a system with modified Negex and Aronow algorithm. They have also implemented Naïve Bayes' and Support Vector Machine based algorithm to automatically learn negation detection process using a set of manually annotated discharge summaries. Based on the observation on related works and simplicity, we implemented a simplified form of NegeX[2] algorithm to detect negated co-morbidities in our work.

For synonym/acronym extraction, thesauri based methods have been found to be useful in query expansion for information retrieval [9]. Domain independent thesauri such as WordNet [10] has proven to be helpful in a generalized information retrieval scenario[11]. However, such thesauri give very poor or no coverage for highly domain specific hyponym extraction scenario such as ours. Manual construction of such domain specific thesauri is expensive. As a result, automatic domain specific synonyms/hyponyms extraction has been started without being successful for substantial accuracy[12][13][14]. Term variation based hyponym detection[15] and distributional similarity[16] based synonym extraction methods have also been explored. In the term variation based method, variations of term such as "mouth cancer" and "cancer of mouth" [17] is studied and analyzed. In the distributional similarity based method, terms occurring in the proximity of known terms are taken to be similar which has been shown to increase recall at the cost of precision[16]. McCrae et al[17] used automated regular expression based patterns starting from few seed patterns to discover more patterns with a heuristic search method. All the above

mentioned methods still do not capture all the problem scenarios. For example, in our case, co-morbidity is mentioned in one form in one report and in another form in another which confronts the challenge of detecting synonyms from unassociated data. Our work is most closely related to the work in Riloff et al[5] which explains different information extraction based algorithms for high precision text classification. The idea is to automatically build domain specific dictionary from the given training corpus which is then used for information extraction task. The method is fully scalable and portable to any domain. On the negative side, this method is good for high precision results only with a compromise on recall.

3 Methods

The identification of obesity and its co-morbidities is a multiclass, multilabel classification problem where each patient may have multiple diseases and each disease can be marked with any of the labels such as Y for "Yes" meaning the patient has the co-morbidity, N for "No" meaning the patient does not have the co-morbidity, "Q" for Questionable meaning questionable whether the patient has the co-morbidity or "U" for unmentioned meaning the co-morbidity is not mentioned in the record. The judgments are provided in two forms; "textual judgments" and "intuitive judgments". Textual judgments are strictly based on text and intuitive judgments are based on implicit information in the narrative text.

The basic idea in our work (as explained in section 1) is to combine the best parts of traditional statistical learning methods and information extraction based methods. Traditional learning methods exhibit high recall with relatively good precision whereas extraction based methods exhibit high (near to perfect) precision. We evaluate various machine learning algorithms to obtain best classification result for the problem domain. We then apply extraction based method to refine the results obtained from statistical method to obtain better precision retaining high recall. In the process, we address specific challenges posed by both the statistical classification method and extraction based method.

Statistical machine learning based methods exhibit promising results in most of the general cases. However, the methods cannot perform to its best when the data exhibits a very high dimension. Moreover, these methods also cannot give accurate results when there is a high interdependence between the classes of a multiclass problem. We discuss these problems and our approach of solution in the following sub-sections. Similarly information extraction based methods also encompass challenges such as representative entity extraction, synonyms/hyponyms identification, negation handling etc which are discussed in the following sub-sections.

3.1 Multiclass Multilabel Classification

Our problem involves both the multilabel classification and multinomial or multiclass classification. Multiclass classification is a classification problem where a document can belong to one of several classes (more than two classes). The classes in multiclass

classification are mutually exclusive. Multilabel classification is defined as a classification problem where a document can belong to several classes simultaneously or to a single class or to none of the classes [18]. Most of the multilabel classification algorithms consider that the classes are independent of each other. With this assumption, the classification problem turns down to multiple binary classification problems. However, this generalization in our case is very expensive as all the co-morbidities are highly related to each other and the presence of one helps to induce the other.

3.2 Information Extraction

Information extraction (IE) is a type of information retrieval which extracts structured information, i.e. categorized and contextually and semantically well-defined data from a certain domain, from unstructured machine-readable documents automatically [19]. A complete translation of the semantics of a document requires an in-depth natural language processing which is computationally expensive. However, information extraction is a more focused and well-defined task. The advantage of information extraction is that the document that is not relevant to the context can be ignored thus reducing the computational cost effectively.

In our work, we tried to extract representative medical terms from the report that could be used as important keywords to characterize the report. We specifically extracted four medical concepts from the report namely; diseases or syndromes, sign or symptoms, body parts and clinical drugs. To extract the medical concepts, we used MetaMap Transfer (MMTx) [20]. This software processes text through a series of modules. First it is parsed into components including sentences, paragraphs, phrases, lexical elements and tokens. Variants are generated from the resulting phrases. Candidate concepts from the UMLS Metathesaurus[21] are retrieved and evaluated against the phrases. The best of the candidates are organized into a final mapping in such a way as to best cover the text.

3.3 Finding Synonyms/Acronyms and Class Interdependence: Riloff Metric

Medical people often have different word choices for the same concept. This semantic relationship can be exploited using tools such as WordNet in most cases when the problem is not domain dependent. However, discovering semantic relationship by nature is an open ended problem and highly domain specific. It is often very expensive or impossible to create a comprehensive resource by hand. Thus a corpus based discovery methods is essential to improve the coverage.

The basic intuition in identifying synonyms/hyponyms here is to find phrases which probably convey the same information. We use pre-classified training corpus to identify such phrases first. We then use the discovered synonymous/hyponymous phrases to make better predictions in test data. For this, the first task is to classify some words/phrases into major semantic categories. We used methods explained in section 3.2 to extract such semantic categories in the report. For the synonym identification task, we extracted all the disease/syndrome names mentioned in the

report. Then for each co-morbidity, we calculate a riloff value defined in equation 1 for each disease/syndrome which represents the similarity of that disease/syndrome to the co-morbidity. For example, a rilloff value of 0.3820 indicates the degree of similarity of the co-morbidity "chf" with the syndrome "cardiomyopathy".

Riloff value is defined by the equation

$$Riloff\ value = \frac{R}{I} \log R \quad \text{..... (i)}$$

Here,
R is the number of occurrence of given syndrome when given co-morbidity is present
I is the number of occurrence of given syndrome when the corresponding co-morbidity is not present.

The same concept can also be used to study the interdependence of co-morbidities. High Riloff value of co-morbidity for another indicates that the co-morbidity is causing the former co-morbidity to occur.

3.4 Negation Handling

Prediction in medical domain cannot be accurate without consideration of negation words. Negative qualifier assigned to a medical condition may indicate the absence of the condition, so the ability to reliably identify the negation status of medical concepts affects the quality of results produced by the classification system. Let us consider a simple example sentence "the patient does not have asthma". In this sentence, if the word 'not' is not given a special attention and just considered as just another feature, the semantics of the sentence could be the exact opposite. We use simplified version of NegEx to handle negation effect in our system. We use the negation module in extraction based approach. Our hypothesis here is that if one of the text fragments is negated, the concept is reversed, but if both are negated the decision is retained (double-negation), and so forth.

4 System Framework

The framework broadly consists of traditional supervised classification system and the extraction based classification system. In the supervised classification method, initially, the corpus is preprocessed to extract important features. Preprocessing includes tokenizing, stemming, stopwords removal etc. The features for classification are represented as unigram words and bigram phrases for baseline classification. Document frequency based feature selection strategy is also applied to select important features from the feature set. Document frequency is defined as the count of number of documents in which given feature occurs. The feature weighting scheme used is the tf-idf value. The weight w for each feature in a document is calculated using the formula:

$$weight = tf \times idf = tf \times \log(\frac{N}{df}) \quad \text{.....(ii)}$$

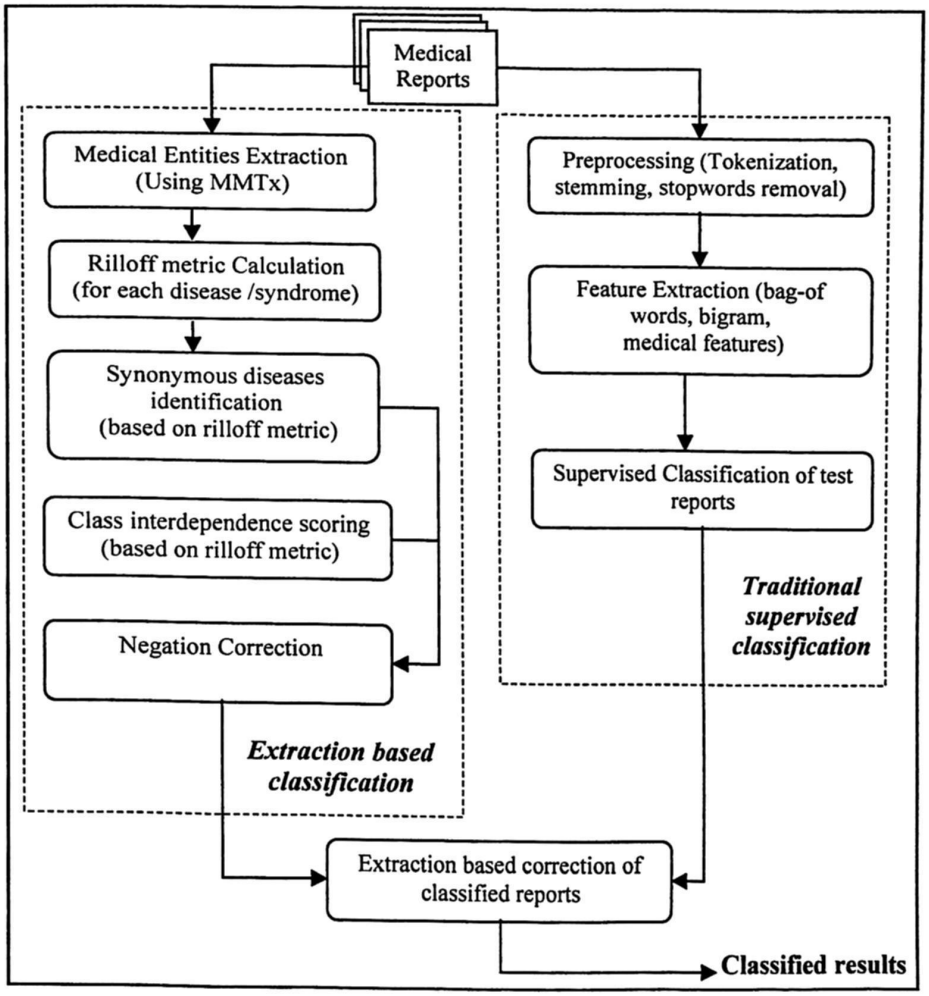

Figure 1. System framework for a combination of both statistical and extraction based classification of medical reports

where *tf* is the term frequency(number of occurrences of the feature in the document) of the feature in the corresponding document, N is the total corpus size and *idf* is the document frequency of the term. For more advanced classification, medical phrases such as disease/syndromes, sign/symptoms, body parts and clinical drugs are also used as features. The same *tf-idf* feature weighting scheme is used for these features too. We used Java based Weka[22] API to implement and evaluated various machine learning algorithms in the work.

For the extraction based classification, initially, medical features are extracted using MetaMap MMTx. The detailed process is explained in section 3.2. Then for each identified disease/syndrome, Riloff value is calculated. The Riloff value is normalized with the corpus size to get uniform Riloff value. Based on Riloff value, synonymous

terms are extracted. For this work, we have considered all the disease/syndromes with a Riloff value greater than 0.3 to be synonymous syndromes. After extracting all synonymous terms, the sentences containing those words are extracted from the reports. These sentences are then checked for negation terms. If the term is negated, it is ignored. If not, the report is considered to have the co-morbidity. The positive cases identified in this process are then used to refine the results obtained from the traditional classification approach.

5 Experimental Observations and Results

We summarize our observations and results in this section. For the evaluation of the supervised algorithms, extraction based synonym/hyponym identification and negation handling; we used 611 narrative medical reports for training the system and 119 reports for testing. The initial unique feature set size was 185527. Features with document frequency greater than 9 were only selected for classification purpose. The final feature set size was 6650.

5.1 Dataset

The dataset used is the de-identified discharge summaries of patients obtained from different healthcare organizations for the obesity challenge. Each document is marked as present, absent, questionable or unmentioned with respect to every co-morbidity and obesity. For each document, both the textual judgments (what the text explicitly states) and intuitive judgments (what the text implies) are provided. Altogether, there are sixteen co-morbidities namely; Obesity, Diabetes mellitus (DM), Hypercholesterolemia, Hypertriglyceridemia, Hypertension (HTN), Atherosclerotic CV disease (CAD), Heart failure (CHF) , Peripheral vascular disease (PVD), Venous insufficiency , Osteoarthritis (OA), Obstructive sleep apnea (OSA), Asthma, GERD, Gallstones / Cholecystectomy, Depression and Gout. There are reports in the dataset.

5.2 Synonym/hyponym Extraction

Table 1 shows some of the synonymous/hyponymous words extracted using the Riloff metric from the medical reports. The Riloff value indicates the degree of similarity of a co-morbidity with the synonym set. The Riloff value for each co-morbidity has different scale as this value depends on the corpus size which is relevant for the co-morbidity.

5.3 Classification Results (Accuracy)

Here, we compare the performance of Naïve Bayes' classification, Support Vector Machine (SVM) classification, J48 decision tree based classification and our system which incorporates the combination of J48 and extraction based classification.

Table 1. Extraction of hyponyms for each co-morbidity based on Riloff metric

Co-morbidity	Synonym set	Riloff val
gout	Abdominal aortic aneurysm	0.44
	Chronic obstructive pulmonary disease	0.61
Diabetes	Nph	0.42
hypercholesterolemia	Hyperlipidemia	0.30
	Elevated cholesterol	0.34
obesity	Fibromyalgia	0.31
	Obese	0.82
depression	chronic pain	0.51
	Migraines	0.45
gerd	gastroesophageal reflux	1.01
	Fibromyalgia	0.33
pvd	peripheral vascular disease	0.67
	Vascular disease	0.66
asthma	Asthma flare	0.8
	Tracheobronchitis	0.48
cad	coronary artery disease	0.49
chf	Congestive heart failure	0.49
	Heart failure	0.79
	Ischemic cardiomyopathy	0.40
osa	obstructive sleep apnea	0.81
	Sleep annea	1.01
	Pulmonary hypertension	0.59
oa	Djd	0.922
	Arthritis	0.67
	Fibromyalgia	0.59
	Degenerative joint disease	0.69
	Osteoarthritis	1.38
Hypertension	Htn	0.54

The graph in figure 2 summarizes the performance of the above mentioned algorithms. Naïve Bayes' algorithm does not show good result in predicting the co-morbidities although it attains accuracy over 90% in the case of hyperglyceridemia. Support vector machine performs relatively better than Naïve Bayes'. However, J48 decision tree performs the best in predicting obesity and its co-morbidities. J48 algorithm has exhibited accuracy over 90% for almost all the co-morbidities, some being nearly perfect. The refinement of the J48 results with the extraction based results has given even better accuracy of although not of significant value.

Similarly, figure 3 below shows the performance of different algorithms on the intuitive judgment. The overall accuracy of all the algorithms is worse than for the textual judgment. Among all the algorithms, Naïve Bayes' performed the worst with accuracy less than 70% for all the co-morbidities.

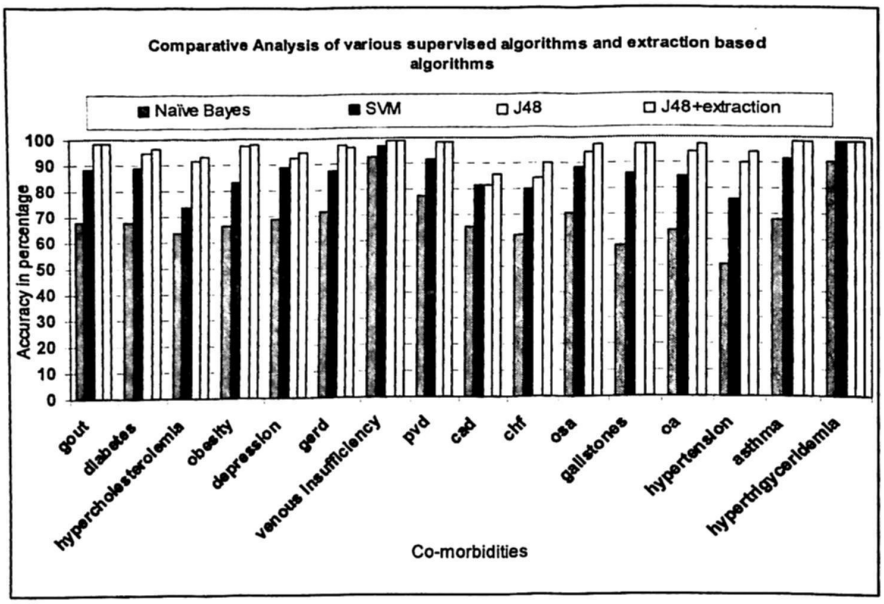

Figure 2. Accuracy of Various algorithms and combination of J48 and extraction based algorithm for co-morbidities classification based on textual judgment

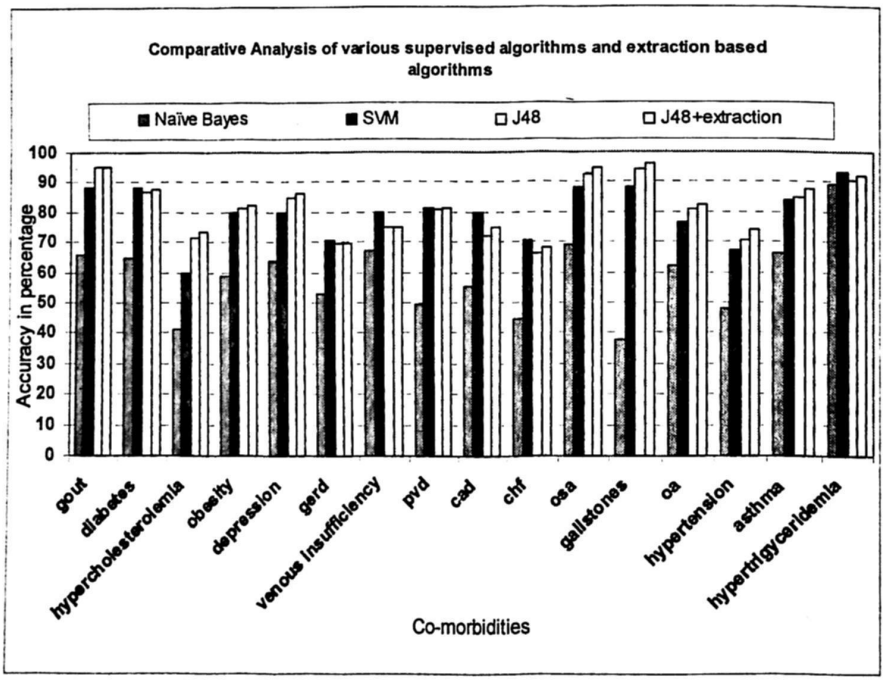

Figure 3. Accuracy of Various algorithms and combination of J48 and extraction based algorithm for co-morbidities classification based on Intuitive judgment

Support Vector Machine (SVM) based classification and J48 decision tree based classification shows comparable and relatively better results. In some co-morbidity, SVM classifier even outperforms J48 and combination of J48 and extraction based classifier.

6 Conclusion

We have explored various problems associated with free text processing of narrative medical reports and have also presented approaches to address some of those. We explored the semantic aspects of medical reports such as synonymy/hyponymy extraction, negation handling and multi-classes interdependence. The experimental results show that this method does help in increasing the accuracy of the results obtained from statistical algorithms. The identification of synonymous/hyponymous words can help not only in document classification, but can be important many other applications. One of the advantages of automated hyponyms extraction to thesauri based extraction is that the technique can be ported to any domain easily and is very domain dependent capturing the best of the context. Another advantage is that it can retain the semantics of the context without a full document analysis (works just by extracting relevant part of the document). This makes it computationally less expensive too. As a future work, we intend to explore other semantic aspects such as temporality, experiencer etc to make better predictions.

References

1. Informatics for Integrating Biology and the Bedside, https://www.i2b2.org/NLP/Main.php
2. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 34, 301--310 (2001)
3. Aronow, D.B., Feng, F., Croft, W.B., Ad Hoc Classification of Radiology Reports. Journal of the American Medical Informatics Association, pp. 393-411 (1999)
4. Lu, C., Probabilistic and machine learning approaches to medical classification problems, ph.d. dissertation (2005)
5. Riloff, E., Lehnert, W., Information Extraction as a Basis for High-Precision Text Classification, ACM Transactions on Information Systems (1994)
6. Bishop, C. M., Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
7. Averbuch, M., Karson, T., Ben-Ami, B., Maimon, O., Rokach, L. Context-Sensitive Medical Information Retrieval. in Proc. of 11th World Congress on Medical Informatics (MEDINFO-2004). San Francisco, CA: IOS Press (2004)
8. Goryachev, S., Sordo, M., Zeng, Q. T., Ngo, L., Implementation and Evaluation of four different methods of Negation Detection (2006)
9. Gonzalo J, Verdejo F, Chugur I, Cigarran J: Indexing with WordNet synsets can improve Text Retrieval. In Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP Montreal, Canada (1998)
10. WordNet: A Lexical Reference System and its Application. MIT Press, Cambridge, MA., pages 265—283 (1998)

11. Pedersen, T., Patwardhan, S., and Michelizzi, J. "WordNet::Similarity - Measuring the Relatedness of Concepts" In Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004), pp. 38-41. Boston (2004)
12. Snow, R., Jurafsky, D., Ng, A., Learning syntactic patterns for automatic hypernym discovery. In Proc of Neural Information Processing Systems Vancouver, Canada 1297-1304 (2004)
13. Pereira, F., Tishby, N., Lee, L., Distributional Clustering of English Words. In Proc of ACL-1993 Columbus, Ohio, USA .pp. 183-190 (1993)
14. Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A., Wilbur WJ: Automatic extraction of gene and protein synonyms from MEDLINE and journal articles, Proc AMIA Symp .pp. 919-923 (2002)
15. Morin, E., Jacquemin, C., Automatic Acquisition and Expansion of Hypernym Links. Computers and the Humanities 363-396 (2004)
16. Dumais, S., Furnais, G., Landauer, T., Indexing by Latent Semantic Analysis, American Society for Information Science, pp. 391-407 (1990)
17. McCrae, J., Collier, N., Synonym set extraction from the biomedical literature by lexical pattern discovery, BMC Bioinformatics (2008)
18. http://nlp.stanford.edu/IR-book/html/htmledition/classification-with-more-than-two-classes-1.html
19. Wikipedia. Information Extraction http://en.wikipedia.org/wiki/Information_extraction
20. MetaMap transfer (MMTx) mmtx.nlm.nih.gov/
21. Unified Medical Language System (UMLS) www.nlm.nih.gov/research/umls/
22. Weka, Data mining with open source machine learning software www.cs.waikato.ac.nz/ml/weka/