# Grading the Quality of Medical Evidence

**Binod Gyawali, Thamar Solorio**
CoRAL Lab
Department of Computer and Information Sciences
University of Alabama at Birmingham, AL, USA
{bgyawali,solorio}@cis.uab.edu

**Yassine Benajiba**
Clinical Decision Support Solutions Department
Philips Research North America, Briarcliff Manor, NY, USA
yassine.benajiba@philips.com

## Abstract

Evidence Based Medicine (EBM) is the practice of using the knowledge gained from the best medical evidence to make decisions in the effective care of patients. This medical evidence is extracted from medical documents such as research papers. The increasing number of available medical documents has imposed a challenge to identify the appropriate evidence and to access the quality of the evidence. In this paper, we present an approach for the automatic grading of evidence using the dataset provided by the 2011 Australian Language Technology Association (ALTA) shared task competition. With the feature sets extracted from publication types, Medical Subject Headings (MeSH), title, and body of the abstracts, we obtain a 73.77% grading accuracy with a stacking based approach, a considerable improvement over previous work.

## 1 Introduction

"Evidence Based Medicine (EBM) is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996). EBM requires to identify the best evidence, understand the methodology and strength of the approaches reported in the evidence, and bring relevant findings into clinical practice. Davidoff et al. (1995) express EBM in terms of five related ideas. Their ideas imply that the conclusions should be derived based on the best evidence available, the clinical decisions should be made based on the conclusions derived, and the performance of the clinical decisions should be evaluated constantly. Thus, physicians practicing EBM should be constantly aware of the new ideas and the best methodologies available based on the most recent literature. But the amount of clinical documents available is increasing everyday. For example, Pubmed, a service of the US National Library of Medicine contains more than 21 million citations for biomedical literature from MEDLINE, life science journals, and online books (last updated on December 7, 2011) [1]. The abundance of digital information makes difficult the task of evaluating the quality of results presented and the significance of the conclusions drawn. Thus, it has become an important task to grade the quality of evidence so that the most significant evidence is incorporated into the clinical practices.

There are several scale systems available to grade medical evidence. Some of them are: hierarchy of evidence proposed by Evans (2003), Grading of Recommendations Assessment, Development, and Evaluation (GRADE) scale by GRADE (2004), and Strength of Recommendation Taxonomy (SORT) scale by Ebell et al. (2004). The SORT scale addresses the quality, quantity, and consistency of evidence and proposes three levels of ratings: A, B, and C. Grade A is recommended based on the consistent, good-quality patient-oriented evidence, grade B is based on the inconsistent or limited-quality patient-oriented evidence, and grade C is based on consensus, disease-oriented evidence, usual practice, expert opinion or case studies.

[1] http://www.ncbi.nlm.nih.gov/books/NBK3827/



*Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*, pages 176–184,
Montréal, Canada, June 8, 2012. 

The Australasian Language Technology Association (ALTA) 2011 organized the shared task competition[2] to build an automatic evidence grading system for EBM based on the SORT grading scale. We carry out our experiments using the data set provided for the competition and compare the accuracy of grading the evidence by applying basic approaches and an ensemble (stacking) based approach of classification. We show that the later approach can achieve 73.77% of grading accuracy, a significant improvement over the basic approaches. We further extend our experiments to show that, using feature sets generated from the method and conclusion sections of the abstracts helps to obtain higher accuracy in evidence grading than using a feature set generated from the entire body of the abstracts.

## 2 Related Work

To the best of our knowledge, automatic evidence grading based on a grading scale was initiated by Sarker et al. (2011). Their work was based on the SORT scale to grade the evidence using the corpus developed by Molla-Aliod (2010). They showed that using only publication types as features could yield an accuracy of 68% while other information like publication types, journal names, publication years, and article titles could not significantly help to improve the accuracy of the grading. Molla-Aliod and Sarker (2011) worked on the evidence grading problem of 2011 ALTA shared task and achieved an accuracy of 62.84% using three sequential classifiers, each trained by one of the following feature sets: word n-grams from the abstract, publication types, and word n-grams from the title. They applied a three way classification approach where the instances classified as A or C were removed from the test set and labeled as such, while instances classified as B were passed to the next classifier in the pipeline. They repeated this process until they reached the end of three sequential classifiers.

Most of the EBM related work is focused on either the identification of important statements from the medical abstracts or the classification of medical abstracts to facilitate the retrieval of important documents. Work by Demner-Fushman et al. (2006), Dawes et al. (2007), Kim et al. (2011) automatically identify the key statements in the medical abstracts and classify them into different levels that are considered important for EBM practitioners in making decisions. Kilicoglu et al. (2009) worked on recognizing the clinically important medical abstracts using an ensemble learning method (stacking). They used different combinations of feature vectors extracted from documents to classify the evidence into relevant or non relevant classes. They approached the problem as a binary classification problem without using any grading scales.

Systematic Reviews (SRs) are very important to support EBM. Creating and updating SRs is highly inefficient and needs to identify the best evidence. Cohen et al. (2010) used a binary classification system to identify the documents that are most likely to be included in creating and updating SRs.

In this work, we grade the quality of evidence based on the SORT scale, that is different from most of the existing works related to classification of abstracts and identification of key statements of abstracts. We work on the same problem as by Molla-Aliod and Sarker (2011) but, we undertake the problem with a different approach and use different sets of features.

## 3 Dataset

We use the data of 2011 ALTA shared task competition that contains three different sets: training, development and test set. The number of evidence instances present in each set is shown in Table 1. Each data set consists of instances with grades A, B, or C based on the SORT scale. The distribution of evidence grades is shown in Table 2.

| Data Set | No. of Evidence Instances |
|---|---|
| Training Set | 677 |
| Development Set | 178 |
| Test Set | 183 |

Table 1: Evidence per data set

The evidence instances were obtained from the corpus developed by Molla-Aliod and Santiago-Martinez (2011). The corpus was generated based on the question and the evidence based answer for the question along with SOR grade obtained from the *"Clinical Inquiries"* section of the Journal of

[2] http://www.alta.asn.au/events/sharedtask2011

| Grades | Training set (%) | Development set (%) | Test set (%) |
|---|---|---|---|
| A | 31.3 | 27.0 | 30.6 |
| B | 45.9 | 44.9 | 48.6 |
| C | 22.7 | 28.1 | 20.8 |

Table 2: Evidence distribution per grade

Family Practice (JFP). A sample question from the JFP Clinical Inquiries section is *"How does smoking in the home affect children with asthma?"*. Each evidence contains at least one or more publications depending upon from which publications the evidence was generated. Each publication is an XML file containing information such as abstract title, abstract body, publication types, and MeSH terms. Each publication is assigned at least one publication type and zero or more MeSH terms. The MeSH terms vocabulary [3] is developed and maintained by the National Library of Medicine and is used in representation, indexing and retrieval of medical documents. Some of the medical document retrieval work emphasizes the use of MeSH terms in the efficient retrieval of documents (Trieschnigg et al., 2009; Huang et al., 2011). MeSH terms are also used in document summarization (Bhattacharya et al., 2011).

```
28092 B 10593430
28094 C 14712967 12269676 12165283 12618157
18163 A 8381089 7972972
18164 C 7972972 8621845
18166 A 8386917
16192 B 10920726
18162 A 8386917
52371 C 11642617 2328431 10532723
16193 B 9569395 12069675
```

Figure 1: Sample data file

Each data set contains an additional grade file with the information related to the evidence instances, their grades, and the publications. A sample of the file is shown in Figure 1. The first column contains the evidence id, the second column contains the grades A, B, or C of the evidence based on the SORT scale, and the remaining columns show the publication id of each publication in the evidence.

[3] http://www.nlm.nih.gov/mesh

The problem in this task is to analyze the publications in each evidence provided and classify them into A, B or C.

The dataset available for our research has abstracts in two different formats. One of them contains abstracts divided into sections: background, objective, method, result, and conclusion. The other format contains abstracts with all the information in a single block without any sections. A sample of an abstract having only four sections in the given data is shown below:

Objectives: To determine the effectiveness of a muscle strengthening program compared to a stretching program in women with fibromyalgia (FM).

Methods: Sixty-eight women with FM were randomly assigned to a 12 week, twice weekly exercise program consisting of either muscle strengthening or stretching. Outcome measures included muscle strength (main outcome variable), flexibility, weight, body fat, tender point count, and disease and symptom severity scales.

Results: No statistically significant differences between groups were found on independent t tests. Paired t tests revealed twice the number of significant improvements in the strengthening group compared to the stretching group. Effect size scores indicated that the magnitude of change was generally greater in the strengthening group than the stretching group.

Conclusions: Patients with FM can engage in a specially tailored muscle strengthening program and experience an improvement in overall disease activity, without a significant exercise induced flare in pain. Flexibility training alone also results in overall improvements, albeit of a lesser degree.

In the abstract above, we see that the approaches applied for the study are described in the method section, and the outcome and its effectiveness are described in the conclusion section.

## 4 Proposed Methodology

In this paper we propose a system to identify the correct grade of an evidence given publications in the evidence. We deal with the problem of evidence grading as a classification problem. In evidence grading, basic approaches have been shown to have poor performance. Molla-Aliod and Sarker (2011) showed that a basic approach of using simple bag-of-word features and a Naive Bayes classifier achieved 45% accuracy and proposed a sequential approach to improve the accuracy at each step. Our preliminary studies of applying the simple classification approach also showed similar results. Here, we propose a stacking based approach (Wolpert,

1992) of evidence grading. Stacking based approach builds a final classifier by combining the predictions made by multiple classifiers to improve the prediction accuracy. It involves two steps. In the first step, multiple base-level classifiers are trained with different feature sets extracted from a dataset and the classifiers are used to predict the classes of a second dataset. Then, a higher level classifier is trained using the predictions made by the base-level classifiers on the second dataset and used to predict the classes of the actual test data. In this approach, base-level classifiers are trained independent of each other and allowed to predict the classes. Based on the predictions made by these base-level classifiers, the higher level classifier learns from those predictions and makes a new prediction that is the final class.

Our stacking based approach of classification uses five feature sets. In the first step of classification, we train five classifiers using different feature sets per classifier and use the classifiers to predict the grades of the development dataset. Thus, at the end of the first step, five different predictions on the development dataset are obtained. In the second step, a new classifier is trained using the grades predicted by the five classifiers as features. This new classifier is then used to predict the grades of the test dataset.

## 5 Features

We extracted six sets of features from the publications to perform our experiments. They are as follows:

1. Publication types
2. MeSH terms
3. Abstract title
4. Abstract body
5. Abstract method section
6. Abstract conclusion section

For feature set 1, we extracted 30 distinct publication types from the training data. For the MeSH terms feature set, we selected 452 unique MeSH terms extracted from the training data. The publications contained the descriptor name of the MeSH terms having an attribute "majortopicyn" with value 'Y' or 'N'. As MeSH terms feature set, we selected only those MeSH term descriptor names having majortopicyn='Y'.

We extracted the last four sets of features from the title, body, method, and conclusion sections of the abstracts. Here, the body of an abstract means the whole content of the abstract, that includes background, objective, method, result, and conclusion sections. We applied some preprocessing steps to generate these feature sets. We also applied a feature selection technique to reduce the number of features and include only the high informative features from these feature sets. The details about preprocessing and feature selection techniques are described in Section 6.

We performed all the experiments on the basis of evidence, i.e. we created a single feature vector per evidence. If an evidence contained more than one publication, we generate its features as the union of the features extracted from all its publications.

The grades of the evidence in the SORT scale are based on the quality of evidence, basis of experiments, the methodologies used, and the types of analysis done. Grades also depend upon the effectiveness of the approach used in the experiments. The method section of an abstract contains the information related to the basis of the experiments, such as randomized controlled trails, systematic review, cohort studies, and the methods used in their research. The conclusion section of the abstract usually contains the assertion statements about how strongly the experiment supports the claims. Analysis of the contents of abstracts shows that the information needed for grading on SORT scale is typically available in the method and conclusion sections, more than in the other sections of the abstracts. Thus, we used the method and conclusion sections of the abstracts to generate two different feature sets so that only the features more likely to be important in grading using the SORT rating would be included.

**Separating method and conclusion sections of the abstracts**

In order to extract features from the method and conclusion sections, we should separate them from the body of abstracts, which is a challenging task for those abstracts without section headers. Of the total number of abstracts, more than one-third of the abstracts do not contain the section headers. In order to separate these sections, we used a very simple approach based on the number of sentences present

in the method and conclusion sections, and the body of the abstracts. We used the following information to separate the method and conclusion sections from these abstracts: i) Order of sections in the abstracts, ii) Average number of sentences in the method and conclusion sections of the abstracts having sections, and iii) Average number of sentences in the entire body of the abstracts not having sections. All the abstracts having section headers contained the sections in the same order: background, objective, method, result and conclusion. From the available training dataset, we calculated:

i. The average number of sentences in the method (4.14) and conclusion (2.11) sections of the abstracts divided into sections

ii. The average number of sentences (8.78) of the abstracts not having sections

Based on these values, we fragmented the abstracts that do not have the section headers and separated the method and conclusion sections from them. Table 3 shows how the method and conclusion sections of those abstracts were generated. For example, the fourth row of the table says that, if an abstract without section headers has 6, 7 or 8 sentences (let it be $n$), then the 3rd, 4th and 5th sentences were considered as the method section, and the $n^{th}$ sentence was considered as the conclusion section.

| Total sentences in Abstracts($n$) | Method | Conclusion |
|---|---|---|
| 1 | None | 1 |
| 2 or 3 | 1 | $n$ |
| 4 or 5 | 2 and 3 | $n$ |
| 6 or 7 or 8 | 2, 3 and 4 | $n$ |
| More than 8 | 3, 4 and 5 | $n$-1 and $n$ |

Table 3: Selecting method and conclusion of the abstracts having a single block

## 6 Experiments and Results

This section describes the two sets of experiments performed to compare the performance of the stacking based approach and the effectiveness of the base-level classifiers used. The first set of experiments was done to provide a baseline comparison against our stacking based approach. The second set consists of five experiments to evaluate different configurations of stack based classifiers. The basic approach of classification implies the use of a single classifier trained by using a single feature vector.

We applied preprocessing steps to generate feature sets from the title, body, method and conclusion sections of the abstracts. The preprocessing steps were: detecting sentences using OpenNLP Sentence Detector[4], stemming words in each sentence using Porter Stemmer (Porter, 1980), changing the sentences into lower-case, and removing punctuation characters from the sentences. After the preprocessing step, we generated features from the unigrams, bigrams and trigrams in each part. We removed those features from the feature sets that contained the stopwords listed by Pubmed[5] or contained any token having a length less than three characters. To remove the less informative features, we calculated the information gain of the features in the training data using Weka (Hall et al., 2009) and selected only the top 500 high informative features for each feature set. We used the Weka SVM classifier for all the experiments. Based on the best result obtained after a series of experiments run with different kernel functions and regularization parameters, we chose the SVM classifier with a linear kernel and regularization parameter equals 1 for all the experiments. We used a binary weight for all the features.

### 6.1 First set of experiments

In the first set, we performed nine experiments using the basic classification approach and one experiment using the stacking based approach. The details of the experiments and the combinations of the features used in them are as shown in Table 4.

The first six experiments in the table were implemented by applying a basic approach of classification and each using only a single set of features. Experiments 7, 8, and 9 were similar to the first six experiments except, they used more than one set of features to create the feature vector. Each feature in the experiments 7, 8, and 9 encode the section of its origin. For example, if feature $abdomen$ is present in method as well as conclusion sections, it is represented as two distinct features $conc\_abdomen$ and $method\_abdomen$. In experiment 10, we applied

[4]http://incubator.apache.org/opennlp

[5]http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/?report=objectonly)

the stacking approach of classification using five base-level classifiers. The base-level classifiers in this experiment are the basic classifiers used in experiments 1 to 5.

| Exp. No. | Features used | Exp. type |
|---|---|---|
| 1. | Publication types | Basic approach |
| 2. | MeSH terms | |
| 3. | Abstract title | |
| 4. | Abstract method | |
| 5. | Abstract conclusion | |
| 6. | Abstract body | |
| 7. | Publication types, MeSH terms | |
| 8. | Publication types, MeSH terms, Abstract title, Abstract body | |
| 9. | Publication types, MeSH terms, Abstract title, Abstract method, Abstract conclusion | |
| 10. | Publication types | Stacking based approach |
| | MeSH terms | |
| | Abstract title | |
| | Abstract method | |
| | Abstract conclusion | |

Table 4: Experiments to compare basic approaches to a stacking based approach

Figure 2 shows the results of the 10 experiments described in Table 4 in the same order, from 1st to 10th place and the result of the experiment by Molla-Aliod and Sarker (2011). The results show that the stacking based approach gives the highest accuracy (73.77%), outperforming all the basic approaches applying any combination of feature sets. The stacking based approach outperforms the baseline of a single layered classification approach (Exp 9) that uses all the five sets of features. Molla-Aliod and Sarker (2011) showed that a simple approach of using a single classifier and bag-of-words features could not achieve a good accuracy (45.9%) and proposed a new approach of using a sequence of classifiers to achieve a better result. Similar to their simple approach, our basic approaches could not achieve good results, but their performance is comparable to Molla-Aliod and Sarker (2011)'s baseline system. The result of our stacking based approach shows that our approach has a better accuracy than the sequencial classification approach (62.84%) proposed by

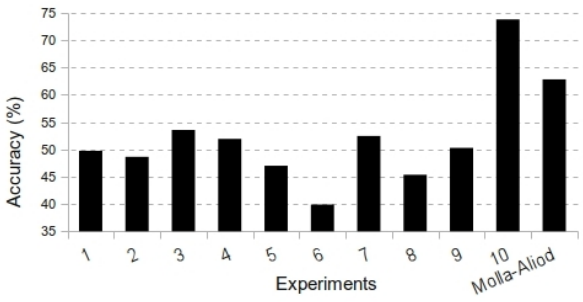


Figure 2: Comparison of accuracy of basic approaches to a stacking based approach. X-axis shows the experiments and Y-axis shows the accuracy of the experiments. The first nine experiments are based on the basic approach and the tenth experiment is based on the stacking based approach.

Molla-Aliod and Sarker (2011).

Our stacking based approach works on two levels. In the first level, the base-level classifiers predict the grades of the evidence. In the next level, these predictions are used to train a new classifier that learns from the predictions to identify the grades correctly. Moreover, the five feature sets used in our experiments were unrelated to each other. For example, the features present in MeSH headings were different from the features used in publication types, and similarly, the features present in the method section of the abstract were different from the features present in the conclusion section. Each base-level classifier trained by one of these feature sets is specialized in that particular feature set. Thus, using the predictions made by these specialized base-level classifiers to train a higher level classifier helps to better predict the grades, this cannot be achieved by a single classifier trained by a set of features (Exp. 1, 2, 3, 4, 5, 6), or a group of different feature sets (Exp. 7, 8, 9).

### 6.2 Second set of experiments

In the second set of experiments, we compared five experiments performed varying the base-level classifiers used in our stack based approach. Experiments 1 and 2 were performed using a single base-level classifier, that means that the second classifier is trained on only one feature. Experiments 3 and 4 were performed by using four base-level classifiers, and experiment 5 was performed using five base-

level classifiers. The 5th experiment in this set is same as the 10th experiment in the first set. The details about the feature sets used in each experiment are shown in Table 5.

| Exp. No. | Features used | No. of Base level classifiers |
|---|---|---|
| 1. | Publication types, MeSH terms, Abstract title, Abstract body | 1 |
| 2. | Publication types, MeSH terms, Abstract title, Abstract method, Abstract conclusion | 1 |
| 3. | Publication types | 4 |
| | MeSH terms | |
| | Abstract title | |
| | Abstract body | |
| 4. | Publication types | 4 |
| | MeSH terms | |
| | Abstract title | |
| | Abstract method, Abstract conclusion | |
| 5. | Publication types | 5 |
| | MeSH terms | |
| | Abstract title | |
| | Abstract method | |
| | Abstract conclusion | |

Table 5: Experiments to compare stacking based approach

Figure 3 shows the accuracy of the five experiments shown in Table 5 in the same order. It shows that the accuracy of 1st and 2nd experiments is lower than the accuracy of 3rd, 4th, and 5th experiments. In these two experiments, a feature vector generated from the prediction of a single base-level classifier is used to train the higher level classifier, that is not sufficient to make a correct decision.

Experiments 3, 4, and 5 show a considerable improvement in the accuracy of the grading. Comparing the results of experiments 3 and 4, we see that the 4th experiment has higher accuracy than the 3rd one. The difference between these experiments was the use of features from the method and conclusion sections of the abstracts in the 4th experiment, while using features from the entire body of abstracts in the 3rd experiment. The higher accuracy in the 4th experiment shows that the method and conclusion sections of the experiment contain high informative text that is important for evidence grading, while

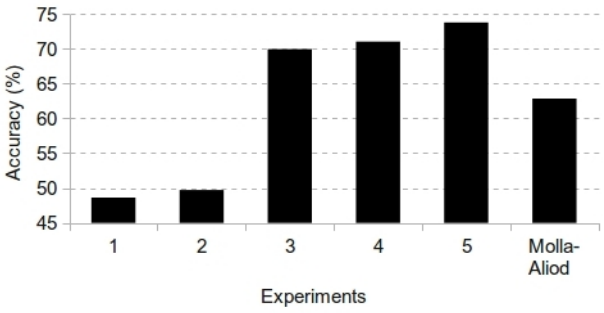


Figure 3: Comparison of accuracy of the stacking based approaches. X-axis shows the experiments and Y-axis shows the accuracy of the experiments. 1st and 2nd experiments use only one base-level classifier, 3rd and 4th experiment are based on four base-level classifiers and 5th one uses five base-level classifiers.

the body of abstracts may contain some information that is not relevant to the task. The same analysis can also be inferred from the results of experiment 8 and 9 in the first set of experiments. The highest accuracy obtained in the 5th experiment of applying 5 base-level classifiers shows that identifying the sections of the abstracts containing high informative features and using a sufficient number of base-level classifiers can help to achieve a good accuracy in evidence grading.

## 7 Error Analysis

The result obtained by the stacking based approach (5th experiment in Table 5) using five base-level classifiers gave a higher error rate in predicting grades A and C, compared to the error rate in predicting grade B. Most of the error is the misclassification of A to C and vice versa. One of the possible reasons of this might be due to the use of the feature set extracted from the conclusion section. Among the five base-level classifiers used in the experiment, the one trained by the features extracted from the conclusion sections has the lowest accuracy (5th experiment in Figure 2). We evaluated the text contained in the conclusion section of the abstracts in our dataset. The section mostly contains the assertion statements having the words showing strong positive/negative meanings. Conclusion of A grade evidence mostly contains the information that strongly asserts the claim *(e.g. emollient treatment*

***significantly*** *reduced the high-potency topical corticosteroid consumption in infants with AD)*, while that of C grade evidence is not strong enough to assert the claim *(e.g. PDL therapy should be considered among the better established approaches in the treatment of warts, although data from this trial suggest that this approach is* ***probably not*** *superior)*. It seems that the problem might be because of not processing the negations appropriately. So, in order to preserve some negation information present in the conclusion sections, we performed another experiment by merging words $no$, $not$, $nor$ with their successor word to create a single token from the two words. This approach still could not reduce the misclassification. Thus, the simple approach of extracting unigram, bigram, and trigram features from the conclusion section might not be sufficient and might need to include higher level analysis related to assertion/certainty of the statements to reduce the misclassification of the evidence.

Other possible reasons of the misclassification of the evidence might be the imbalanced data set. Our dataset (Table 2) contains higher number of instances with grade B than those with grades A and C. Moreover, the number of publications per evidence is not uniform, that ranges from 1 to 8 publications per evidence in the test data. Analyzing the results, we found that misclassification of evidence having only one publication is higher than that of the evidence having more than one publication. If an evidence contains only one publication, the features of the evidence extracted from a single publication might not be sufficient to accurately grade the evidence and might lead to misclassification.

In order to evaluate the appropriateness of our approach in extracting the method and conclusion sections, we performed a manual inspection of abstracts. We could not revise all the abstracts to verify the approach. Thus, we randomly selected 25 abstracts without section headers from the test data and viewed the content in them. We found that the conclusion section was appropriately extracted in almost all abstracts, while the selection of method section was partially effective. Our approach was based on the assumption that all the abstracts having many sentences have all the sections (background, objective, method, result, and conclusion). But we found that the abstracts do not follow the same format, and the start sentence of the method section is not consistent. Even a long abstract might sometimes start with the method section, and sometimes the objective section might not be present in the abstracts. This could lead to increase the error in our grading system.

## 8 Conclusion

This paper presents an approach of grading the medical evidence applying a stacking based classifier using the features from publication types, MeSH terms, abstract body, and method, and conclusion sections of the abstracts. The results show that this approach achieves an accuracy of 73.77%, that is significantly better than the previously reported work. Here, we present two findings: 1) We show that the stacking based approach helps to obtain a better result in evidence grading than the basic approach of classification. 2) We also show that the method and conclusion sections of the abstracts contain important information necessary for evidence grading. Using the feature sets generated from these two sections helps to achieve a higher accuracy than by using the feature set generated from the entire body of the abstracts.

In this work, all the information available in the method and conclusion sections of the abstracts is treated with equal weight. Evidence grading should not depend upon specific disease names and syndromes, but should be based on how strong the facts are presented. We would like to extend our approach by removing the words describing specific disease names, disease syndromes, and medications, and giving higher weight to the terms that describe the assertion of the statements. In our current work, we apply a simple approach to extract the method and conclusion sections from the abstracts not having sections. Improving the approach by using a machine learning algorithm that can more accurately extract the sections might help to increase the accuracy of grading. Including the information about the strength of assertions made in the conclusion sections could also help in boosting the accuracy. Future work would also include testing the effectiveness of our approach on other diverse data sets having complex structures of the evidence, or on a different grading scale.

## References


Sanmitra Bhattacharya, Viet HaThuc, and Padmini Srinivasan. 2011. Mesh: a window into full text for document summarization. *Bioinformatics*, 27(13):i120–i128.

Aaron M. Cohen, Kyle Ambert, and Marian McDonagh. 2010. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. *AMIA Annu Symp Proc.*, 2010:121 – 125.

Frank Davidoff, Brian Haynes, Dave Sackett, and Richard Smith. 1995. Evidence based medicine. *BMJ*, 310(6987):1085–1086, 4.

Martin Dawes, Pierre Pluye, Laura Shea, Roland Grad, Arlene Greenberg, and Jian-Yun Nie. 2007. The identification of clinically important elements within medical journal abstracts: Patient-Population-Problem, Exposure-Intervention, Comparison, Outcome, Duration and Results (PECODR). *Informatics in Primary Care*, 15(1):9–16.

Dina Demner-Fushman, Barbara Few, Susan E. Hauser, and George Thoma. 2006. Automatically Identifying Health Outcome Information in MEDLINE Records. *Journal of the American Medical Informatics Association*, 13(1):52 – 60.

M. H. Ebell, J. Siwek, B. D. Weiss, S. H. Woolf, J. Susman, B. Ewigman, and M. Bowman. 2004. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *American Family Physician*, 69(3):548–56+.

David Evans. 2003. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, 12(1):77–84.

GRADE. 2004. Grading quality of evidence and strength of recommendations. *BMJ*, 328(7454):1490, 6.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1).

Minlie Huang, Aurlie Nvol, and Zhiyong Lu. 2011. Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667.

Halil Kilicoglu, 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*, 16(1):25–31.

Su Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, 12(Suppl 2):S5.

Diego Molla-Aliod and Maria Elena Santiago-Martinez. 2011. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop*.

Diego Molla-Aliod and Abeed Sarker. 2011. Automatic Grading of Evidence: the 2011 ALTA Shared Task. In *Proceedings of Australasian Language Technology Association Workshop*, pages 4–8.

Diego Molla-Aliod. 2010. A Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop*, volume 8.

MF Porter. 1980. An algorithm for sufx stripping. *Program*, 14(3):130–137.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1.

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

Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska de Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. 2009. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241 – 259.