Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes

Mahdi Abdollahi[1], Xiaoying Gao[1], Yi Mei[1], Shameek Ghosh[2], and Jinyan Li[3]

[1] Victoria university of Wellington, Wellington, New Zealand
{mahdi.abdollahi, xiaoying.gao, yi.mei}@ecs.vuw.ac.nz
[2] Medius Health, Sydney, Australia
shameek.ghosh@mediushealth.org
[3] University of Technology Sydney, Sydney, Australia
Jinyan.Li@uts.edu.au

Abstract. Text classification is a challenging task for allocating each document to the correct predefined class. Most of the time, there are irrelevant features which make noise in the learning step and reduce the precision of prediction. Hence, more efficient methods are needed to select or extract meaningful features to avoid noise and overfitting. In this work, an ontology-guided method utilizing the taxonomical structure of the Unified Medical Language System(UMLS) is proposed. This method extracts concepts of appeared phrases in the documents which relate to diseases or symptoms as features. The efficiency of this method is evaluated on the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set. The obtained experimental results show significant improvement by the proposed ontology-based method on the accuracy of classification.

Keywords: Coronary Artery Disease Notes · Text Classification · Feature Selection · Conceptualization · Ontology.

1 Introduction

Text mining is one of the important topics in artificial intelligence which deals with analyzing different types of unstructured text to extract useful knowledge. There are many tasks in text mining such as text classification, text clustering, entity extraction, document summarization and semantic analysis. Text classification is one of the extensively studied natural language processing tasks. In text classification, the goal is to automatically classify text documents into one or more predefined classes. For example, detecting spam and non-spam emails, automatically tagging client queries and categorizing news articles are some applications of text classification. The main steps of text classification consist of preprocessing, representing text, weighting features, selecting features, training, testing and evaluating. The aim of the feature selection step is choosing useful features to use in text classification.

In text classification, using all of the unique words in the documents as features is one of the simplest way of text representation. The first problem with this method is that the number of features is big and the second issue is the existence of noisy features. The massive data can reduce the speed of the learning step and increase the time cost. The noisy data can have negative effects on the learned model which lead to poor accuracy of label prediction. Hence, feature selection plays a significant role which can improve the performance of classification by selecting meaningful features and at the same time reducing the number of noisy features.

The most utilized feature selection algorithm in analyzing text is single feature ranking [14], which is a filter technique of choosing m features as a sub set from the n features by considering the top m features based on their rank. The significance of each feature is defined by its contribution to the classification task, and some basic measures of relevance are: Logistic Regression [18], Statistical Testing [17], Pearson's correlation [15], and distinctive data hypothesis based measures [16]. Single feature ranking is a simple method with a low computation cost which encourages many researchers to apply it in their research, however, it does not consider interactions between features. As a matter of fact, most filter methods assess features separately which cannot distinguish interactions between features [15].

One of the solutions can be detecting meaningful phrases from the original documents as features and then extracting their concepts by considering the domain of documents and labels. In this approach, we use phrases consisting of different words which keep the interactions between features, and meanwhile, reduce the number of extracted features.

This paper proposes a method which applies ontology by referring to Unified Medical Language System(UMLS) for entity recognition, and then aggregates frequent entities to create features. The proposed method is integrated with five common text classification methods to answer the following research questions:

1. Whether the proposed method can reduce the number of features and keep the meaningful features; and
2. Whether the proposed method can increase the accuracy in classification of the targeted clinical text.

The rest of the paper is organized as follows: section 2 describes the problem and summaries the related work. Our method is presented in section 3. The obtained experimental results are shown in section 4. At the end, the conclusions and future work are presented in section 5.

2 Related work

Text classification is the task of assigning label l_i to document d_j, where $l_i \in L = \{l_1, \cdots, l_{|L|}\}$ and $d_j \in D = \{d_1, \cdots, d_{|D|}\}$, using a function F:

$$F : D \to L \tag{1}$$

In formula 1, function F is a classifier which gets documents (D) as input and allocates labels (L) as output to each of the input documents. In this paper, we focus on binary classification. Hence, the set L is $\{0, 1\}$, where 1 is disease, and 0 is no disease.

The extraction and selection of features for document classification problems has received a lot of interest in the past. Typically, a lot of these algorithms rank features using statistics from the distribution of features in the given corpus [11, 12]. Existing methods have employed metrics associated with word frequency, information gain, mutual information, term frequency-inverse document frequency (tf-idf) for extracting textual features [13]. However, the afore-mentioned techniques tend to treat each feature separately, i.e they ignore the dependencies between features.

In clinical text mining, the text describes a set of clinical events within a narrative, with the goal of producing an explanation as precise and comprehensive as possible when describing the health status of a patient. Generally, such texts include a heavy use of domain specific terminology and the frequent inclusion of acronyms, which makes clinical text analysis very different from standard text mining. Specially, a discriminative combination of domain-specific medical events reported within a clinical note can be highly indicative of a patient's condition.

There has been research that applies text classification on clinical text. Previously, Pratt et al [1] employed words, medical phrases, and their combinations as features for medical document classification. Multi-label classification performance based on an associative classifier is examined on medical articles [2]. In another study, Hidden Markov models were used for classification [3]. In a recent study, an approach using support vector machines and latent semantic indexing was applied to some data sets including the ones consisting of medical abstracts [6]. The performances of classifiers on medical document classification was analyzed for two cases where stemming was applied and not applied [7]. The impact of different text representations of biomedical texts on the performance of classification were analyzed [8]. Feature selection methods using Gini Index were employed along with models like bayesian networks and decision trees to improve medline document classification [9].

Besides, there exist a number of studies in the literature where ontology-based classification approaches have been applied [4, 5]. The use of ontologies like Unified Medical Language System(UMLS), Systematized Nomenclature of Medicine(SNOMED), and Medical Subject Headings(MeSH) have proved very useful for improving classification performance [10, 11, 26].

In addition, some work has used clinical records for prominent tasks such as finding risk factors for diabetic patients [22], extracting Framingham risk score(FRF) for target population [23], using rule-based and dictionary-based methods to identify heart disease risk factors [24], and applying a rule-based method by combining with regular expression and UMLS to spot risk of heart disease [25]. our approach applies an ontology as a feature selection method for

text classification and our target is to identify Coronary Artery Disease (CAD) disease.

By analyzing the previous work, it is noticeable that the majority of disease-targeted systems have tended to develop static rule-based systems which require human interventions every time the model is updated with new features. Such systems are not scalable for practical machine learning purposes. Our system allows an easier and flexible selection of different types of medical concepts to enable automatic extraction of features or combinations and generation of a prediction model. This is an easier way to investigate domain concepts and determine which concepts are most discriminative to a classification problem. Furthermore, our system provides the ability to allow a domain user to interactively change the concepts and auto-build machine learning models for Coronary Artery Disease (CAD) investigation.

3 Our ontology based approach

In this section, the proposed method and the used tools for extracting concepts of phrases are described in details. Additionally, our way of labeling the candidate data set is introduced.

3.1 Overview

One of the important points in text classification problems is to investigate the domain of documents which should be classified and the domain of classes that documents should be labeled with. This can help to select only related features of the documents to the domain for training phase and improve the accuracy of prediction for unseen documents. In the clinic text classification task all of the documents are discharge notes of patients in medical domain. The candidate class is whether or a disease such as that Coronary Artery Disease (CAD) is present. Our goal is to select features that have relations with the disease. In this case, the performance of the learned model can be improved.

To achieve the above goal, our proposed algorithm employs the knowledge in the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set and UMLS library. For this purpose, the MetaMap tool is used to extract all the concepts of existing phrases for each document using the UMLS. As shown in Fig. 1, the concepts extraction step is employed on both the training and the test documents. Then, by considering the medical domain, the concept selection step is performed on the obtained concepts. As a first step, two concepts are selected among all the concepts: "Disease or Syndrome" and "Sign or Symptom". By following this way of concept selection, the meaningful concepts will be selected which will assist the training phase to learn better in order to increase the accuracy of classification.

3.2 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [19] was introduced for modeling the language of health and biomedicine. UMLS is a source of knowledge which improves the performance of information systems in the biomedical area. It provides three main resources: the Metathesaurus, the Semantic Network and the

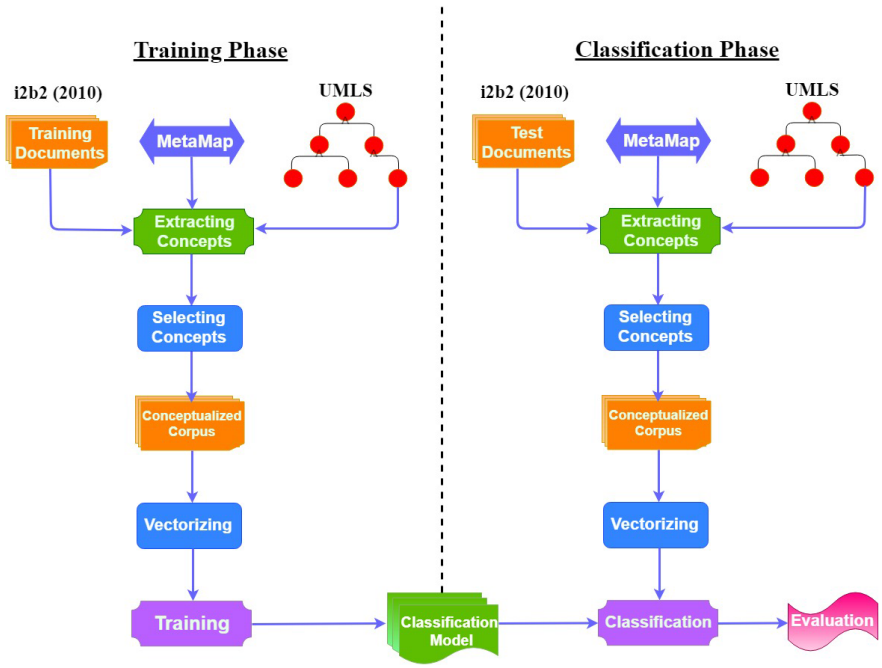

Fig. 1. The flowchart of the architecture of using MetaMap and UMLS for text classification.

SPECIALIST Lexicon. The largest component of UMLS is the Metathesaurus. It gives services such as finding biomedical concepts of phrases and relationships between concepts (e.g. SNOMED-CT, Mesh, etc.). The Semantic Network includes a collection of extensive topic classes, and different types of Semantics, which cover a matchable classification of concepts provided in the UMLS Metathesaurus, and a category of relationships and Semantic Relations between Semantic Types. The SPECIALIST lexicon includes a specialised English vocabulary of biomedical words.

3.3 MetaMap Tool

MetaMap is an exceptionally configurable program created by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM). It maps text to the UMLS Metathesaurus to find Metathesaurus concepts in text. MetaMap is a knowledge concentrated approach that utilizes computational-linguistic, natural-language processing (NLP) and symbolic methods. MetaMap is applied broadly in Information Retrival (IR), data mining applications. Furthermore, it is utilized for automatically biomedical literature indexing at the U.S. National Library of Medicine (NLM). It allows the mapping between text content and related concepts in UMLS. To achieve this goal, MetaMap breaks the content into expressions and after that, for each expression, it selects the mapping alternatives based on the ranking of mapping quality.

3.4 Conceptualization

Two sentences are given below as a sample to show how MetaMap works on the input notes and what output it provides in classification process.

> *"Hyperlipidemia: The patient's Lipitor was increased to 80 mg q.d. A progress note in the patient's chart from her assisted living facility indicates that the patient has had shortness of breath for one day."*

Fig. 2 shows a segment of the returned results from MetaMap. Table 1 summaries the extracted concepts of detected meaningful phrases from the sample sentences using MetaMap. As can be observed, the phrase "hyperlipidemia" belongs to "[Disease or Syndrome]" and "[Finding]" concepts. The phrase "shortest of breath" is allocated to the "[Sign or Symptom]", [Clinical Attribute] and "[Intellectual Product]" concepts. Considering the medical domain and the type of the classes in the selected data set, we choose concepts that appear in the "[Disease or Syndrome]" or "[Sign or Symptom]" categories. First we identify these two categories which are in square brackets, then the phrase that is within the round parentheses at the same line will be extracted as the main phrase. For example, the phrase "Dyspnea" is extracted in line 19 of Fig. 2 for the phrase "shortness of breath". After finishing the concept selection step, the obtained phrases will be used instead of the original documents in the binary classification problem. In order to give weights to the extracted terms of the documents, TF-IDF is applied in the vectorization step and each document is represented as a vector of weights based on the TF-IDF function.

```
   ---------------------------------------------------------------------------------------------------
1  Phrase: hyperlipidemia .
2  >>>>> Phrase
3  hyperlipidemia
4  <<<<< Phrase
5  >>>>> Mappings
6  Meta Mapping (1000):
7    1000   Hyperlipidaemia, NOS (Hyperlipidemia) [Disease or Syndrome]
8  Meta Mapping (1000):
9    1000   Hyperlipidemia (Serum lipids high (finding)) [Finding]
10 <<<<< Mappings
11 Processing 00000000.tx.7: MEDICATIONS ON ADMISSION : Lipitor , Flexeril ,
12 hydrochlorothiazide and Norvasc .
   ---------------------------------------------------------------------------------------------------
13 Phrase: shortness of breath
14 >>>>> Phrase
15 shortness of breath
16 <<<<< Phrase
17 >>>>> Mappings
18 Meta Mapping (1000):
19   1000   SHORTNESS OF BREATH (Dyspnea) [Sign or Symptom]
20 Meta Mapping (1000):
21   1000   Shortness of breath (Shortness of breath:-:Point in time:^Patient:-) [Clinical Attribute]
22 Meta Mapping (1000):
23   1000   Shortness of breath (How Often Shortness of Breath) [Intellectual Product]
24 <<<<< Mappings
   ---------------------------------------------------------------------------------------------------
```

Fig. 2. A segment of returned results of extracted concepts using MetaMap.

3.5 Data Preprocessing and Labelling

The idea of the paper is tested on the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set. This is the first time that the data set is used for text classification problem. This paper focuses on binary classification, so all the documents are labeled based on whether or not the Coronary Artery Disease (CAD) is present. Each document in the original data set has three files

Table 1. The extracted concepts of example sentences using MetaMap.

Sentences	Detected Phrases	Extracted Concepts	Selected
First Sentence	hyperlipidaemia	**[Disease or Syndrome]**	✓
		[Finding]	×
	patient	[Patient or Disabled group]	×
	Lipitor	[Organic Chemical, Pharmacologic Substance]	×
	80%	[Quantitative Concept]	×
	mg++ increased	[Finding]	×
Second Sentence	progress note	[Clinical Attribute]	×
		[Intellectual Product]	×
	patient chart	[Manufactured Object]	×
	assisted living facility	[Healthcare Related Organization, Manufactured Object]	×
	patient	[Patient or Disabled group]	×
	shortness of breath	**[Sign or Symptom]**	✓
		[Clinical Attribute]	×
		[Intellectual Product]	×
	one day	[Temporal Concept]	×

consisting of "Concepts.con", "Relations.rel", and "Assertions.ast" which were provided by the i2b2 organization for Relations Challenge. We used the content of "Assertions.ast" file of each document to determine the label of it. As shown in Fig. 3, there are a number of problem names inside each Assertion file. To label all of the documents, at the first step, all the lines of the file is searched for the "Coronary Artery Disease" phrase. If the phrase is found by the search, the second step will be checking whether the disease is present or not. If the name of illness appears with the phrase "present" in the same line, we will consider that the document is in the CAD class. First line of Fig. 3 indicates a sample of both the phrases "Coronary Artery Disease" and "present" occurring in the same line. By following this rule, all of the labels of 170 training documents and 256 test documents are extracted.

```
---------------------------------------------------------------------
Assertions

c="coronary artery disease" 24:8 24:10||t="problem"||a="present"
c="myocardial infarction" 24:19 24:20||t="problem"||a="absent"
c="cyanosis" 44:6 44:6||t="problem"||a="absent"
c="hypertension" 24:0 24:0||t="problem"||a="present"
         .
         .
         .
c="chest pain" 47:11 47:12||t="problem"||a="present"
---------------------------------------------------------------------
```

Fig. 3. A subpart of the Assertions file.

4 Results and Discussions

The performance of the proposed method is assessed on the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set. All of the documents are labeled based on whether or not Coronary Artery Disease(CAD) is present as a binary classification problem. The 2010 i2b2 data set includes 426 documents in different topics. The original data set split the documents into 170

train set and 256 test set. Among all the topics, class CAD is considered to form a binary classification. Five popular classifiers are used in the experimental comparison. The classifiers are Naive Bayes, Linear Support Vector Machine (SVM), K-Nearest Neighbor(KNN), Decision Tree and Logistic Regression. The performance of the classifiers are evaluated based on three main metrics (Precision, Recall, F1-measure) using micro-average and macro-average.

Some of the parameters of these classifiers are turned to get better results. For this purpose, the number of the neighbors in the KNN is set to 28 for the "n_neighbors" parameter. In the Decision Tree classifier, the maximum depth of the tree and the random number generator are set to 14 for the "max_depth" and 11 for the "random_state" parameters, respectively. The inverse of regularization strength in the Logistic Regression is set to "1e1" for the "C" parameter. Furthermore, early stopping rule is selected to avoid overfitting in training Linear SVM and Logistic Regression classifiers. Other parameters of the classifiers are their default values.

Tables 2 and 3 compare the obtained micro-average and macro-average results of the classifiers without using MetaMap and with using MetaMap, respectively. The best results are highlighted in the tables. It can be concluded from the experimental results that the accuracies of all classifiers are increased significantly after applying the proposed method. In Table 2, K-Nearest Neighbor using MetaMap achieved better performance (with 94.86% accuracy) in comparison with the other classifiers. Additionally, the Linear SVM without using MetaMap achieved the highest accuracy (with 87.35%) which is less than the accuracy of the Linear SVM with using MetaMap (with 93.28%).

Table 2. The obtained Micro-average results for i2b2 2010. (Bold values indicate the best performance in each metric measure.)

Method	**without MetaMap**			**with MetaMap**		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Naive Bayes	77.47	77.47	77.47	81.42	81.42	81.42
Linear SVM	**87.35**	**87.35**	**87.35**	93.28	93.28	93.28
KNN	84.98	84.98	84.98	**94.86**	**94.86**	**94.86**
Decision Tree	85.77	85.77	85.77	90.12	90.12	90.12
Logistic Regression	86.96	86.96	86.96	92.89	92.89	92.89

Table 3. The obtained Macro-average results for i2b2 2010. (Bold values indicate the best performance in each metric measure.)

Method	**without MetaMap**			**with MetaMap**		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Naive Bayes	50.55	50.20	48.33	68.50	62.21	64.00
Linear SVM	84.44	70.66	74.67	91.07	86.28	88.41
KNN	85.33	62.01	65.08	91.92	91.24	**91.58**
Decision Tree	77.17	**74.47**	**75.67**	82.78	**91.51**	85.93
Logistic Regression	**86.39**	68.02	72.31	**92.38**	83.64	87.15

Fig. 4 shows the comparison of the classifiers accuracy without MetaMap and with MetaMap. By analyzing Fig. 4, Naive Bayes and Decision Tree classifiers

are improved approximately 4% using the proposed method. Furthermore, Linear SVM and Logistic Regression achieved 6% more precision. The biggest improvement is achieved by K-Nearest Neighbor (10%). Overall, all of the learned models by utilizing the concept of phrases instead of the original documents achieved on average a 6.1% improvement in classifying the i2b2 2010 data set. Moreover, the number of features has been reduced from 7554 to 788 by the conceptualization approach, which is about 90% reduction.

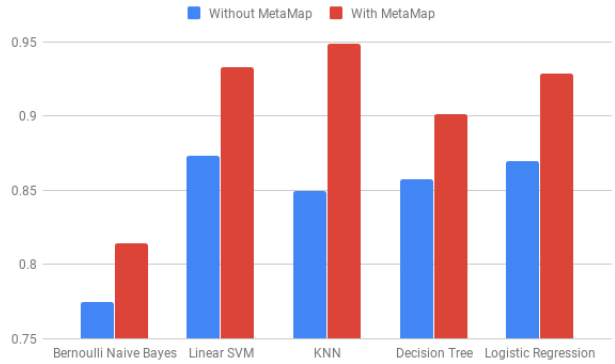

Fig. 4. Comparison of the classifiers' accuracy in test data set without MetaMap and with MetaMap.

To further evaluate our approach, instead of the original training-testing split given by the data set, we used 10-fold cross validation. we shuffle the documents and run the experiment 30 times, and each time is 10-fold cross validation. We did significance test using the experiment results of the 30 runs. Table 4 details the mean and the standard deviation of the suggested method with MetaMap and the method without MetaMap over the i2b2 data set. The classification accuracy is the average of 30 times 10-fold cross validation test. The Wilcoxon signed ranks test [20] is applied to check whether the proposed method has made significant difference in classification accuracy. According to Table 4, "T" column shows the significance test of the without MetaMap method against the suggested method, where "+" implies the proposed technique is significantly more accurate, "=" implies no significant difference, and "-" implies significantly less accurate.

From Table 4, it can be concluded that the proposed method is able to achieve considerably higher classification accuracy than the other method. Our approach gains significantly better classification accuracy in four cases. Only in the case of Decision Tree classifier, the method shows not significantly difference of classification accuracy.

Table 4. Comparison of classification accuracy and standard deviation averages using 30 independent runs. The highlighted entries are significantly better(Wilcoxon Test, $\alpha = 0.05$)

Dataset	Classifier	Without MetaMap	Highest Mean (Lowest STD)	With MetaMap	Highest Mean (Lowest STD)	T
i2b2 2010	Naive Bayes	80.49±0.055	81.34(0.036)	**84.26±0.053**	85.64(0.029)	+
	Linear SVM	88.96±0.046	89.49(0.031)	**92.56±0.038**	93.08(0.016)	+
	KNN	86.76±0.051	87.80(0.023)	**91.61±0.039**	92.82(0.028)	+
	Decision Tree	90.36±0.037	92.60(0.016)	89.14±0.042	91.39(0.029)	=
	Logistic Regression	88.51±0.047	89.02(0.027)	**92.63±0.038**	93.32(0.021)	+

Table 5. The ranking of the classifiers by Fredman's test

Classifier	Ranking
Logistic Regression	1,467
Linear SVM	1,600
KNN	2,933
Decision Tree	4,000
Naive Bayes	5,000

4.1 Further Analysis

For further analysis the methods, we checked the outputs and detected two documents with names "0101.txt" and "0302.txt" and label CAD which all the classifiers in the method without MetaMap have been labeled incorrectly, whereas all of the classifiers in the proposed method have been labeled correctly. By checking carefully the documents, we found two main reasons for this case. The first reason is that our work decreases the number of noisy data significantly. It assists classifiers to learn better. The second reason is that the new method maps phrases to their concepts which are meaningful and most of the time shorter than the original phrases. Since all the words in the documents stand alone as features, a phrase consists more than one word will lose its meaning. For example, the phrase "shortness of breath" in the method without MetaMap will be two features: "shortness" and "breath". But in our method, it will be phrase "Dyspnea" which still has meaning when it stand alone (See line 19 in Fig. 2). These reasons improve the quality of features and lead to high prediction accuracy in the proposed method.

To more analysis the classifiers, the Friedman test [20] is applied to check whether the proposed method has made significant difference in classification accuracy. Furthermore, the Holm procedure [20] is utilized for pairwise comparisons of the used classifiers. Table 5 shows the ranking of the classifiers by Friedman's test where smaller implies better. It is obvious that Logistic Regression is the best classifier among others. Additionally, Table 6 shows the pairwise comparison by Holm's test. There are significantly differences between classifiers of all of the highlighted rows except the last row. For example, based on first row, Logistic Regression classifier is significantly better than Nave Bayes. From last row of Table 6, Logistic Regression is not significantly better than linear SVM. In summary, Logistic Regression and Linear SVM are the most suitable classifiers for the target binary classification.

Table 6. Pairwise comparison by Holm's test (Bold values indicate the significant difference.)

Classifiers	Holm
Naive Bayes vs. Logistic Regression	**0.0050**
Naive Bayes vs. Linear SVM	**0.0056**
Decision Tree vs. Logistic Regression	**0,0063**
Linear SVM vs. Decision Tree	**0.0071**
Naive Bayes vs. KNN	**0.0083**
KNN vs. Logistic Regression	**0.0100**
Linear SVM vs. KNN	**0.0125**
KNN vs. Decision Tree	**0.0167**
Naive Bayes vs. Decision Tree	**0.0250**
Linear SVM vs. Logistic Regression	0.0500

5 Conclusion and Future Work

The current study proposed a medical ontology driven feature engineering approach to reduce the number of features as well as persist with meaningful features. In conjunction with the MetaMap tool, we map meaningful phrases in medical text to specific UMLS medical concepts. The related concepts to the problem domain are selected as features. The number of features is reduced significantly by selecting "Disease or Syndrome" and "Sign or Symptom" concepts, which are the most important in the domain of clinical notes. Experimental and statistical results show that the suggested approach can accomplish significantly better classification accuracy.

As our future work, we will consider relations between diseases and symptoms, and include the ones that are interconnected as pairs [21]. Furthermore, we are planning to use concepts of sentences instead of phrases as features, hopefully to further reduce the number of features and increase the accuracy. We will find temporal relations between events to increase the classification accuracy. Finally, all of the suggested ideas will apply on other data sets for further analysis.

References

1. Yetisgen-Yildiz, M., Pratt, W.: The effect of feature representation on MEDLINE document classification. In: AMIA annual symposium proceedings, pp. 849–853. American Medical Informatics Association (2005)
2. Rak, R., Kurgan, L. A., Reformat, M.: Multilabel associative classification categorization of MEDLINE articles into MeSH keywords. IEEE engineering in medicine and biology magazine, **26**(2), 47–55 (2007)
3. Yi, K., Beheshti, J.: A hidden Markov model-based text classification of medical documents. Journal of Information Science, **35**(1), 67–81 (2009)
4. Camous, F., Blott, S., Smeaton, A. F.: Ontology-based MEDLINE document classification. In: Bioinformatics Research and Development, pp. 439–452. Springer, Berlin, Heidelberg (2007)
5. Dollah, R. B., Aono, M.: Ontology based approach for classifying biomedical text abstracts. International Journal of Data Engineering (IJDE), **2**(1), 1–15 (2011)
6. Uysal, A. K., Gunal, S.: Text classification using genetic algorithm oriented latent semantic features. Expert Systems with Applications, **41**(13), 5938–5947 (2014)
7. Parlak, B., Uysal, A. K.: Classification of medical documents according to diseases. In: 23th Signal Processing and Communications Applications Conference (SIU), pp. 1635–1638. IEEE (2015)

8. Yepes, A. J. J., Plaza, L., Carrillo-de-Albornoz, J., Mork, J. G., Aronson, A. R.: Feature engineering for MEDLINE citation categorization with MeSH. BMC bioinformatics, **16**(1), 113–124 (2015)
9. Parlak, B., Uysal, A. K.: On Feature Weighting and Selection for Medical Document Classification. In: Developments and Advances in Intelligent Systems and Applications, pp. 269–282. Springer, Cham, (2018)
10. Garla, V. N., Brandt, C.: Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. Journal of the American Medical Informatics Association, **20**(5), 882–886 (2012)
11. Shah, F. P., Patel, V.: A review on feature selection and feature extraction for text classification. In: Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on, pp. 2264–2268. IEEE (2016)
12. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), **34**(1), 1–47 (2002)
13. Ko, Y.: A study of term weighting schemes using class information for text classification. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 1029–1030. ACM (2012)
14. Lewis, D. D.: Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on Speech and Natural Language, pp. 212–217. Association for Computational Linguistics, (1992)
15. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection, Journal of machine learning research, **3**(Mar), 1157-1182 (2003)
16. Tran, B., Xue, B., Zhang, M.: Genetic programming for feature construction and selection in classification on high-dimensional data. Memetic Computing, **8**(1), 3-15 (2016)
17. Tang, J., Alelyani, S., Liu, H.: Feature selection for classification: A review. Data Classification: Algorithms and Applications, pp. 37. (2014)
18. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering, **40**(1), 16-28 (2014)
19. Unified Medical Language System (UMLS®), http://www.nlm.nih.gov/research/umls/. Last updated 20 Apr 2016
20. Demar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, **7**, 1-30 (2006)
21. Ernst, P., Siu, A., Weikum, G.: Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC bioinformatics, **16**(1), 157–169 (2015)
22. Khalifa, A., Meystre, S.: Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. Journal of biomedical informatics, **58**, S128–S132 (2015)
23. Jonnagaddala, J., Liaw, S. T., Ray, P., Kumar, M., Chang, N. W., Dai, H. J.: Coronary artery disease risk assessment from unstructured electronic health records using text mining. Journal of biomedical informatics, **58**, S203–S210 (2015)
24. Yang, H., Garibaldi, J. M.: A hybrid model for automatic identification of risk factors for heart disease. Journal of biomedical informatics, **58**, S171–S182 (2015)
25. Shivade, C., Malewadkar, P., Fosler-Lussier, E., Lai, A. M.: Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. Journal of biomedical informatics, **58**, S103–S110 (2015)
26. Buchan, K., Filannino, M., Uzuner, .: Automatic prediction of coronary artery disease from clinical narratives. Journal of biomedical informatics, **72**, 23–32 (2017)