


Ontology-guided data augmentation for medical document classification

Mahdi Abdollahi¹[0000-0002-6115-1342], Xiaoying Gao¹[0000-0002-6326-7947], Yi Mei¹[0000-0003-0682-1363], Shameek Ghosh², and Jinyan Li³[0000-0003-1833-7413]

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. Extracting meaningful features from unstructured text is one of the most challenging tasks in medical document classification. The various domain specific expressions and synonyms in the clinical discharge notes make it more challenging to analyse them. The case becomes worse for short texts such as abstract documents. These challenges can lead to poor classification accuracy. As the medical input data is often not enough in the real world, in this work a novel ontology-guided method is proposed for data augmentation to enrich input data. Then, three different deep learning methods are employed to analyse the performance of the suggested approach for classification. The experimental results show that the suggested approach achieved substantial improvement in the targeted medical documents classification.

Keywords: Ontology · Data Augmentation · Medical Document Classification.

1 Introduction

Medical document classification is different from the commonly considered document classification in terms of text terminology and their repetitiveness. In medical document classification, the content explains a set of medical events in a discharge note, with the objective of providing a clarification as accurately and comprehensively as conceivable when explaining the health condition of a patient. Mainly, such text massively uses domain-specific vocabulary and acronyms, making medical note analysis significantly different from commonly considered document classification. In addition, different combinations of domain-specific clinical events in a medical discharge note can explain a patient’s health status completely differently. Hence, extracting important information to analyze clinical documents is exceptionally imperative.

One of the important factors which has effect on the classification accuracy is the size of the data set for training the model. Generally, there is a lack of adequate data in medical area [1]. When the training data set is not big enough,

the trained classification model has not sufficient instances to learn. Hence, the prediction of the classifier will not be satisfactory. This issue can be worse when the data set has not enough text inside of the documents such as document abstracts. One possible solution to address the issue is to augment data for training the model.

Data augmentation is a methodology that empowers experts to fundamentally build the assorted variety of data accessible for training models, without really gathering new data. Data augmentation has many applications in image classification, sound and speech classification [2]. But there is not much work for text. In terms of text, it is not appropriate to augment the text by utilizing signal transformations as commonly used in image or speech classification. Because the order of words in text is important and may has semantic meaning. Hence, the best approach for doing data augmentation is to paraphrase the sentences in the documents by human. But this is very expensive due to the large size of instances in the data set. Replacing words and expressions with their synonyms can be a reasonable choice in data augmentation [3]. However, these methods are using normal dictionaries for augmentation and some domain specific terms or acronyms do not have synonyms in normal dictionaries.

As there are domain-specific vocabulary and acronyms in medical discharge notes, finding synonyms is not trival and this requires domain knowledge. In this paper, an ontology-based method is introduced for data augmentation by targeting concepts of words and expressions in the documents. This method will replace all of the words and phrases with their scientific names if they belong to a concept in medical field. This paper plans to study the following research questions:

1. Whether the ontology-guided approach can produce new discriminative instances from the original document set; and
2. Whether the proposed method can improve the classification accuracy in the targeted medical documents classification task.

2 Related work

2.1 Data augmentation in classification

Data augmentation is a technique to deal with data scarcity in training models for different tasks such as classification. There are some common methods such as adding spelling errors, paraphrasing by utilizing syntax trees or regular expressions, adding textual noise and replacing with synonyms. Among these methods, synonyms replacement is one of the common approaches in textual data augmentation.

Zhang et al. [3] applied data augmentation in Convolutional Neural Network (CNN) for text classification by utilizing English thesaurus obtained from WordNet. They replaced the words and expressions with their synonyms in the text to make new text based on the main data set. Rosario [2] introduced a method to data augmentation for short texts classification by producing similar words for each short texts to make a longer text by considering a semantic space. Quijas has investigated the effect of data augmentation in training CNNs and RNNs

for text classification [4]. Kobayashi has suggested “contextual augmentation” method which produces counterparts of words by using a bidirectional language model and replaces words with their counterparts in sentences. They examined the method on different data set and showed improvements [5]. Coulombe in [6] has introduced another textual data augmentation by applying different methods including paraphrase generation, spelling errors, textual noise, back-translation and synonyms replacement. The methods were tested on different neural network architectures. Jungiewicz has proposed an approach to textual data augmentation for training CNNs by applying on sentence classification task. The researcher transformed sentences by keeping their lengths the same as their original lengths. The author has employed a thesaurus which belongs to Princeton University’s WordNet [7]. However, these methods are using normal dictionaries for augmentation and some domain specific terms or acronyms do not have synonyms in normal dictionaries.

2.2 Feature extraction in medical document classification

Shah and Patel have used statistical approaches from features distribution in document classification to rank features [8]. The introduced methods used information gain (IG), mutual information, word frequency and term frequency-inverse document frequency (tf-idf) metrics for textual feature extraction. Nevertheless, these methods weight each feature separately without considering the relationship between features. Ontology-based classification methods is introduced in [9]. Dollah and Aono have introduced ontology-based classification approaches for biomedical abstract text classification [9]. Authors in [10–13] utilize different ontologies such as Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine (SNOMED) and Medical Subject Headings (MeSH) to increase text classification accuracy.

Medical documents have been utilized in different tasks such as analyzing Framingham risk score (FRF), assessing risk factors in diabetic patients, discriminating heart disease risk factors, and finding risk factors for heart disease patients [14]. In this paper, we employ ontology as a feature extraction approach to detect meaningful words and expressions for augmenting documents.

3 Our ontology-based method

In this section, we illustrate a novel data augmentation method and the utilized tools for extracting concepts of words and expressions for producing new documents. The suggested approach targets concepts of words and expressions to replace them with their scientific names. Fig. 1 shows the flowchart of the suggested ontology-based approach for data augmentation.

The input of the suggested system is a set of clinical documents. Firstly, the method parses each document and tokenize the context based on sentences. Then, MetaMap tool [15] is employed to detect the meaningful phrases and their concepts in each sentence from the Unified Medical Language System (UMLS). After finding the phrases with a concept, the scientific name of the detected words or expressions are used to replace their corresponding phrases in the sentence. All of the new documents are created by applying the method. Next, all

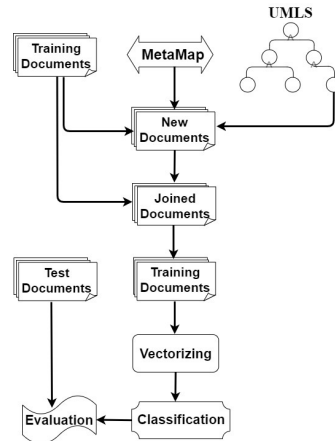


Fig. 1. The proposed data augmentation for medical document classification

of the features are extracted from the original data set and the new created data set. Then, three different neural network approaches including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Hierarchical Attention Network (HAN) are employed for classification. The output predicts the label of a document.

It is expected that the suggested approach which produces meaningful documents and keeps their class name based on the original documents, can enhance the classification accuracy.

3.1 Data augmentation method

There are many domain-specific words and expressions in medical document and data augmentation requires domain knowledge. In this section, an ontology-guided approach is introduced for short text augmentation as a preprocessing stage.

UMLS is a domain-specific dictionary in the biomedical field. It provides an ontology structure of medical terminology concepts. In the suggested approach ("SciName"), each document in the data set (D) is analyzed independently. Firstly, the x th document (D_x) is tokenized based on the sentences (S). Then, the i th sentence (S_i) is sent to the UMLS by using the MetaMap tool. MetaMap extracts all of the concepts of the detected meaningful expressions in S_i from the UMLS. Next, all of the detected phrases are replaced with their extracted scientific names from the UMLS. Finally, S_i is updated in D_x . This process is repeated on all of the sentences of documents to make new documents.

A document segment is given below to illustrate how MetaMap works on the input medical documents and what output it returns in the data augmentation process. The following is a sample of a clinical note.

"Early resistance to pathogens requires a swift response from nk cells. In largeint giorgio trinchieri identified an nk growth factor and activator later called inter-

leukin 12 il 12. This discovery helped reveal the regulatory link between innate and adaptive immunity.”

```

-----
1 Phrase: Early resistance to pathogens
2 >>>> Phrase
3 early resistance to pathogens
4 <<<< Phrase
5 >>>> Mappings
6 Meta Mapping (834):
7 570 Pathogen (Pathogenic organism) [Organism]
8 <<<< Mappings
-----
9 Phrase: a swift response from nk cells.
10 >>>> Phrase
11 a swift response from nk cells
12 <<<< Phrase
13 >>>> Mappings
14 Meta Mapping (719):
15 586 Swift (Family Apodidae) [Bird]
16 753 Response (Response process) [Organism Attribute]
17 623 NK cells (Natural Killer cells) [Cell]
18 <<<< Mappings
-----
19 Phrase: an nk growth factor
20 >>>> Phrase
21 nk growth factor
22 <<<< Phrase
23 >>>> Mappings
24 Meta Mapping (901):
25 660 NK (Natural Killer Cells) [Cell]
26 <<<< Mappings
-----
27 Phrase: activator later called interleukin 12 il 12.
28 >>>> Phrase
29 activator later called interleukin 12 il 12
30 <<<< Phrase
31 >>>> Mappings
32 Meta Mapping (745):
33 595 call (Decision) [Mental Process]
34 795 IL (Illinois (geographic location)) [Geographic Area]
35 <<<< Mappings
-----
36 Phrase: helped
37 >>>> Phrase
38 helped
39 <<<< Phrase
40 >>>> Mappings
41 Meta Mapping (966):
42 966 Help (Assisted (qualifier value)) [Qualitative Concept]
43 <<<< Mappings
-----
44 Phrase: the regulatory link between innate
45 >>>> Phrase
46 the regulatory link between innate
47 <<<< Phrase
48 >>>> Mappings
49 Meta Mapping (695):
50 593 regulatory [Regulation or Law]
51 760 Link (Links List) [Intellectual Product]
52 <<<< Mappings
-----

```

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

Fig. 2 shows the output of the MetaMap for the sample document. Table 1 presents the detected expressions with their concepts and scientific names for each phrase of the sample document. The concepts and scientific names of each detected phrase in the table is extracted by analyzing the lines 7, 15, 16, 17 for the first sentence, lines 25, 33, 34 for the second sentence and lines 42, 50 and 51 for the third sentence (in Fig. 2). Firstly, the phrase appeared in square brackets is extracted as a concept of the detected expression in the sentence. Then, the phrase appeared within the round parentheses at the same line is extracted as a scientific name of the detected expression. Finally, the extracted scientific name is used to replace the original expression in the sentence. This process is applied on all of the three sentences of the sample note. Below is the final output of the proposed method for the example clinical note.

“Early resistance to pathogenic organisms requires a family apodidae response process from natural killer cells. In largeint giorgio trinchieri identified an nat-

Table 1. The detected phrases of the example notes using MetaMap.

Sentences	Detected Phrases	Extracted Concepts	Replaced Phrases
First Sentence	pathogens	[Organism]	Pathogenic organism
	swift	[Bird]	Family Apodidae
	response	[Organism Attribute]	Response process
	nk cells	[Cell]	Natural Killer Cells
Second Sentence	nk	[Cell]	Natural Killer Cells
	call	[Mental Process]	Decision
	il	[Geographic Area]	Illinois (geographic location)
Third Sentence	help	[Qualitative Concept]	Assisted (qualifier value)
	regulatory	[Regulation or Law]	regulatory
	link	[Intellectual Product]	Links List

ural killer cells growth factor and activator later decisioned interleukin 12 illinois (geographic location) 12. This discovery assisted (qualifier value) reveal the regulator links list between innate and adaptive immunity.”

The length of the output is longer than the input due to the more specific knowledge provided by the UMLS. For example, the acronym "nk" is changed to "natural killer" and the acronym "il" is replaced with "illinois (geographic location)". The proposed method can easily provide more informative knowledge by using the UMLS. Finally, the new produced documents are used for the training stage together with the original documents to improve the performance of medical document classification.

4 Experiment design

4.1 Classification methods

In the ontology-based data augmentation, the new documents are made and mixed with the original documents to use for classification. We use three deep learning (DL) models, including a convolutional neural network (CNN), a recurrent neural network (RNN), and a hierarchical attention network (HAN) [16]. The performance is calculated by evaluating macro F1-measure metric for all of the used ML methods:

$$F1 \text{ measure} = \frac{1}{N} \sum_{i=1}^N 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \quad (1)$$

where N indicates the number of classes. Word2Vec word embedding is used to represent word tokens into numerical vectors. Word embedding represents the semantic meaning of each word in a numerical vector form. Word2Vec makes word embedding by utilizing a feed-forward neural network to anticipate the vicinage words for an input word. The word embedding is trained on all documents and transformed each word to its corresponding embedding. Then, the learned word embedding is used to generate the input for CNN, RNN and HAN. The size of word embedding is 350.

4.2 Dataset and preprocessing

The performance of the suggested ontology-guided data augmentation is evaluated on the 2010 Informatics for Integrating Biology and the Bedside (i2b2),

CAD (Coronary Artery Disease) task of the 2008 Informatics for Integrating Biology and the Bedside (i2b2) and the PubMed data set. The labels of i2b2(2008) and i2b2(2010) data set are CAD and non-CAD that form an imbalanced binary classification task. The numbers of CAD instances for i2b2(2008) training and testing sets are 391 and 272 documents, respectively. The numbers of CAD instances for i2b2(2010) training and testing sets are 25 and 48 documents, respectively. The labels of the PubMed data set are metabolism, physiology, genetics, chemistry, pathology, surgery, psychology, and diagnosis. The data set includes 8000 documents, each class has 1000 documents with 70% of documents for training and 30% for testing. The size of the training sets of all data sets will double by adding the new produced documents to the original ones.

The i2b2(2008) data set has 20701 different terms. It contains 1113 documents which 656 documents for training and 457 documents for testing. The i2b2(2010) data set has 7481 different terms. It contains 426 documents which 170 documents for training and 256 documents for testing. The PubMed data set has 30178 various terms. The number of the train documents is increased to 16000 in PubMed, 2226 in i2b2(2008) and 852 in i2b2(2010) by adding the new augmented documents. The 2223 input documents contain 27248 various terms. The 852 input documents contain 14920 various terms. The 16000 input documents include 59151 different terms.

4.3 Parameter Settings

Three different ML methods are used to evaluate the proposed idea. The suggested parameters in [16] are used for the employed neural network approaches. The early stopping approach by considering the validation accuracy (three epochs without any change) is used to terminate the training step.

The used CNN architecture has 3 parallel convolutional layers with 100 channels for each one. The window of layers are 3, 4, and 5 words, respectively. The output of this architecture for each input document is 300 channels \times number of words. The applied dropout rate is 50% [16].

HAN [16] is a deep learning model developed for document classification. It contains two hierarchies. The lower hierarchy analyzes a line in word level and it feeds with a word embedding. Then, it uses a bidirectional GRU to apply an attention mechanism to find more important words. The output is a line embedding which is feed to the upper hierarchy to analyze a document in line level. A dropout is applied on the produced document embedding and finally, a softmax function is employed to predict a label of each document.

The RNN architecture uses an attention mechanism (which is similar to a single hierarchy of HAN method). A bidirectional GRU with attention and 200 number of neurons is utilized with dropout and softmax. The used optimizer is Adam with learning rate of 0.0002. The applied dropout rate is 50%.

5 Results and discussion

The performance of the methods are evaluated based on macro F1-measure and accuracy metrics for the PubMed data set and macro F1-measure metric for the i2b2(2010) data set.

Three different approaches are applied on the original documents in each data set. In first approach(SynName), we used WordNet dictionary to extract all of the synonyms of the main word appeared inside of a document. Then, the most similar synonym is found by using the GloVe pre-trained model(from the GloVe website <http://nlp.stanford.edu/data/glove.6B.zip>) and used to replace the main word to augment new documents. The used GloVe model provides 100-dimensional vector which is trained on Wikipedia data with 6 billion tokens and a 400,000 word vocabulary. In second approach(our proposed method), UMLS is employed to find scientific names of the appeared phrases in the documents based on their concepts to replace with the original phrase in the document to augment a new document. In third approach, we combined the augmented document from the two introduced approaches with the original data set. The proposed augmentation method(SciName) is applied on the training set only. Then, experimental results are calculated using 30 independent runs on the original test set.

Tables 2-5 compare the statistical results for three methods. The average and standard deviation of accuracies and F1-measure are provided for each ML method and the significance test is done utilizing the experiment results of the 30 runs to compute the three approaches. The Wilcoxon signed ranks test with significance level of 0.05 is used to assess whether the suggested approach has made significant difference in classification performance. In tables 2-5, "T" column indicates the significance test of the best approach(the third combined approach) against the other methods, where "+" indicates the suggested method is significantly better, "=" mentions no significant difference, and "-" points significantly less accurate. The best results are highlighted in the tables.

By analyzing tables 2 and 3, it is clear that neural network methods are improved in accuracy and F1-measure by using combination of the original data set with the obtained augmented data sets from the two introduced augmentation methods. The highest accuracy and F1-measure in tables 2 and 3 belong to RNN. Tables 4 and 5 provide the statistical results for i2b2(2010) and i2b2(2008) data sets, respectively. In table 4, RNN shows high F1-measure in the method which is using most similar synonyms to augment new data and HAN indicates good performance with combination approach(SynName+SciName). The highest F1-measure in both data sets belongs to RNN with 97.97% and 98.22%, respectively. In Tables 2, 3 and 5, the suggested ontology-based approach(SciName) shows better performance in comparison with the SynName method [3].

5.1 The value of the work

As indicated in this paper, in numerous practical works on modeling, data augmentation is extremely important. This is a situation that we encounter when in practical settings, real life patient cases are unavailable to feed data-hungry models (a rare disease is an example where available cases are few). In fact, synthetic data synthesis and augmentation has strong advantages with respect to advancing healthcare models research by protecting patient confidentiality, and is a promising tool for situations where real world data is difficult to obtain or

unnecessary. At that time, in combination with data augmentation we can also perform simulations to generate digital patient cases.

Table 2. Comparison of classification accuracy and standard deviation averages using 30 independent runs for PubMed data set. The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original		SynName		SciName		SynName+SciName	
Classifiers	Accuracy		Accuracy		Accuracy		Accuracy	
	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN	71.64±0.55 (72.67)	+	80.64±0.63 (81.84)	+	80.80±0.51 (82.08)	+	84.16±0.66 (85.42)	
RNN	71.53±1.03 (73.50)	+	84.42±0.90 (86.38)	+	85.57±0.62 (86.75)	+	90.80±0.45 (91.63)	
HAN	71.29±0.69 (72.88)	+	84.29±0.85 (85.38)	+	85.00±0.94 (86.92)	+	90.75±0.49 (91.79)	

Table 3. Comparison of classification F1-measure and standard deviation averages using 30 independent runs for PubMed data set. The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original		SynName		SciName		SynName+SciName	
Classifiers	F1-measure		F1-measure		F1-measure		F1-measure	
	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN	71.54±0.76 (72.64)	+	80.42±0.69 (81.42)	+	80.48±0.71 (81.81)	+	84.34±0.54 (85.36)	
RNN	71.62±0.73 (72.97)	+	84.07±0.88 (85.64)	+	85.37±0.90 (87.00)	+	90.91±0.58 (91.98)	
HAN	70.96±0.82 (72.21)	+	84.21±1.23 (86.16)	+	84.95±0.73 (86.36)	+	90.85±0.79 (91.99)	

Table 4. Comparison of classification F1-measure and standard deviation averages using 30 independent runs for i2b2 2010 data set. The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original		SynName		SciName		SynName+SciName	
Classifiers	F1-measure		F1-measure		F1-measure		F1-measure	
	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN	81.83±11.19 (95.62)	+	94.40±2.20 (96.84)	-	93.41±2.12 (97.10)	-	94.38±2.25 (97.20)	
RNN	88.48±10.36 (96.33)	+	97.97±1.61 (99.24)	-	93.50±1.90 (96.41)	-	96.96±1.27 (98.58)	
HAN	86.60±11.20 (93.88)	+	94.40±3.75 (99.24)	+	92.72±1.79 (95.73)	+	96.55±1.08 (98.60)	

Table 5. Comparison of classification F1-measure and standard deviation averages using 30 independent runs for i2b2 2008 data set(CAD Task). The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original		SynName		SciName		SynName+SciName	
Classifiers	F1-measure		F1-measure		F1-measure		F1-measure	
	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN	92.34±1.27 (94.94)	+	93.85±1.77 (95.72)	+	94.69±0.47 (95.57)	=	94.69±1.50 (96.82)	
RNN	95.72±0.86 (96.81)	+	98.20±0.49 (98.65)	+	96.69±0.38 (97.46)	+	98.22±0.35 (99.09)	
HAN	95.49±1.05 (97.27)	+	96.46±1.59 (97.75)	+	96.72±0.43 (98.15)	+	98.08±0.41 (99.09)	

6 Conclusions and Future Work

This paper proposes a new ontology-based data augmentation method by replacing meaningful expressions with their scientific names to deal with the data shortage issue in medical document classification. The introduced approach is able to improve the precision of classification in the neural network models. Experimental results for accuracy and f1-measure show that the suggested method can increase the performance of the CNN, RNN and HAN models by using the suggested ontology-based approach to provide more samples in the training phase. This paper shows promise in utilizing an ontology-guided data augmentation approach in clinical document classification, however, it is still necessary to do more research to improve the classification performance. We will explore

other ways to do data augmentation for medical discharge notes. Meanwhile, we will investigate different combination of data augmentation methods to enhance the classification precision.

References

1. Sánchez, D., Batet, M. & Viejo, A. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics* **52**, 189–198 (2014).
2. Rosario, R. R. *A Data Augmentation Approach to Short Text Classification* PhD thesis (UCLA, 2017).
3. Zhang, X., Zhao, J. & LeCun, Y. *Character-level convolutional networks for text classification* in *Advances in neural information processing systems* (2015), 649–657.
4. Quijas, J. K. *Analysing the effects of data augmentation and free parameters for text classification with recurrent convolutional neural networks* (The University of Texas at El Paso, 2017).
5. Kobayashi, S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201* (2018).
6. Coulombe, C. Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs. *arXiv preprint arXiv:1812.04718* (2018).
7. Jungiewicz, M. & Smywiński-Pohl, A. Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science* **20** (2019).
8. Shah, F. P. & Patel, V. *A review on feature selection and feature extraction for text classification* in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (2016), 2264–2268.
9. Dollah, R. B. & Aono, M. Ontology based approach for classifying biomedical text abstracts. *Int. J. Data Eng* **2**, 1–15 (2011).
10. Buchan, K., Filannino, M. & Uzuner, Ö. Automatic prediction of coronary artery disease from clinical narratives. *Journal of biomedical informatics* **72**, 23–32 (2017).
11. Abdollahi, M., Gao, X., Mei, Y., Ghosh, S. & Li, J. *Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes* in *Australasian Joint Conference on Artificial Intelligence* (2018), 104–110.
12. Abdollahi, M., Gao, X., Mei, Y., Ghosh, S. & Li, J. *An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation* in *2019 IEEE Congress on Evolutionary Computation(CEC)* (2019), 1–8.
13. Abdollahi, M., Gao, X., Mei, Y., Ghosh, S. & Li, J. *Stratifying Risk of Coronary Artery Disease Using Discriminative Knowledge-Guided Medical Concept Pairings from Clinical Notes* in *Pacific Rim International Conference on Artificial Intelligence* (2019), 457–473.
14. 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).
15. Aronson, A. R. & Lang, F.-M. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* **17**, 229–236 (2010).
16. Gao, S. *et al.* Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association* **25**, 321–330 (2017).

A Appendix

In this section, we test the proposed ontology-based approach in two settings: combined and mix. In the combined one, context of each augmented document is joined with the context of its corresponding original document. In the mix one, the augmented data set is added to the original data set which each original and augmented document stands alone without any combination. Based on these scenarios, six different cases are analyzed by the suggested approach as follows:

- Case 1: the original data set is used without any augmentation. (5600 Training/2400 testing)
- Case 2: the new augmented train set is concatenated in context with the original train set, but the original test set is used without any changes. (5600 Training/2400 testing)
- Case 3: the new augmented train set is mixed with the original train set, but the original test set is used without any changes. (11200 Training/2400 testing)

Cases 1, 2 and 3 are checking the performance of the proposed idea on the original test set without any changes. Tables 6 and 7 show the performance of the applied new data augmentation approach for PubMed data set on the three different mentioned cases based on classification accuracy and macro f1-measure metrics, respectively.

Table 6. Accuracy results on PubMed data set

Method	Case 1	Case 2	Case 3
CNN	71.54	70.46	80.80
RNN	71.62	74.88	85.57
HAN	70.96	74.46	85.00

Table 7. F1-measure results on PubMed data set

Method	Case 1	Case 2	Case 3
CNN	69.65	68.42	80.48
RNN	73.07	73.94	85.37
HAN	72.43	73.73	84.95

The results indicate that the performance of ML methods is improved in cases 3 in comparison with the original data set. It shows better performance by applying data augmentation to increase the train set size. HAN and RNN methods show better performance than CNN. On the other hand, RNN has the highest precision in comparison with HAN and CNN due to using long short term memory to remember relation between words. As case 3 uses the original data set for testing and show better performance for all the used methods, it is applied on all of the data sets to calculate the experimental results in the tables 2, 3, 4 and 5.