Ontology-based Enriched Concept Graphs for Medical Document Classification

Niloofer Shanavas[a,*], Hui Wang[a], Zhiwei Lin[a], Glenn Hawe[a]

[a] *Ulster University, School of Computing, Jordanstown, BT370QB, United Kingdom*

Abstract

The rapidly increasing volume of medical text data, including biomedical literature and clinical records, presents difficulties to biomedical researchers and clinical practitioners. Automatic text classification is an important means for managing medical text data. The main challenge in medical text classification is the complex terminology used in these documents. Therefore, it is critical to handle synonymy, polysemy, and multi-word concepts so that classification is based on the meaning of these documents. The solution to this problem of complex terminology helps in building systems with better access to relevant data, resulting in more effective utilisation of the existing information. In this paper, we present a simple and effective approach to address this challenge. A concept graph is automatically constructed and enriched for each medical text document with the help of a domain-specific similarity matrix that is built using Unified Medical Language System (UMLS) concepts in the training documents. Medical text documents are compared based on their enriched concept graphs using a graph kernel. Classification is then done based on the comparison result. The benefit of this approach is that it allows the incorporation of domain knowledge into the classification framework. The experiments on biomedical abstracts and clinical reports classification show the effectiveness of the proposed approach. Based on evaluation metrics of precision, recall and F1-scores, our method achieves a significantly higher classification performance than other widely used similarity measures for similarity-based text classification.

Keywords: Medical text classification, SVM, Graph-based text

*Corresponding author: Tel.: +44-7586750660;
Email address: shanavas-n@ulster.ac.uk (Niloofer Shanavas)

representation, UMLS, Similarity measure, Graph kernel

1. Introduction

As the volume of biomedical literature and clinical records continues to grow quickly, the complexity of medical terminology is increasing. Biomedical literature reports new research discoveries and theories. Clinical records contain data about the patients' symptoms, family history, diagnosis, treatment and medication. These documents serve as critical information for clinical decision making. Due to the rapid growth of medical data, efficient tools are needed to discover knowledge from the huge data and put them into use for the diagnosis, treatment and prevention of diseases [1].

Text mining solves the overload problem that results from abundant medical information. Automatic text classification, which is a text mining task, assigns class labels to documents based on their content. Medical text classification helps in organising the growing amount of medical data, making it easier to locate and extract the relevant data. It assists biomedical researchers and clinicians to make practical use of the findings by helping them easily find the relevant information hidden within the large amount of data. The main challenges in medical text classification are the identification of medical entities, and handling synonyms and polysemous words to classify the documents accurately [2].

An important and challenging part of text classification is the effective representation of text. The bag-of-words approach is the most commonly used text representation scheme. It is based on the term independence assumption and considers a document as a set of independent terms. The similarity computed between text documents is usually based on the exact matching of terms in the documents. Hence, it does not handle synonyms and polysemous words, thus resulting in an inaccurate similarity value. In this paper, we address this challenge and present a graph-based method to represent medical documents for the accurate calculation of similarity between medical text documents that improves the performance of text classification. Graphs are powerful mathematical constructs that are capable of modelling complex data. Our proposed method automatically constructs an enriched concept graph for each medical document from the initial set of concepts identified. Each node in a graph represents a unique medical concept, whilst each edge represents an association between related concepts.

Our approach utilises the hierarchical relationship in the UMLS semantic network to add related concepts, link the associated concepts and compute the weight of nodes/edges. The similarity between any two concept graphs is then calculated using a graph kernel. Experiments involving the classification of biomedical abstracts and clinical records are performed. Our evaluation shows that the proposed graph-based method for calculating similarity significantly outperforms other commonly used similarity measures.

The rest of the paper is organised as follows. Section 2 gives an overview of the related works. Section 3 introduces the proposed method for automatic construction of enriched concept graphs using domain knowledge from UMLS. Section 4 describes the calculation of similarity between the enriched concept graphs using a graph kernel and Section 5 presents the system architecture. Section 6 presents the experiments and results. Finally, Section 7 concludes the paper.

2. Related Works

In this section, we focus on works that utilize a graph-based representation of text and then classify documents based on the similarity value computed between the text graphs, and on semantic kernels that incorporate knowledge into the text classification framework. We also provide a review of text classification applications in the medical domain.

In a graph-based text representation for text classification, nodes usually correspond to terms in text and the edges can denote syntactic relations, semantic relations and statistical relations such as word co-occurrences. The graph-based text representation can then be classified without explicitly converting it into vectors by graph matching [3] [4] or by using graph kernels [5] [6] [7].

The graph distance measure based on maximum common subgraph can be used for calculating the similarity between documents represented as graphs [3] [4]. Schenker *et al.* [3] extended the k-NN classifier to classify web documents represented as graphs where each node represents a unique term and a labeled directed edge links two adjacent terms. A graph-based distance measure based on maximum common subgraph is used to compute the similarity between graphs. Although the classification performance increased when the graph size was set to a larger number of nodes, it resulted in an increase in the time complexity. Wu *et al.* [4] built a co-occurrence graph model that

considers structural information to represent text where each term is represented by a node; the edges denote co-occurrence relationships and the edge weights indicate the strength of the relationship. The edge weights increase as the number of times the terms that appear together increases. The graph similarity is calculated using a maximum common subgraph based similarity measure that considers the contribution of common nodes, common edges and weights. The average similarity between the document to be classified and the text documents in each class is computed to find the class that the document belongs to. The disadvantage with this approach is the increase in time complexity with increase in the size of the graph.

Graph kernels help in comparing graphs by decomposing it into substructures (such as nodes, edges, random walks, shortest path, cycles, subtrees) and computing the similarity between these substructures. In [5], the semantic information in text documents is represented as Discourse Representation Structures (DRS) from Discourse Representation Theory and then classified using the direct product kernel with a SVM classifier. The construction of the semantic representation is more time consuming than the bag-of-words approach, but there is no significant difference in the classification performance. In [6], the biomedical documents are represented as concept graphs where nodes correspond to biomedical concepts in the UMLS database and edges denote semantic relationship between the concepts. The weighted concept graphs are classified using a set kernel and a simple linear kernel. The set kernel measures similarity between graphs based on the number of shared edges. The linear kernel based similarity is the cosine similarity between the edge weight vectors of a pair of graphs. Kernel functions were used with both SVM and k-NN classifier, and the set-based-kernel SVM classifier outperformed the bag-of-words with tf-idf weighting approach for biomedical document classification. Nikolentzos *et al.* [7] defines a document similarity measure based on a graph kernel between graph-based representation of documents. A modified shortest path graph kernel is proposed for the comparison of a pair of documents. In their graph-based representation, nodes correspond to words and edges connect nodes that have the shortest distance less than a particular threshold d. The similarity value computed with shortest distance based graph kernel is based on the number of matching nodes/terms and the sum of the products of the labels of matching edges. The edge label is the inverse of the shortest distance between the nodes that the edge connects.

Walk-based kernels that are the products of node kernels have been pro-

posed that captures semantic similarity between words using word embeddings. The approach in [8] considers both syntactic and semantic similarity through a random walk-based kernel. It uses word embeddings (SENNA) to represent words and extends beyond label matching. In [9], a convolution sentence kernel based on word2vec embeddings is proposed. They smooth the delta word kernel to capture the semantic similarity of words. The similarity between sentences is obtained by combining the similarity of all the phrases. Although these approaches go beyond label matching, there is a high computational cost due to the calculation of distance between all possible pairs of words in the sentences.

The information in knowledge bases such as Wordnet and Wikipedia can be utilized to improve the performance of text classification. Siolas *et al.* [10] introduced semantic smoothing by incorporating a-priori knowledge from Wordnet into text classification. The semantic smoothing of tf-idf feature vectors is performed using a smoothing matrix that contains the semantic similarity between words obtained using Wordnet. This results in the increase in the feature value of the terms that are related semantically. The introduction of semantic prior knowledge in the SVM kernel or k-NN improves the classification performance [10]. Another work that used Wordnet for designing a semantic smoothing kernel for text classification is [11]. They calculated the similarity between words based on the shared superconcepts of these terms. Cristianini *et al.* [12] incorporated into a kernel the semantic relations between terms calculated using LSI. Wang *et al.* [13] enriched the bag-of-words representation by embedding the knowledge from Wikipedia into a semantic kernel to consider synonyms, polysemous words and concepts.

Supervised semantic smoothing kernels exist that utilize class information in building a semantic matrix [14; 15; 16]. A sprinkled diffusion kernel that uses both co-occurrence information and class information for word sense disambiguation is presented in [14]. In this approach, the smoothing helps in increasing the semantic relationship between terms in the same class. But, it does not distinguish the common terms between classes. Class Meaning Kernel (CMK) [15] is a supervised semantic kernel that considers the meaningfulness of terms in the classes using Helmholtz principle from Gestalt theory. In order to increase the importance of class specific terms compared to common terms, the semantic smoothing is done using the semantic matrix built from class-based meaning values of terms. Class Weighting Kernel (CWK) [16] smooths the representation of documents using class-based term weights that calculates the importance of the terms in the classes. Hence,

there are different variants of semantic kernels with variations in the design of the semantic smoothing matrix. Since a document is represented as a vector and is based on a term independence assumption, these semantic kernels ([10], [11], [12], [13], [14], [15], [16]) do not consider term dependencies such as the order of words or the distance between words in the computation of similarity between documents.

Medical text classification can aid in decision making and thus improve the quality of healthcare. Applications of text classification within the medical domain include classification of biomedical articles [6]; classification of clinical notes based on the medical sub-domain [17]; diagnosis of Autism Spectrum Disorder based on the information in patients' medical forms [18]; classification of hospital records based on diseases [19]; determination of Rheumatoid Arthritis Disease activity status from clinical notes of Rheumatoid Arthritis patients [20]; classification of ICU/NICU notes to identify the procedures and diagnosis [21]; determination of adverse drug reactions from social media text [22]; classification of clinical reports to identify cases of lung cancer [23] and classification of epilepsy diagnosis based on ICD-9 codes [24].

As text classification has several critical applications in the medical domain, it is necessary to classify the medical documents accurately by considering the relevant concepts and relationships in the documents. In this paper, we introduce a novel method for the automatic construction of concept graphs, which utilises both class information in the pre-classified training documents and knowledge from the UMLS semantic network to weight and enrich the graphs. The graph enrichment method presented in the paper helps in developing a similarity measure that goes beyond exact matching of terms and relationships. The advantage of our method over other approaches is that it is a simple and effective method to compute semantic similarity and does not require the computation of distance between all possible pairs of words for each document.

3. Automatic construction of enriched concept graphs

This section explains the process of converting a medical text document to an enriched graph representation. The steps in this process are illustrated in Figure 1. The initial step of identifying medical concepts within a text document using QuickUMLS is explained in section 3.1. The second step assigns a weight to each medical concept using a supervised concept weighting

scheme, and is presented in section 3.2. Finally, section 3.3 introduces the automatic construction of concept graph and its enrichment using similarity matrix **S**.

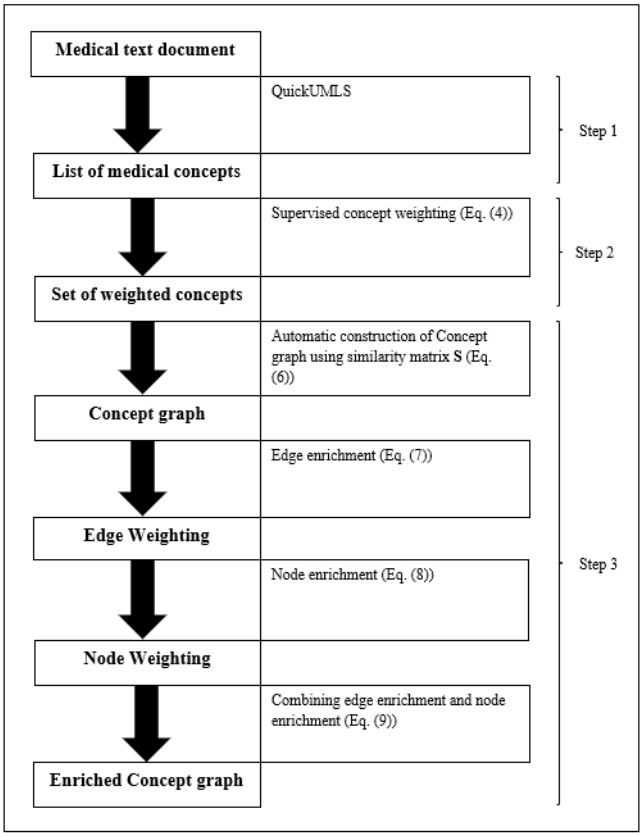

Figure 1: Steps to convert a medical text document to an enriched concept graph

3.1. Identification of medical concepts in text documents

The first step in our approach is to convert each medical document to a list of medical concepts, as shown in Figure 2 [25]. It is important to identify the medical entities in a medical document for accurate processing of the document. There are many tools available such as MetaMap [26], CTAKES [27] and QuickUMLS [28] that help in obtaining UMLS (https://uts.nlm.nih.gov /home.html) concepts from medical text documents. We have used QuickUMLS for mapping a medical text document to a set of medical concepts as

it is fast and effective. As QuickUMLS has high efficiency in the extraction of medical information, it can be applied to large medical datasets [28].

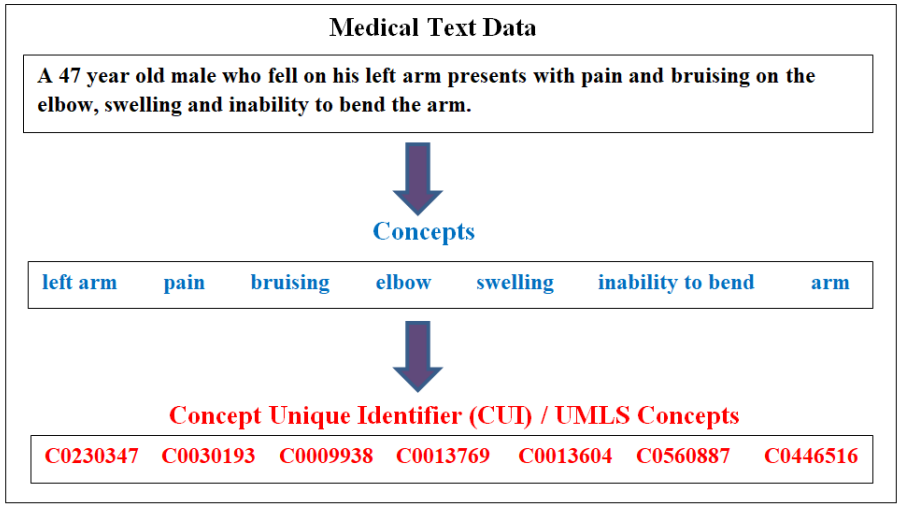

Figure 2: Medical Concept Identification in medical text data

QuickUMLS is a system based on approximate matching for the extraction of medical concepts in the UMLS meta-thesaurus. It identifies the text spans in documents approximately matching concepts in UMLS, and returns the concepts associated with each text span. The algorithm used in QuickUMLS for approximate dictionary matching is CPMerge, introduced in [29]. In QuickUMLS, the similarity functions such as Jaccard similarity (the default choice), cosine similarity, dice or overlap can be used for string matching. We can also set the threshold for the similarity value between strings (the default value is 0.7). The number of tokens to be considered for matching can be varied (the default is set to 5). The main advantages of QuickUMLS are its speed, applicability to large datasets and ability to capture variation of terms e.g. tumor and tumour [28]. The automatic construction of a graph for a document from the list of concepts obtained using QuickUMLS, and its enrichment, are explained in Section 3.3.

In the proposed concept graph representation of a document, a node corresponds to a medical concept in UMLS and an edge links the associated concepts within the document. We represent the rich information contained within each medical text document using a graph. The automatically constructed graph considers the following information:

(i) The medical concepts in the document.

(ii) The concepts related to these medical concepts by a parent-child relationship in the UMLS semantic network.

(iii) The frequency of each medical concept.

(iv) The relevance of each medical concept, as determined using a supervised concept weighting scheme.

(v) The similarity between each of the concepts in the document.

(vi) The association between each of the concepts in the document to determine the meaning of the document.

Each medical concept is assigned a concept unique identifier (CUI) which groups together terms that are synonymous. Hence, it considers terms that have the same meaning, resulting in a reduction of the number of terms required to represent a document. For example, the synonyms such as multiple myeloma, plasma cell myeloma and myelomatosis that are multi-word concepts are represented by a single node in the concept graph. The edges connect concepts in the document that are related and hence, the meaning of a document is determined by the connections between the nodes. Since the similarity between documents consider the relationship between concepts in each document, the similarity value is based on the meaning of the concepts in the documents, thereby solving the problem of polysemy.

Each document is initially represented as a concept vector $\mathbf{v}$, whose components correspond to the weights of the concepts. The proposed weighting is explained in Section 3.2. Since we utilise the distribution of the training documents in the classes to weight the concepts, the concept weighting approach is supervised.

3.2. Supervised Concept Weighting

The medical concepts identified for each document are weighted based on their relevance to distinguish the documents in different classes. Our approach to weighting concepts assigns higher weights to class specific concepts, thus reducing the weights of unimportant concepts within each document. To determine the relevance of the medical concepts, we utilize the supervised relevance weight (srw), an effective supervised term weight factor that we developed in [30] to determine the relevance of a term to the classification task. The calculation of srw for a concept m is given below.

We initially calculate the concentration of concept m in class C_i compared to its concentration in other classes. $class\_rel\_prob(m, C_i)$ in Eq. (1) denotes the concentration of concept m in class C_i where a, b and c denote the

number of documents in class C_i that contain the concept m, the number of documents in class C_i that do not contain the concept m and the number of documents not in class C_i that contain the concept m respectively.

$$class\_rel\_prob(m, C_i) = \log_2\Big(2 + \frac{a}{\max(1, c)}\Big) \times \log_2\Big(2 + \frac{a}{\max(1, b)}\Big) \quad (1)$$

To reduce the overweighting of unimportant concepts, the average density of the concept m in the classes is calculated as shown in Eq. (2) where C is the total number of classes and N_i is the total number of documents in class C_i.

$$avg\_density(m) = \frac{\sum_{i=1}^{C}\Big(\frac{a}{N_i}\Big)}{C} \quad (2)$$

The calculation of srw for a concept m is computed as shown in Eq. (3) where $max\_class\_rel(m)$ is the maximum of the $class\_rel\_prob(m, C_i)$ values for a concept m.

$$\text{srw}(m) = max\_class\_rel(m) \times \log_{10}\Big(\frac{1}{avg\_density(m)}\Big) \quad (3)$$

The weight of each concept m in the document is a product of the frequency of the concept denoted as $f(m)$ and srw of the concept as shown below.

$$w(m) = f(m) \times \text{srw}(m) \quad (4)$$

In section 3.3, we explain the conversion of the set of concepts in a document to a graph using an ontology-based similarity matrix.

3.3. Enriched Concept Graph Representation of Document

A similarity matrix $\mathbf{S} = (s_{ij})_{p\times p}$ is a square matrix of dimension $p \times p$ where p is the number of unique medical concepts in the training documents. The values in the similarity matrix correspond to the similarity between medical concepts determined using UMLS semantic network. If there is a parent-child relationship (is-a relation) between concept m_i and concept m_j in the UMLS semantic network, then the element s_{ij} in row i and column j of $\mathbf{S}$ should have a value greater than 0 and is set to 0.5. Since it is a symmetric matrix, the similarity between m_i and m_j is equal to the similarity between

m_j and m_i. We utilise this ontology-based similarity matrix to automatically construct a graph-based representation of text.

The initial representation of a document as a concept vector $\mathbf{v}$ of dimension p with components corresponding to the weight of the concepts obtained using Eq. (4) is converted to a concept graph using the similarity matrix $\mathbf{S}$. The matrix operations in Eq. (5)-(9) given below convert a concept vector to a concept graph, where nodes correspond to medical concepts and edges connect medical concepts with a parent-child relationship in the semantic network of UMLS. It also creates new nodes that are related to the initial concepts extracted from the document using QuickUMLS. To assign lower weights for edges connecting newly created nodes, $\mathbf{S_{mod}}$ is created for each document d from the similarity matrix $\mathbf{S}$. Hence, more importance is given to the associations between main concepts within the document by reducing the weights for the links with the newly added concepts using a reduction factor $x \in (0, 1]$ (the lesser the value of x, the higher is the weight reduction).

A vector $\mathbf{v_{mod}}$ (of dimension p) is created from $\mathbf{v}$ as given below in Eq. (5) where w_{mod} and w correspond to the elements of the vectors $\mathbf{v_{mod}}$ and $\mathbf{v}$ respectively and $x \in (0, 1]$ is the reduction factor.

$$w_{mod} = \begin{cases} 1 & \text{if } w > 0 \\ x & \text{if } w = 0 \end{cases} \tag{5}$$

$\mathbf{S_{mod}}$ (of dimension $p \times p$) is obtained by computing the element-wise multiplication of each row of $\mathbf{S}$ with $\mathbf{v_{mod}}$. The concept vector of a document d denoted as $\mathbf{v}$ is converted to a diagonal matrix $\mathbf{V}$ which is a matrix created with the elements of $\mathbf{v}$ on the diagonal. The matrix $\mathbf{V}$ is then multiplied with $\mathbf{S_{mod}}$ as in Eq. (6) to build a concept graph for document d.

$$\mathbf{A_1} = \mathbf{V} \times \mathbf{S_{mod}} \tag{6}$$

The average of the product $\mathbf{A_1}$ and its transpose $\mathbf{A_1^T}$ is computed as in Eq. (7) to assign weights to edges based on the average weight of the concepts that the edge connects, the similarity between the nodes (or concepts) and the reduction factor x. The weights assigned to edges correspond to the strength of association between the concepts in the document.

$$\mathbf{A_2} = \frac{(\mathbf{A_1} + \mathbf{A_1^T})}{2} \tag{7}$$

The next step is assigning weights to nodes using the similarity matrix $\mathbf{S}$ as in Eq. (8), resulting in node weights based on the weights of similar nodes.

$$\widehat{\mathbf{v}} = \mathbf{v} \times \mathbf{S} \tag{8}$$

$\widehat{\mathbf{V}}$ is a diagonal matrix obtained by setting $\widehat{\mathbf{v}}$ as the diagonal elements. $\widehat{\mathbf{A_2}}$ is obtained by setting the diagonal elements of $\mathbf{A_2}$ to 0. The adjacency matrix $\mathbf{E}$ of the enriched concept graph is obtained by adding $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{A_2}}$.

$$\mathbf{E} = \widehat{\mathbf{V}} + \widehat{\mathbf{A_2}} \tag{9}$$

Hence, the weight of each node is represented by a self-loop and corresponds to the importance of the medical concept.

4. Similarity calculation between enriched concept graphs

The proposed enriched concept graphs represent the rich information hidden in text. Since the information is represented by the edges of the proposed graph, we need a similarity measure that compares the structure of the graph representation of the documents. To utilise the information represented in the graphs to classify documents, we use an effective graph similarity measure. A graph kernel measures the similarity between graphs. We use an edge walk graph kernel [31; 7] that compares walks of length one in the graphs. The input to the kernel is a pair of concept graphs and the output is the similarity value based on the matching edges in the graphs.

If $G_i = (V_i, E_i)$ corresponds to the enriched concept graph of document d_i and $G_j = (V_j, E_j)$ corresponds to the enriched concept graph of document d_j, the similarity between the documents d_i and d_j is calculated using the edge walk kernel function [31; 7] as explained in Eq. (10)-(13). $k^{(1)}_{walk}$ is a kernel that compares edge walks of length 1 in the graphs G_i and G_j [31]. It is the product of kernel functions on the edge and the two nodes that the edge connects. It is defined in Eq. (10) for graphs G_i and G_j where u_i and v_i are the vertices that belong to the set of vertices V_i in G_i, e_i is the edge linking u_i and v_i in G_i, u_j and v_j are the vertices that belong to the set of vertices V_j in G_j and e_j is the edge connecting u_j and v_j in G_j. As defined in Eq. (11), k_{node} compares nodes in the graphs and returns one if the medical concepts corresponding to the nodes are equal, otherwise returns zero. We use $\mathrm{C}(u_i)$ and $\mathrm{C}(u_j)$ to denote the medical concepts that correspond to the nodes u_i and u_j respectively. k_{edge} compares edges in the graphs and is the

product of the weights of the edges in the graphs compared. If the weight of the edge e_i in G_i is $w_{edge}(e_i)$ and the weight of the edge e_j in G_j is $w_{edge}(e_j)$, then $k_{edge}(e_i, e_j)$ is the product of $w_{edge}(e_i)$ and $w_{edge}(e_j)$ as given in Eq. (12). Since the similarity value depends on the number of edges, a product of the Frobenius norms of the adjacency matrices A_i and A_j of the graphs G_i and G_j denoted as $||A_i||_F$ and $||A_j||_F$ respectively is considered as given in Eq. (13) so that the document similarity is not affected by the size of the graphs [7]. Hence, the numerator in Eq. (13) is equivalent to the sum of the elements in the element-wise product of the adjacency matrices A_i and A_j.

$$k^{(1)}_{walk}(e_i, e_j) = k_{node}(u_i, u_j) \times k_{edge}(e_i, e_j) \times k_{node}(v_i, v_j) \tag{10}$$

$$k_{node}(u_i, u_j) = \begin{cases} 1 & \text{if } \mathrm{C}(u_i) = \mathrm{C}(u_j) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$k_{edge}(e_i, e_j) = \begin{cases} w_{edge}(e_i) \times w_{edge}(e_j) & \text{if } e_i \in E_i \wedge e_j \in E_j \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$k(d_i, d_j) = \frac{\sum_{e_i \in E_i, e_j \in E_j} k^{(1)}_{walk}(e_i, e_j)}{||A_i||_F \times ||A_j||_F} \tag{13}$$

The similarity between every pair of documents is computed by calculating the similarity between enriched concept graph representations of documents using the edge walk kernel. These similarity values are used to build the kernel matrix $\mathbf{K}$ where a matrix element k_{ij} corresponds to similarity between i^{th} and j^{th} document. The matrix created satisfies the two important mathematical properties of kernel matrix i.e. symmetry and positive semi-definiteness. This matrix is then used to train SVM classifier and the classification model built is used for the classification of medical documents.

5. Graph Kernel Based Medical Document Classification

The proposed graph-kernel based medical text classification pipeline is shown in Figure 3. The documents are initially represented as a set of concepts obtained using QuickUMLS. The supervised term weight factor is utilised to assign weight to concepts. These concepts are converted to enriched graphs automatically using a similarity matrix built with ontology-based similarity values between concepts. A graph kernel based on edge

matching is then employed to calculate the similarity between a pair of documents. The similarity values are then used to build a kernel matrix. The kernel matrix is fed to a SVM to learn and predict the classes of the documents.

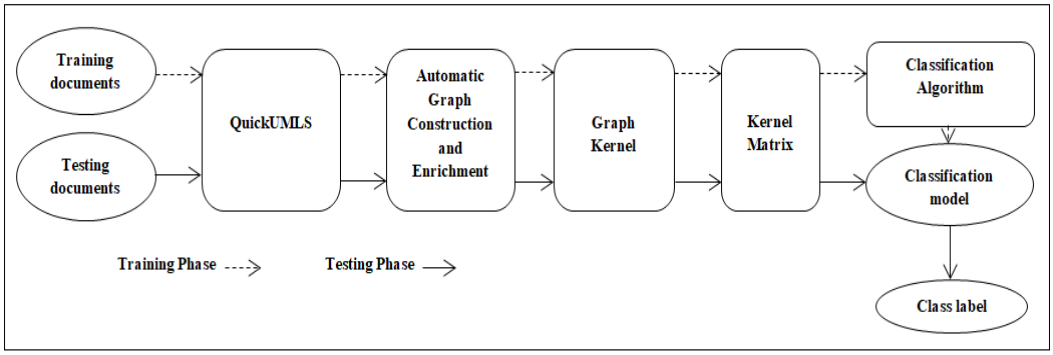

Figure 3: Medical text classification pipeline

6. Experiments and Results

In this section, we describe the experiments performed to evaluate the performance of our proposed approach on the classification of medical documents. The datasets used for medical text classification are listed below.

- Ohsumed Dataset[1] - This dataset contains medical abstracts classified into 23 cardiovascular disease categories. It is a multi-label dataset and the total number of abstracts is 13,929. We removed all the documents with more than one label to obtain a single label dataset, thereby reducing the size of the dataset to 7,400 documents.

- Medical Notes Dataset - This dataset contains 1,669 medical transcription reports obtained from https://www.mtsamples.com. It is classified into 11 specialities such as Cardiovascular/Pulmonology, Dermatology, ENT/Otolaryngology, Gastroenterology, Nephrology, Neurology, Obstetrics/Gynaecology, Ophthalmology, Orthopaedic, Psychiatry/Psychology and Urology.

[1] http://disi.unitn.it/moschitti/corpora.htm

Using QuickUMLS, each medical document is converted to a set of medical concepts. Table 1 shows the number of unique terms and concepts in each dataset. The size of a document is reduced considerably by mapping it to concepts. In QuickUMLS, we used the default similarity function i.e. Jaccard similarity measure, the default threshold of 0.7 for the minimum similarity between strings and the default window size of 5 for the limit on the number of tokens to be considered for matching. An enriched concept graph is automatically constructed for a document from the set of concepts using a similarity matrix as explained in Section 3. The similarity matrix containing the similarity values between medical concepts in the training documents, has the similarity of each concept to its parents/children in the hierarchical relationship in UMLS set to 0.5. The weight reduction factor x in Eq. (5) is set to 0.3. Therefore, the association with the related concepts added during enrichment is less than the association with the initial concepts. The edges are weighted based on the features of the concepts that they connect such as frequency, relevance, similarity between the concepts and the weight reduction factor x. The nodes that represent concepts group together synonyms and the association between concepts help to determine the context of the document. The graph similarity using edge walk kernels compare the concepts and the associations within each document. As the nodes and edges are weighted effectively, the matching nodes and edges contribute to document similarity based on their relevance.

Table 1: Number of unique terms and concepts

Dataset	**No. of unique terms**	**No. of unique concepts**
Ohsumed	31079	13039
Medical Notes	17985	9361

The enriched graph-based similarity measure is compared with linear kernel, cosine similarity, Tanimoto similarity, Sorensen similarity, radial basis function (RBF) kernel, shortest path graph kernel (with depth equal to 1) and the supervised semantic kernels i.e. CMK and CWK.

In shortest path graph kernel (spgk) with depth equal to 1, the graph is equivalent to an unweighted co-occurrence graph representation where nodes correspond to terms in the document and the edges connect terms that co-occur within a pre-defined sliding window of size 2. The Tanimoto similarity [32] and Sorensen similarity [32] are computed for boolean vectors of documents. Boolean vector is a bag-of-words representation with binary weights (either 1 or 0) to indicate the presence or absence of the term. The linear kernel, RBF kernel and cosine similarity are applied to tf-idf weighted vector representations of documents. tf-idf weighted vector is a bag-of-words representation with tf-idf weights. tf-idf is a product of tf (term frequency) and idf (inverse document frequency) which assigns more weight to rare terms than common terms.

The macro-averaged measure calculates the performance metric (such as precision, recall or F1 score) for each class and then computes the average of these metrics. Since this measure gives equal importance/weight to each class, the value of the measure is affected when there is a class imbalance [33]. When the performance metrics for small classes are high, the macro-averaged results could be high even when majority of the documents are not classified correctly. So we have used the weighted average that assigns weight to each class to evaluate the classification performance. The weights are based on the number of instances in the class. It calculates the metrics for each class and then computes the weighted average of these metrics. Hence, it can handle class imbalances unlike unweighted macro-averaged measure that gives equal importance for all the classes. Tables 2, 3 and 4 show the precision, recall and F1-scores (weighted average) obtained for the classification of the medical documents using the proposed similarity measure, linear kernel, cosine similarity, Tanimoto similarity, Sorensen similarity, radial basis function (RBF) kernel and shortest path graph kernel (with depth equal to 1). The results reported in these tables are obtained by 10-fold cross validation. The validation set is 20 percent of the training set and is used to optimize the value of the parameter C in SVM. The best value of C from the set of values $\{0.01,0.1,1,10,100,1000\}$ is then used to classify the documents in the testing set. Tables 5, 6 and 7 compare the performances (using train/test split) of the proposed method and the supervised semantic kernels i.e. CMK and CWK for the classi-

fication of the medical documents. Since CMK and CWK require long training time, the performance is evaluated by splitting the dataset into training and testing set in the 80:20 ratio. The default value of 1 for parameter C in SVM is then used to classify the documents. In text classification with CMK and CWK, attribute selection (as reported in their experiments [15] [16]) is applied using mutual information to select the best 2000 terms. CWK_{wfs} and CMK_{wfs} correspond to the semantic kernels CWK and CMK without performing this feature selection. There is a considerable improvement in the performance of these semantic kernels without feature selection. The precision, recall and F1-scores (weighted average) show the superior performance of the proposed approach compared to the baseline similarity measures for medical document classification.

Table 2: Precision values obtained for medical document classification using different similarity measures

Dataset	**Linear**	**Cosine**	**Sorensen**	**Tanimoto**	**RBF**	**Spgk**	**Proposed method**
Ohsumed	71.48	73.55	62.69	61.30	72.29	57.78	**74.94**
Medical Notes	83.70	84.23	81.40	79.06	83.76	75.96	**86.71**

Table 3: Recall values obtained for medical document classification using different similarity measures

Dataset	**Linear**	**Cosine**	**Sorensen**	**Tanimoto**	**RBF**	**Spgk**	**Proposed method**
Ohsumed	69.37	70.30	61.32	58.69	69.51	53.97	**73.51**
Medical Notes	83.48	84.14	81.63	79.72	83.54	76.81	**86.89**

Table 4: F1 scores obtained for medical document classification using different similarity measures

Dataset	**Linear**	**Cosine**	**Sorensen**	**Tanimoto**	**RBF**	**Spgk**	**Proposed method**
Ohsumed	69.33	70.50	60.49	57.13	69.63	51.50	**73.29**
Medical Notes	83.09	83.70	81.04	78.73	83.15	75.41	**86.41**

Table 5: Comparison of precision values obtained for medical document classification using supervised semantic kernels

Dataset	**CWK**	**CWK$_{wfs}$**	**CMK**	**CMK$_{wfs}$**	**Proposed Method**
Ohsumed	58.36	64.79	54.32	53.47	**73.19**
Medical Notes	85.14	84.45	79.91	80.26	**86.98**

Table 6: Comparison of recall values obtained for medical document classification using supervised semantic kernels

Dataset	**CWK**	**CWK$_{wfs}$**	**CMK**	**CMK$_{wfs}$**	**Proposed Method**
Ohsumed	58.92	64.12	54.80	53.85	**71.62**
Medical Notes	84.73	84.13	79.04	80.24	**86.53**

Table 7: Comparison of F1 scores obtained for medical document classification using supervised semantic kernels

Dataset	**CWK**	**CWK$_{wfs}$**	**CMK**	**CMK$_{wfs}$**	**Proposed Method**
Ohsumed	58.19	63.80	53.57	52.48	**71.48**
Medical Notes	84.56	83.99	79.22	79.86	**86.32**

The effectiveness of the text similarity measure depends on the method for document representation. The proposed automatic graph construction method converts a document to an enriched concept graph, which helps to consider the semantic relationship between terms in computing document similarity. It encodes information about the relevant medical concepts and their associations within the document. The information represented in the graphs is useful to take into account the relevant content in the documents for calculating the similarity between documents. Other commonly used similarity measures is based on term overlap. Medical documents contain complex medical terms and hence, it is important to understand the medical terminology to determine the content of the document. Our method helps to easily incorporate the medical terms and their relationships from the medical knowledge base for semantic classification of medical documents. Therefore, the proposed approach provides (i) a simple technique to convert a document to a concept graph (ii) an automatic method for graph enrichment that results in adding related nodes, linking the nodes and weighting the nodes and edges based on their relevance and (iii) a similarity measure that goes beyond exact matching of terms to consider relevant content in the documents. This method easily integrates knowledge into the text classification framework increasing the performance of classification of text documents. The proposed domain-specific text classification framework can be adapted for different domains by designing the similarity matrix based on the domain.

7. Conclusion

We developed an effective approach to represent and compare the main content of medical text documents, solving the challenges due to the complex terminology used in these documents. The set of medical concepts in each document are identified and then converted to an enriched concept graph using a similarity matrix. The enrichment adds related concepts, links the associated concepts and weights the concepts and their associations based on their relevance. The similarity between the enriched concept graphs is computed using a graph kernel and is used to classify medical text documents. The proposed method easily incorporates the knowledge from UMLS semantic network into the text

classification framework. The enriched graph-based similarity measure clearly outperforms the widely used similarity measures for medical document classification. The automatic graph enrichment method can be further explored by embedding information from different knowledge bases to design more effective similarity matrix and evaluate its effect on text classification.

References

[1] A. M. Cohen, W. R. Hersh, A survey of current work in biomedical text mining, Briefings in Bioinformatics 6 (1) (2005) 57–71 (March 2005).

[2] G. Mujtaba, L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh, H. F. Nweke, Clinical text classification research trends: Systematic literature review and open issues, Expert Systems with Applications (2018).

[3] A. Schenker, M. Last, H. Bunke, A. Kandel, Classification of web documents using a graph model., in: ICDAR, IEEE Computer Society, 2003, pp. 240–244 (2003).

[4] J. Wu, Z. Xuan, D. Pan, Enhancing text representation for classification tasks with semantic graph structures, International Journal of Innovative Computing, Information and Control (ICIC) 7 (5) (2011).

[5] T. Gonçalves, P. Quaresma, Using graph-kernels to represent semantic information in text classification., in: P. Perner (Ed.), MLDM, Vol. 5632 of Lecture Notes in Computer Science, Springer, 2009, pp. 632–646 (2009).

[6] S. Bleik, M. Mishra, J. Huan, M. Song, Text categorization of biomedical data sets using graph kernels and a controlled vocabulary., IEEE/ACM Trans. Comput. Biology Bioinform. 10 (5) (2013) 1211–1217 (2013).

[7] G. Nikolentzos, P. Meladianos, F. Rousseau, Y. Stavrakas, M. Vazirgiannis, Shortest-path graph kernels for document similar-

ity, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1890–1900 (2017).

[8] S. Srivastava, D. Hovy, E. Hovy, A walk-based semantically enriched tree kernel over distributed word representations, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1411–1416 (2013).

[9] J. Kim, F. Rousseau, M. Vazirgiannis, Convolutional sentence kernel from word embeddings for short text categorization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 775–780 (2015).

[10] G. Siolas, F. d'Alché Buc, Support vector machines based on a semantic kernel for text categorization, in: Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, Vol. 5, IEEE, 2000, pp. 205–209 (2000).

[11] S. Bloehdorn, R. Basili, M. Cammisa, A. Moschitti, Semantic kernels for text classification based on topological measures of feature similarity, in: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE, 2006, pp. 808–812 (2006).

[12] N. Cristianini, J. Shawe-Taylor, H. Lodhi, Latent semantic kernels, Journal of Intelligent Information Systems 18 (2-3) (2002) 127–152 (2002).

[13] P. Wang, C. Domeniconi, Building semantic kernels for text classification using wikipedia, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 713–721 (2008).

[14] T. Wang, W. Li, F. Liu, J. Hua, Sprinkled semantic diffusion kernel for word sense disambiguation, Engineering Applications of Artificial Intelligence 64 (2017) 43–51 (2017).

[15] B. Altınel, M. C. Ganiz, B. Diri, A corpus-based semantic kernel for text classification by using meaning values of terms, Engineering Applications of Artificial Intelligence 43 (2015) 54–66 (2015).

[16] B. Altınel, B. Diri, M. C. Ganiz, A novel semantic smoothing kernel for text classification with class-based weighting, Knowledge-Based Systems 89 (2015) 265–277 (2015).

[17] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, H. C. Chueh, Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach., BMC Med. Inf. Decision Making 17 (1) (2017) 155:1–155:13 (2017).

[18] J. Yuan, C. Holtz, T. H. Smith, J. Luo, Autism spectrum disorder detection from semi-structured and unstructured medical data., EURASIP J. Bioinformatics and Systems Biology 2017 (2017) 3 (2017).

[19] S. Kocbek, L. Cavedon, D. Martinez, C. Bain, C. Mac Manus, G. Haffari, I. Zukerman, K. Verspoor, Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources, Journal of biomedical informatics 64 (2016) 158–167 (2016).

[20] C. Lin, H. Canhao, T. Miller, D. Dligach, R. M. Plenge, E. W. Karlson, G. K. Savova, Feature engineering and selection for rheumatoid arthritis disease activity classification using electronic medical records, in: ICML Workshop on Machine Learning for Clinical Data Analysis, 2012 (2012).

[21] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, R. A. Dudley, J. Boscardin, N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit, Journal of the American Medical Informatics Association 21 (5) (2014) 871–875 (2014).

[22] A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, Journal of biomedical informatics 53 (2015) 196–207 (2015).

[23] S. Kocbek, L. Cavedon, D. Martinez, C. Bain, C. M. Manus, G. Haffari, I. Zukerman, K. Verspoor, Evaluating classification

power of linked admission data sources with text mining, in: Annual Conference in Big Data in Health analytics (David Hansen 20 October 2015 to 21 October 2015), Vol. 1468, 2015, p. 17 (2015).

[24] L. Pereira, R. Rijo, C. Silva, M. Agostinho, Icd9-based text mining approach to children epilepsy classification, Procedia Technology 9 (2013) 1351–1360 (2013).

[25] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, [PowerPoint presentation], Available at: http://medir2016.imag.fr/data/slides_paper16.pdf.

[26] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17 (2001).

[27] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, Journal of the American Medical Informatics Association 17 (5) (2010) 507–513 (2010).

[28] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, in: MedIR workshop, sigir, 2016 (2016).

[29] N. Okazaki, J. Tsujii, Simple and efficient algorithm for approximate dictionary matching, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 851–859 (2010).

[30] N. Shanavas, H. Wang, Z. Lin, G. I. Hawe, Supervised graph-based term weighting scheme for effective text classification., in: ECAI, Vol. 2016, 2016, pp. 1710–1711 (2016).

[31] K. M. Borgwardt, H.-P. Kriegel, Shortest-path kernels on graphs, in: Fifth IEEE international conference on data mining (ICDM'05), IEEE, 2005, pp. 8–pp (2005).

[32] L. Ralaivola, S. J. Swamidass, H. Saigo, P. Baldi, Graph kernels for chemical informatics, Neural networks 18 (8) (2005) 1093–1110 (2005).

[33] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data mining and knowledge discovery handbook, Springer, 2009, pp. 667–685 (2009).