Semantic Term Weighting for Clinical Texts

Ryosuke Matsuo[a,*], Tu Bao Ho[a,b]

[a] *Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan*
[b] *John von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam*

Abstract

This paper introduces a new semantic term weighting method for clinical texts that are the core of electronic medical records (EMRs). The proposed method exploits a ranking of death causes and well-known resources UMLS and ICD-10 to identify the medical importance of terms regarding the severity of patients' conditions appearing in a clinical text. This semantic term weighting method is shown to be appropriate in classification such as mortality prediction using clinical texts.

Keywords: semantic term weighting, ontology, causes of death ranking, vector space model, mortality prediction, severity of patients' conditions

1. Introduction

In recent years, the prevalence of clinical texts such as electronic medical records (EMRs) and electronic health records (EHRs) opens new chances for developing methods to solve many significant problems in medicine [1, 2, 3]. The prevalence has led to the need for processing the clinical texts. The clinical texts are narratives about patients with their diagnosis and treatment data at hospitals. This is naturally different from other common medical literature such as medical digitalized books and research articles that are usually about diseases and treatments. Representing the clinical texts in a computable form is required for further tasks of text processing.

*Corresponding author
Email addresses: matsuor@jaist.ac.jp (Ryosuke Matsuo), bao@jaist.ac.jp (Tu Bao Ho)

The vector space model (VSM) is powerful for various language processing tasks in which term weighting - giving a numerical weight to each term appearing in a document representing its importance regarding the document - plays a crucial role. Term weighting has been used in various language processing methods, including text classification, clustering, sentiment analysis, recommendation and information retrieval. The traditional measure used in term weighting is TFIDF [4], derived from the term frequency and the inverse document frequency. TFIDF and its variants are commonly used as weights of document terms in VSM. TFIDF is simple and effective, and it forms a popular base for advanced algorithms in spite of its age [5].

However, the term importance captured by term weighting methods using TFIDF and its variants does not relate to the term meanings but only to the frequencies. These methods are not suitable for applications that require considering the term meanings. Therefore, semantic term weighting methods have been developed aiming at assigning weights to document terms based on their meanings.

Ontology-based term weighting has been pursued by many researchers. Ontologies systematically represent domain knowledge as a hierarchical structure by concepts and relationships that can exist between terms [6]. Tar and Nyunt [7] and Sureka and Punitha [8] exploited ontologies for concept weighting by focusing on the length of words from the association between two concepts, the correlation coefficient of the word depending on the concept's existence in the ontology and concept probability derived from the number of occurrences of the concept and all concepts. Zakos and Verma [9] and Sakre et al. [10] exploited four types of conceptual information in WordNet to determine term importance. The four types: number of senses, number of synonyms, level number and number of children, were employed to determine the term's generality or specificity, unlike IDF depending on the document collection's statistics. Some work demonstrated the semantic relationship of terms based on their conceptual similarity [11, 12, 13, 14]. After the calculation through TFIDF, the term weight was adjusted in accordance with the semantic similarity of other terms in the

same vector. Luo et al. [15] augmented term weights based on the relevance of terms to categories in the WordNet ontology.

Medical ontologies such as UMLS or MeSH have been exploited in semantic term weighting for medical literature. Zhang et al. conducted semantic term weighting by considering the semantic relationship of terms using the MeSH ontology [13, 14]. The medical ontology UMLS was employed to expand queries by utilizing categories such as the UMLS concept and the UMLS synonym. The method exploiting UMLS augmented the query terms from the IDF weights based on the categories [16]. Zhu et al. utilized UMLS to augment term weights based on the selected major UMLS semantic types for TREC 2004 Genomics Ad Hoc Retrieval Task [17].

TFIDF and its variants were applied to the clinical texts such as EMRs [18, 19]. Semantic features such as named entities and semantic predications were additionally considered exploiting the clinical texts [20]. The medical ontology UMLS was employed to identify concepts for the semantic features in EMRs.

As mentioned above, clinical texts mainly contain narratives about the patient diagnosis and treatment, and thus the exploitation of such texts also has its own characteristics to support evidence-based diagnosis, patient information retrieval and medication safety of patients, among others. The importance of words in each EMR clinical text therefore not only relates to the medical meanings but also to the patients' status. In order to consider the practical usage of term weighing regarding patients, we basically deal with the phenomenon that the term's medical importance is not fixed, but it depends on the context or aspect under consideration. Regarding different aspects under consideration, the importance of a term can increase or decrease.

This work aims to develop a new semantic term weighting method for EMR clinical texts where importance is considered regarding the severity of patients' conditions. The key idea of our method is to employ both medical ontology UMLS as well as ICD-10 codes and a ranking of causes of death to additionally adjust the weights of terms that are initialized by TFIDF. The use of the knowledge allows us to classify words in EMRs into several categories with dif-

ferent importance in terms of the severity of patients' conditions. A preliminary version of the proposed method showed high performance in an experiment on mortality prediction described in a recent paper [21]. In this work, we further elucidate adequate parameters of Δ and α. The parameter Δ is regarded as the different degrees of the weight importance of each rank in the death causes ranking. The parameter α is a coefficient of the TFIDF weight and the medical importance weight when the two weights are combined for the final term weight. The corresponding coefficients are α_1 and α_2, respectively. Moreover, we execute a statistical test to confirm the effectiveness of the proposed method using the adequate parameters.

This paper uses the well-known database MIMIC II [22] of EMRs to evaluate the proposed method on mortality prediction. The experimental results showed that the proposed method outperformed the TFIDF-based method. Moreover, the statistical test results showed a significant difference between the proposed method and the TFIDF-based method. It was found that the proposed method was not as dependent on parameter α as on parameter Δ. Specifically, the smaller the parameter of Δ is, the greater the performance of the mortality prediction.

Although, methods for mortality prediction have been developed by using scores such as sequential organ failure assessment (SOFA) and simplified acute physiology score (SAPS) or some algorithms without the scores [23, 24, 25, 26], term weighting methods representing clinical data were not exploited for mortality prediction. Therefore the proposed semantic term weighting method is a novel approach to represent clinical data into the VSM form for mortality prediction.

This work contributes a solution to semantic term weighting for representing clinical texts. Our proposed method gives a weight to a term regarding patients, especially the severity of patients' conditions, by exploiting the UMLS ontology, ICD-10 and the ranking of causes of death. Since the proposed method is based on the the severity of patients' conditions concerning various types of diseases, it can be applied to a comprehensive prediction of patients' risk on the clinical

texts such as EMRs. Moreover, the proposed method can be applied to various computational problems in medicine because it simply represents the clinical texts into the VSM form. This advantage also can be utilized for sharing and integrating clinical data in different systems.

2. Proposed Method

2.1. The framework

The proposed framework determines a semantic weight for each term appearing in a given EMR's clinical text. In fact, each EMR's term is assigned a weight w that is a combination of an initial TFIDF weight w_0 and a weight w_m based on its medical importance. In this work, we consider the medical importance of clinical terms regarding the severity of patients' conditions. It is worth noting that it is infeasible to give distinct weights to all distinct terms that can occur in clinical texts as the number of such terms is very large.

The key idea is to hierarchically divide the terms into categories; the terms in each category are reasonably considered to have the same medical importance (weight) regarding the severity of patients' conditions. Especially, we pay more attention to the terms that strongly relate to disease severity. To this end, we choose widely accepted medical knowledge on a ranking of causes of death [27], and identify 18 categories of terms each contain terms with the same importance about disease severity.

The proposed framework for identifying the medical importance of EMR's terms regarding disease severity consists of two stages:

1. *To classify EMR's terms into 18 categories characterizing different levels of the medical importance of terms.*

2. *To appropriately determine the medical importance of these 18 categories (i.e., weights of all the terms in the category).*

Note that the framework can be applied when considering other medical important aspects.

2.2. Classification of clinical terms into medical importance categories

The proposed method to classify EMR's terms into 18 medical importance categories with increasing medical importance is described in Figure 1 and illustrated in Figure 2. Each term is classified into one of 18 categories by the four steps in the method.

The first step is to determine whether the given term in an EMR is a medical term. To this end, we employ the Unified Medical Language System (UMLS). UMLS is composed of the Metathesaurus which is a repository of biomedical concepts, the semantic network which provides 135 categories of the concepts and lexical resources [28]. We identity the Concept Unique Identifiers (CUI) that contain the letter C followed by seven numbers from the UMLS Metathesaurus to identify if the term appearing in the EMR is a medical term. We use the well-known tool MetaMap that maps biomedical text to the UMLS Metathesaurus [29]. During the mapping process, we focus on the terms regarding disease severity such as disorders or drugs for developing patient severity-based term weighting. Finally, we regard the term as a medical term if the term has a CUI code, otherwise, the term is a non-medical term. If the term is recognized as a non-medical term, it is put into category C_1. If not, we go to step 2. The terms in C_1 don't have the medical importance from the viewpoint of the severity of patients' conditions.

The second step is to determine whether a medical term is one term in the well-known classification ICD-10. The International Statistical Classification of Diseases and Related Health Problems (ICD) is an international standard diagnostic classification for all general epidemiological and many health management purposes [30]. The classification provides the alphanumeric codes of medical terms for such diagnoses where the codes are structured in a hierarchy. This work utilizes CUI codes to obtain the ICD-10 codes from the identified medical terms in the previous step. We identify whether the medical term has an ICD-10 code by using the interoperable code of the UMLS concept on BioPortal that is an open repository of biomedical ontologies and can map the CUI code to the ICD-10 classification [31]. If the term is not an ICD-10 term, it is

put into category C_2. If not, we go to step 3. The terms in C_2 have medical importance in terms of the severity of patients' conditions, but correspond to a low weight.

The third step is to determine whether the ICD-10 term is in the list of ranked terms regarding the severity of patients' conditions. To this end, this paper employs a ranking of causes of death that provides the 15 leading causes of death with each of the corresponding ICD-10 codes [27]. The combination of the ranking of causes of death and the ICD-10 hierarchical structure is accomplished by connecting the ICD-10 code of the ICD-10 term in the hierarchical structure with the ICD-10 code of each rank in the ranking. Thus, the ranking gives the medical importance weights of each rank to the corresponding ICD-10 terms in the hierarchical structure of ICD-10. If the term is not a ranked term, it is put into category C_3. If not, we go to step 4. The terms in C_3 have more medical importance than the terms in C_2 regarding disease severity, as the ICD's original use was to classify causes of mortality and it has been applied for dealing with mortality data [30].

The fourth step is to classify a ranked term into one of 15 categories in the ranking of causes of death according to the category it matches. The terms in the categories $C_4, C_5, ..., C_{18}$ have increasingly higher medical importance regarding the severity of patients' conditions in comparison with the categories C_1, C_2 and C_3. The higher the rank of a category the higher the weights of terms in the corresponding category. Therefore, category C_{18} gives the highest medical importance to the corresponding terms among 18 categories.

Note that we evidently have the following relations among categories of ranked terms those ensure that every EMR's term will be classified into one of the 18 categories.

$$\{\text{ICD-10 ranked terms}\} \cup \{\text{ICD-10 non-ranked terms}\} = \{\text{ICD-10 terms}\}$$

$$\{\text{ICD-10 terms}\} \cup \{\text{non-ICD-10 terms}\} = \{\text{medical terms}\}$$

$$\{\text{medical terms}\} \cup \{\text{non-medical terms}\} = \{\text{EMR clinical terms}\}$$

Figure 3 indicates some examples of sentences in EMRs. Figure 4 describes the processing of those sentences to classify them into medical importance categories. For instance, terms 'female' and 'episode' belong to the category C_1 as non-medical terms according to MetaMap. In contrast, terms 'paroxysmal nocturnal dyspnea' and 'hypercholesterolemia' belong to the category C_2 because these do not have ICD-10 codes even they are medical terms. The term 'shortness of breath' is an ICD-10 term which is not ranked in the ranking of causes of death. Accordingly, this term is classified into the category C_3. The ICD-10 ranked terms in the death causes ranking will belong to the categories between C_4 and C_{18}. The term 'congestive heart failure' where the ICD-10 code is I50 corresponds to the top rank. Hence, this term belongs to the category C_{18}. Terms 'diabetes mellitus' and 'hypertension' are ICD-10 ranked terms where the ICD-10 codes are E10-E14.9 and I10-I15.9, respectively. As each of these terms is positioned as the rank 7 and rank 13, these belong to the category C_{12} and C_6, respectively.

2.3. Determination of the medical importance of each category

The problem in this stage is to appropriately determine the degrees of medical importance for 18 categories of EMR's terms. Denote by $w(C_i)$ the weight (a real number) regarding the medical importance of the category C_i from 18 term categories to be determined. The determination of $w(C_i)$ should obey the following constraint:

Proposition. *The value of $w(C_i)$ can be arbitrarily determined but should preserve the linear relation*

$$w(C_1) < w(C_2) < ... < w(C_{18})$$

Noting from the above Proposition, we determine the $w(C_i)$ by equal width intervals

$$w(C_i) = w(C_{i+1}) - \Delta \tag{1}$$

where Δ is a parameter to change the different degrees of medical importance among 18 categories. In this work, we consider four degrees: 0.04, 0.03, 0.02 and 0.01 for the parameter Δ. Since category C_1 does not contain medical terms, the weight $w(C_1)$ regarding the medical importance is zero. Table 1 indicates the weights of each category with the corresponding name of cause of death as well as the ICD-10 code(s). Different category weights can be described by varying the parameter Δ.

2.3.1. Transferring the degree of the ranked term's medical importance to the subordinate term.

In order to utilize the ICD-10 hierarchy, we exploit the hierarchical structure of ICD-10 to transfer each rank's medical importance from the corresponding ICD-10 terms to the subordinate ICD-10 terms. That is, a subordinate term of a ranked term is given the same weight as the ranked term. For example, the term 'malignant neoplasm' is positioned as the second rank in the ranking of causes of death where the rank covers ICD-10 codes among C00-C97. The second rank's medical importance weight is then transferred to the subordinate ICD-10 terms in the range of C00-C97.

2.3.2. Example of a relation among ICD-10 terms associated with the medical importance.

Given terms 'coronary artery disease', 'hypertension' and 'peripheral vascular disease', which correspond to the ICD-10 codes I25.1, I10-I15.9 and I73.9, respectively. Although these ICD-10 codes belong to the same class as diseases of the circulatory system (I00-I99) in ICD-10, the proposed method distinguishes these ICD-10 codes in terms of the severity of patients' conditions as follow

$$\text{I73.9 } (w(C_3)) < \text{I10-I15.9 } (w(C_6)) < \text{I25.1 } (w(C_{18}))$$

Thus, the proposed method identifies the medical importance of ICD-10 terms regarding the severity of patients' conditions by exploiting ICD-10 codes between the ranking of causes of death and the ICD-10 hierarchy.

2.3.3. Combining medical importance weights with TFIDF weights.

The medical importance weight w_m is finally combined with the TFIDF weight w_0 for the final weight w under consideration by following Equation

$$w = \alpha_1 \times w_0 + \alpha_2 \times w_m \tag{2}$$

where α_1 and α_2 are the coefficients of the two weights, respectively. These are employed to balance the TFIDF weight and the medical importance weight. In this work, we investigate 7 cases of the α_1 and α_2 combination as shown in Table 2.

Note that the combination of the two weights is executed for terms appearing in clinical texts after preprocessing of the clinical texts such as stop word removal, chunking and removing the terms correspond to negation words.

3. Experimental evaluation

3.1. Objectives

This paper compares the proposed semantic term weighting method with the TFIDF-based method as a baseline to verify the effectiveness of the proposed method. Moreover, this paper elucidates adequate parameters of Δ and α for the proposed weighting.

3.2. Experimental settings

This paper conducts an experiment on mortality prediction for the evaluation as the proposed weighting is based on the severity of patients' conditions or death causes. EMRs of elderly patients are used from the well-known database MIMIC II [22]. The patients belong to two groups, one is people who died in hospital and the other is people who remained in hospital.

Regarding the statistical aspect, this work uses a total of 13,026 EMRs that contain information about patients who are more than 60 years old. The numbers of EMRs corresponding to the two groups' labels are 2,158 and 10,868, respectively. Table 3 describes the distribution of the document frequencies

of terms and the number of terms where each term belongs to one of the 18 categories.

This paper evaluates the proposed method in its four options corresponding to four different values of the parameter Δ, namely, TFIDF + MED ($\Delta = 0.04$), TFIDF + MED ($\Delta = 0.03$), TFIDF + MED ($\Delta = 0.02$) and TFIDF + MED ($\Delta = 0.01$). These options are compared to the baseline (TFIDF) when varying the parameter α.

In this experiment, we use five classifiers: SVM(rbf), SVM(linear), Naive Bayes, Random forest and Decision tree. Each classifier is executed after the feature selection by using L2 regularization where the parameters of each classification method and L2 regularization are default. A 5-fold stratified cross validation is performed to compute F1 scores on a dataset of 500 EMRs. F1 scores are computed by Equation 3.

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

This paper repeats 100 times the 5-fold stratified cross validation and takes the average value of F1 for the comparison. The dataset is randomly selected with the same ratio of the two labels. The ratio of the training and testing dataset is 80 % and 20 % in each cross validation.

3.3. Results and discussion

The result of the baseline (TFIDF) is compared to the result of the proposed method's options: TFIDF + MED ($\Delta = 0.04$), TFIDF + MED ($\Delta = 0.03$), TFIDF + MED ($\Delta = 0.02$) and TFIDF + MED ($\Delta = 0.01$) when varying the parameter α. The results are in Tables 4 - 8.

The experimental results showed that the proposed method when varying the parameters was generally better than the TFIDF-based method. More specifically, the proposed method's results using the classifiers except Naive Bayes outperformed the results of the TFIDF-based method. This suggests that the proposed semantic term weighting method based on the severity of patients' conditions is appropriate for the mortality prediction task.

On the whole, the smaller Δ is, the greater the prediction performance. The smallest Δ brought about the best F1 score among the proposed method's options in varying the parameter Δ where the classifiers were SVM (rbf), SVM (linear) and Naive Bayes. Moreover, there was a great difference of the scores between TFIDF + MED ($\Delta = 0.04$) and TFIDF + MED ($\Delta = 0.01$). For example, the score's difference was approximately 14 % where SVM (rbf) was employed in the best case of α. On the other hand, the difference of the scores was not remarkable when the classifiers were Random forest and Decision tree. As the small Δ accentuates the medical importance weights, we found that emphasizing the influence of the semantic weighting by the parameter Δ worked well for improving the prediction performance.

The different values of α_1 and α_2 did not significantly effect the performance. However, the combination of α_1 (= 1) and α_2 (= 0.25 or 0.5) was generally better than other cases where the small Δ was employed. This suggests that the proposed method does not depend on parameter α as much as it depends on parameter Δ.

3.4. Statistical test

In order to test whether the proposed method is significantly better than the TFIDF-based method, a t-test was executed under the condition that the scores of the all weighting methods were derived from the same experimental settings.

Figure 5 indicates the p-values by the t-test between the TFIDF-based method and the proposed methods using the best case of α in each of the four options of Δ. Four classifiers were described in Figure 5 as the proposed method's options using these classifiers in varying the parameters α and Δ overwhelmed the TFIDF-based method. The results showed that the p-values were less than 0.05 where the proposed method was better than the TFIDF-based method. Hence, there was a significant difference between the proposed method and the TFIDF-based method in terms of their performance.

On the other hand, the p-value was higher than 0.9 where Δ (= 0.03) and

SVM (rbf) were employed in the best α because there was little difference among the F1 scores under the condition. Although the proposed method was better than the TFIDF-based method, the p-value was approximately 0.1 where Δ (= 0.04) and SVM (linear) were employed in the best α. This suggests that in cases where there was little improvement using the proposed method, its performance was still comparable to that of the TFIDF-based method.

4. Conclusion

This paper proposed a novel semantic term weighting method for EMR clinical texts based on the severity of patients' conditions. In order to capture term importance regarding the severity of patients' conditions, this paper combined mapping information from the medical ontology UMLS, a hierarchal relation of medical terms in the ICD-10 classification and a linear relation from a ranking of causes of death.

The experiments of mortality prediction were performed to verify the effectiveness of the proposed semantic term weighting method and to elucidate adequate parameters of Δ and α. The experimental results showed that the proposed method in varying the parameters generally outperformed the TFIDF-based method. Its effectiveness was verified by T-test because there was a significant difference between the proposed method and the TFIDF-based method in terms of their performance.

As the proposed method was developed based on the medical ontology UMLS, the medical classification ICD-10 and the ranking of causes of death, the proposed method's results were based solely on medical knowledge.

It was found the smaller the parameter of Δ is, the greater the performance of the mortality prediction in most conditions. Since small Δ accentuates the effect of the medical importance, the significance of the proposed semantic term weighting was revealed by the experiments. Although, the different values of α_1 and α_2 did not significantly effect prediction performance, the combination of α_1 (= 1) and α_2 (= 0.25 or 0.5) was generally better than other cases where

a small Δ value was employed. It was also found that the proposed method's performance was less dependent on parameter α than on parameter Δ.

Since the proposed method does not need to specify the severity of diseases or symptoms, the proposed term weighting uniformly captures the medical importance in terms of the severity of patients' conditions. The proposed method can be applied to a comprehensive prediction of patients' risk or severity-based similar case retrieval on clinical documents such as EMRs. Because of the prevalence of EMRs, simply representing the clinical texts into the VSM form by the proposed term weighting method contributes to the medical field from the perspective of various applications in exploiting clinical texts such as EMRs. The proposed method also contributes to share and integrate clinical data in different systems by using the VSM form.

References

[1] D. Blumenthal, M. Tavenner, The meaningful use regulation for electronic health records, New England Journal of Medicine 363 (6) (2010) 501–504.

[2] C. C. Yang, P. Veltri, Intelligent healthcare informatics in big data era, Artificial intelligence in medicine 65 (2) (2015) 75–77.

[3] R. L. Richesson, J. Sun, J. Pathak, A. N. Kho, J. C. Denny, A survey of clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods, Artificial Intelligence in Medicine 71 (2016) 57–61.

[4] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (5) (1988) 513–523.

[5] J. Ramos, Using tf-idf to determine word relevance in document queries, Technical report, Department of Computer Science, Rutgers University (2003).

[6] T. R. Gruber, A translation approach to portable ontology specifications, Knowledge acquisition 5 (2) (1993) 199–220.

[7] H. H. Tar, T. T. S. Nyunt, Ontology-based concept weighting for text documents, World Academy of Science, Engineering and Technology 81 (2011) 249–253.

[8] V. Sureka, S. Punitha, Approaches to ontology based algorithms for clustering text documents, International Journal of Computer Technology and Applications 3 (5) (2012) 1813–1817.

[9] J. Zakos, B. Verma, Concept-based term weighting for web information retrieval, International Journal of Computational Intelligence and Applications 6 (02) (2006) 193–207.

[10] M. M. Sakre, M. M. Kouta, A. M. Allam, Weighting query terms using wordnet ontology, International Journal of Computer Science and Network Security 9 (2009) 349–358.

[11] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, E. E. Milios, Semantic similarity methods in wordnet and their application to information retrieval on the web, in: Proceedings of the 7th annual ACM international workshop on Web information and data management, ACM, 2005, pp. 10–16.

[12] L. Jing, L. Zhou, M. K. Ng, J. Z. Huang, Ontology-based distance measure for text clustering, in: Proceedings of SIAM SDM workshop on text mining, Bethesda, Maryland, USA, 2006.

[13] X. Zhang, L. Jing, X. Hu, M. Ng, X. Zhou, A comparative study of ontology based term similarity measures on pubmed document clustering, in: International Conference on Database Systems for Advanced Applications, Springer, 2007, pp. 115–126.

[14] X. Zhang, L. Jing, X. Hu, M. Ng, J. X. Jiangxi, X. Zhou, Medical document clustering using ontology-based term similarity measures, International Journal of Data Warehousing and Mining (IJDWM) 4 (1) (2008) 62–73.

[15] Q. Luo, E. Chen, H. Xiong, A semantic term weighting scheme for text categorization, Expert Systems with Applications 38 (10) (2011) 12708–12716.

[16] H. Yu, Y.-G. Cao, Using the weighted keyword models to improve information retrieval for answering biomedical questions, AMIA summit on translational bioinformatics.

[17] W. Zhu, X. Xu, X. Hu, I.-Y. Song, R. B. Allen, Using umls-based reweighting terms as a query expansion strategy, in: IEEE International Conference on Granular Computing, 2006, pp. 217–222.

[18] M. Hoogendoorn, P. Szolovits, L. M. Moons, M. E. Numans, Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer, Artificial intelligence in medicine 69 (2016) 53–61.

[19] G. Napolitano, A. Marshall, P. Hamilton, A. T. Gavin, Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction, Artificial intelligence in medicine 70 (2016) 77–83.

[20] R. Kavuluru, A. Rios, Y. Lu, An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records, Artificial intelligence in medicine 65 (2) (2015) 155–166.

[21] R. Matsuo, T. B. Ho, Mortality prediction using a semantic term weighting method based on severity of patients (in japanese), in: Proceedings of The 30th Annual Conference of the Japanese Society for Artificial Intelligence, 4D1-4in2, 2016.

[22] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, R. G. Mark, Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database, Critical care medicine 39 (5) (2011) 952–960.

[23] G. Richards, V. J. Rayward-Smith, P. Sönksen, S. Carey, C. Weng, Data mining for indicators of early mortality in a database of clinical records, Artificial intelligence in medicine 22 (3) (2001) 215–231.

[24] F. Jiménez, G. Sánchez, J. M. Juárez, Multi-objective evolutionary algorithms for fuzzy classification in survival prediction, Artificial intelligence in medicine 60 (3) (2014) 197–219.

[25] V. J. R. Ripoll, A. Vellido, E. Romero, J. C. Ruiz-Rodríguez, Sepsis mortality prediction with the quotient basis kernel, Artificial intelligence in medicine 61 (1) (2014) 45–52.

[26] R. Houthooft, J. Ruyssinck, J. van der Herten, S. Stijven, I. Couckuyt, B. Gadeyne, F. Ongenae, K. Colpaert, J. Decruyenaere, T. Dhaene, et al., Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores, Artificial intelligence in medicine 63 (3) (2015) 191–207.

[27] S. L. Murphy, J. Xu, K. D. Kochanek, Deaths: final data for 2010, National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System 61 (4) (2013) 1–117.

[28] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (suppl 1) (2004) D267–D270.

[29] 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, pp. 17–21.

[30] World Health Organization, International statistical classification of diseases and related health problems, Vol. 1, World Health Organization, 2004.

[31] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, et al., Bioportal: ontologies

and integrated data resources at the click of a mouse, Nucleic acids research
37 (2009) W170–W173.

1. **Step 1**: If the term is not a medical term identified by UMLS, put it to category C_1. If it is a medical term, go to the next step.
2. **Step 2**: If this medical term is not in ICD-10, put it to category C_2. If it is an ICD-10 term, go to the next.
3. **Step 3**: If this ICD-10 term is not ranked by a ranking of causes of death, put it to category C_3. If it is a ranked term, go to the next.
4. **Step 4**: If this ICD-10 term is ranked by the death causes ranking, put it into one of categories C_4, C_5, ..., C_{18} by according the rank of the death cause.

Figure 1: The description to classify each EMR's term into one of medical importance categories.

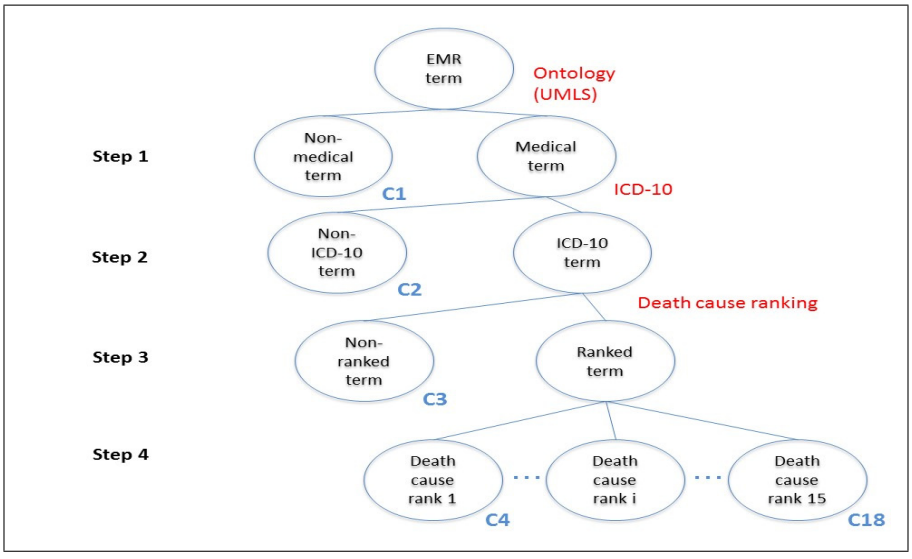


Figure 2: The illustration to classify each EMR's term into one of medical important categories

1. *Mrs. [**Known patient lastname 4483**] is an 81 year old female with congestive heart failure. She has been medically managed but has gradually experienced worsening symptoms of dyspnea on exertion and paroxysmal nocturnal dyspnea.*

2. *She did have that one episode of shortness of breath which was most likely due to acute pulmonary edema.*

3. *As the patient has risk factors of diabetes mellitus, hypertension, and hypercholesterolemia and possible old inferior myocardial infarction on electrocardiogram it was felt that ischemia was the likely cause of her conduction system abnormalities.*

Figure 3: Example of sentences in EMRs

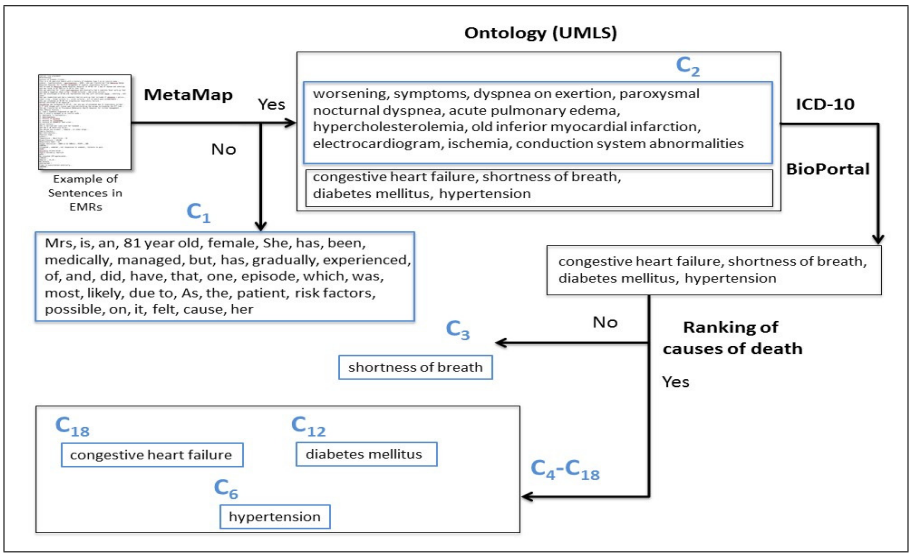


Figure 4: Example of the classification of terms appearing in EMRs

Table 1: The ranking-based medical importance weights in terms of the severity of patients' conditions

Rank	Name of cause of death	ICD-10 code(s)	Weight ($\Delta = 0.04$)	Weight ($\Delta = 0.03$)	Weight ($\Delta = 0.02$)	Weight ($\Delta = 0.01$)	Category
1	Disease of heart	I00-I09, I11, I13, I20-I51	0.7	0.7	0.7	0.7	C_{18}
2	Malignant neoplasms	C00-C97	0.66	0.67	0.68	0.69	C_{17}
3	Chronic lower respiratory diseases	J40-J47	0.62	0.64	0.66	0.68	C_{16}
4	Cerebrovascular diseases	I60-I69	0.58	0.61	0.64	0.67	C_{15}
5	Accidents (unintentional injuries)	V01-X59, Y85-Y86	0.54	0.58	0.62	0.66	C_{14}
6	Alzheimer's disease	G30	0.49	0.55	0.6	0.65	C_{13}
7	Diabetes mellitus	E10-E14	0.45	0.52	0.58	0.64	C_{12}
8	Nephritis, nephritic syndrome and nephrosis	N00-N07, N17-N19, N25-N27	0.41	0.49	0.56	0.63	C_{11}
9	Influenza and pneumonia	J09-J18	0.37	0.46	0.54	0.62	C_{10}
10	Intentional self-harm (suicide)	U03, X60-X84, Y87.0	0.33	0.43	0.52	0.61	C_9
11	Septicemia	A40-A41	0.29	0.4	0.5	0.6	C_8
12	Chronic liver disease and cirrhosis	K70, K73-K74	0.25	0.37	0.48	0.59	C_7
13	Essential hypertension and hypertensive renal disease	I10, I12, I15	0.21	0.34	0.46	0.58	C_6
14	Parkinson's disease	G20-G21	0.16	0.31	0.44	0.57	C_5
15	Pneumonitis due to solids and liquids	J69	0.12	0.28	0.42	0.56	C_4
16	ICD-10 non-ranked terms	none	0.08	0.25	0.4	0.55	C_3
17	Medical terms (No ICD-10 code)	none	0.04	0.22	0.38	0.54	C_2
18	Non-medical terms	none	0	0	0	0	C_1

Table 2: 7 cases of the parameter α's combination

Case (α_1, α_2)	α_1	α_2
Case 1 (1, 1)	1	1
Case 2 (1, 0.75)	1	0.75
Case 3 (0.75, 1)	0.75	1
Case 4 (1, 0.5)	1	0.5
Case 5 (0.5, 1)	0.5	1
Case 6 (1, 0.25)	1	0.25
Case 7 (0.25, 1)	0.25	1

Table 3: Distribution of the document frequencies of terms and the number of terms in each category

Category	Sum of document frequencies of terms	Average of document frequencies of terms	Percentage of document frequencies of terms	Number of terms
C_1	4222157	324.133	0.840461229	102068
C_2	675397	51.8499	0.134444312	12418
C_3	74904	5.7503	0.014910366	1507
C_4	0	0	0	0
C_5	1238	0.095	0.000246436	13
C_6	10304	0.791	0.002051111	24
C_7	389	0.03	7.74342E-05	17
C_8	1340	0.103	0.00026674	19
C_9	106	0.01	2.11003E-05	1
C_{10}	2633	0.202	0.000524124	39
C_{11}	2571	0.197	0.000511782	20
C_{12}	4699	0.361	0.000935381	37
C_{13}	211	0.02	4.20016E-05	12
C_{14}	0	0	0	0
C_{15}	3053	0.234	0.000607729	30
C_{16}	3317	0.255	0.000660281	23
C_{17}	780	0.06	0.000155267	2
C_{18}	20520	1.5753	0.004084705	239

Table 4: Results using SVM (rbf)

Case (α_1, α_2)	TFIDF	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
(1, 0.25)	0.78494	**0.67538**	**0.78375**	**0.81484**	**0.81974**
(1, 0.5)	0.78494	0.64641	0.77372	0.80764	0.81481
(1, 0.75)	0.78494	0.63572	0.77214	0.8057	0.81365
(1, 1)	0.78494	0.62981	0.76917	0.8051	0.81397
(0.75, 1)	0.78494	0.62937	0.76813	0.80449	0.81321
(0.5, 1)	0.78494	0.63004	0.76744	0.80396	0.81222
(0.25, 1)	0.78494	0.633	0.76672	0.80333	0.81176

Table 5: Results using SVM (linear)

Case (α_1, α_2)	TFIDF	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
(1, 0.25)	0.78496	0.70442	0.84683	**0.86332**	**0.86486**
(1, 0.5)	0.78496	0.72928	**0.86053**	0.85722	0.84981
(1, 0.75)	0.78496	0.77238	0.85434	0.8473	0.8384
(1, 1)	0.78496	**0.80113**	0.84795	0.83845	0.83124
(0.75, 1)	0.78496	0.79111	0.84625	0.83763	0.8297
(0.5, 1)	0.78496	0.78291	0.8451	0.83657	0.82857
(0.25, 1)	0.78496	0.77534	0.84292	0.83547	0.82767

Table 6: Results using Naive Bayes

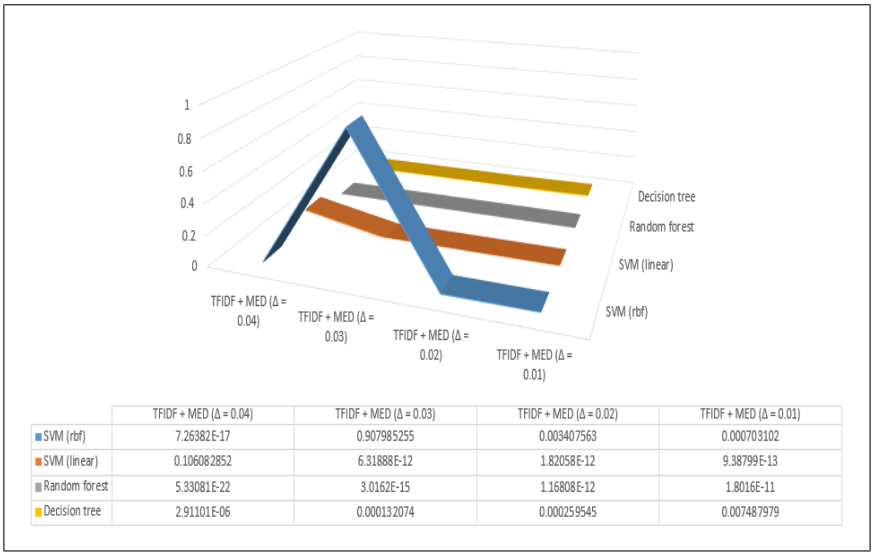
Case (α_1, α_2)	TFIDF	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
(1, 0.25)	0.87986	**0.80568**	**0.84309**	0.84995	**0.85283**
(1, 0.5)	0.87986	0.78021	0.84226	**0.85118**	0.85219
(1, 0.75)	0.87986	0.77223	0.84194	0.84976	0.8513
(1, 1)	0.87986	0.77105	0.84055	0.84789	0.84961
(0.75, 1)	0.87986	0.7633	0.83866	0.84709	0.84891
(0.5, 1)	0.87986	0.75427	0.8377	0.8468	0.84859
(0.25, 1)	0.87986	0.74611	0.83674	0.84628	0.84806

Table 7: Results using Random forest

Case (α_1, α_2)	TFIDF	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
(1, 0.25)	0.84755	0.85075	0.85559	0.85977	0.85958
(1, 0.5)	0.84755	0.8527	0.86024	**0.86158**	0.85916
(1, 0.75)	0.84755	0.85544	0.85858	0.86137	**0.85962**
(1, 1)	0.84755	0.85597	0.85999	0.85823	0.85555
(0.75, 1)	0.84755	0.85837	0.86113	0.85692	0.85565
(0.5, 1)	0.84755	0.86131	**0.86235**	0.8547	0.85406
(0.25, 1)	0.84755	**0.8654**	0.85849	0.8489	0.84005

Table 8: Results using Decision tree

Case (α_1, α_2)	TFIDF	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
(1, 0.25)	0.82661	0.82747	0.83038	0.83085	0.83125
(1, 0.5)	0.82661	0.82917	0.83083	**0.83248**	0.83061
(1, 0.75)	0.82661	0.82957	0.83048	0.83141	0.83058
(1, 1)	0.82661	0.83071	0.83247	0.83174	0.829
(0.75, 1)	0.82661	0.82894	0.83155	0.83103	**0.83142**
(0.5, 1)	0.82661	0.83166	0.8342	0.83219	0.83069
(0.25, 1)	0.82661	**0.83493**	**0.83472**	0.83004	0.8293



	TFIDF + MED (Δ = 0.04)	TFIDF + MED (Δ = 0.03)	TFIDF + MED (Δ = 0.02)	TFIDF + MED (Δ = 0.01)
SVM (rbf)	7.26382E-17	0.907985255	0.003407563	0.000703102
SVM (linear)	0.106082852	6.31888E-12	1.82058E-12	9.38799E-13
Random forest	5.33081E-22	3.0162E-15	1.16808E-12	1.8016E-11
Decision tree	2.91101E-06	0.00132074	0.000259545	0.007487979

Figure 5: The t-test's results between the TFIDF-based method and the proposed methods where the best parameter of α in each Δ was employed