Computers in Biology and Medicine 94 (2018) 1–10

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives

Susan Sabra *, Khalid Mahmood Malik, Mazen Alobaidi

Computer Science and Engineering Department, Oakland University, 2200 N. Squirrel Rd, Rochester, MI 48309, USA

ARTICLE INFO

Keywords:
Venous thromboembolism
Risk factor assessment
Natural language processing
Semantic enrichment
Sentiment analysis
Prediction through classification
Support vector machine

ABSTRACT

Venous thromboembolism (VTE) is the third most common cardiovascular disorder. It affects people of both genders at ages as young as 20 years. The increased number of VTE cases with a high fatality rate of 25% at first occurrence makes preventive measures essential. Clinical narratives are a rich source of knowledge and should be included in the diagnosis and treatment processes, as they may contain critical information on risk factors. It is very important to make such narrative blocks of information usable for searching, health analytics, and decision-making. This paper proposes a Semantic Extraction and Sentiment Assessment of Risk Factors (SESARF) framework. Unlike traditional machine-learning approaches, SESARF, which consists of two main algorithms, namely, ExtractRiskFactor and FindSeverity, prepares a feature vector as the input to a support vector machine (SVM) classifier to make a diagnosis. SESARF matches and maps the concepts of VTE risk factors and finds adjectives and adverbs that reflect their levels of severity. SESARF uses a semantic- and sentiment-based approach to analyze clinical narratives of electronic health records (EHR) and then predict a diagnosis of VTE.

We use a dataset of 150 clinical narratives, 80% of which are used to train our prediction classifier support vector machine, with the remaining 20% used for testing. Semantic extraction and sentiment analysis results yielded precisions of 81% and 70%, respectively. Using a support vector machine, prediction of patients with VTE yielded precision and recall values of 54.5% and 85.7%, respectively.

1. Background and introduction

Venous thromboembolism (VTE) is the third most common cardiovascular disorder. It affects people of both genders, at ages as young as 20 years [1]. There are two types of VTE cases: provoked and unprovoked. VTE comprises deep vein thrombosis (DVT), pulmonary embolism (PE), or both. In addition to genetic risk factors, the increasing prevalence of lifestyle and diet risk factors indicates a potential increase in the number of VTE cases in the near future. A study that was published in 2000 predicted that in the US, 201000 first lifetime cases of VTE would occur in the following year, 25% of these patients would die within 7 days of the occurrence, and death would be so rapid in 22% of all cases that there would be insufficient time for intervention to save the life of the patient [2]. Studies have shown that the fatality rate associated with PE and the rate of recurrent VTE, remain unacceptably high [3,4]. Preventive efforts should begin with lifestyle measures to reduce the risk of VTE [5,6]. Cancer, diabetes, obesity and air pollution are critical VTE risk factors that have been increasing in recent years. More importantly, they are projected to increase dramatically in the near future [2,7,8]. Lifestyle-related risk factors include air pollution, lack of sleep, travel, and stress [9–11]. A thorough analysis of VTE risk factors identifies negative changes in lifestyle, diet, and environment as main causes. Fig. 1 shows a graph that depicts a collective analysis and review of the biomedical literature on cardiovascular and VTE risk factors, which describes the increase in the number of new cases for the top three diseases, which are considered top VTE risk factors.

Obesity and diabetes are strongly correlated abnormalities that result mainly from an unhealthy diet and an abnormal lifestyle.

The graph in Fig. 1 shows a collective analysis of 15 recent studies, in which statistical analyses show the rapidly increasing numbers of cases for both diseases [2,7–11]. For cancer, the number of cases remains approximately constant. However, treatments for cancer such as radiation and chemotherapy are of major concern in terms of blood clot formation. There is a higher risk of occurrence than recurrence of VTE, regardless of whether it is unprovoked or provoked, in patients who are receiving chemotherapy and radiation treatments. Hence, prevention is necessary. Instead of relying only on treatment of VTE after its occurrence, the focus needs to be shifted to preventing such a fatal incidence

* Corresponding author.
E-mail addresses: sabra@oakland.edu (S. Sabra), mahmood@oakland.edu (K. Mahmood Malik), malobaid@oakland.edu (M. Alobaidi).

https://doi.org/10.1016/j.compbiomed.2017.12.026
Received 10 December 2017; Received in revised form 29 December 2017; Accepted 29 December 2017

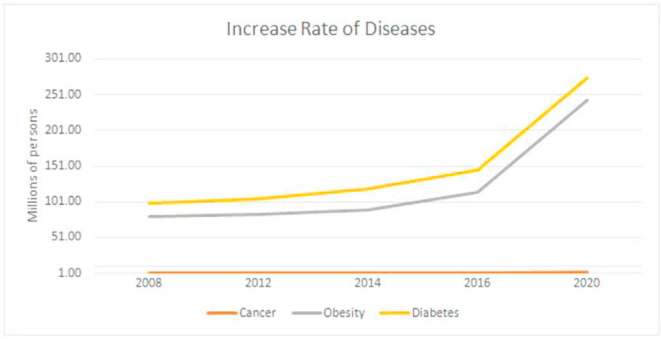


Fig. 1. Increasing rates of diseases.

by predicting it. This new approach to prevention falls under the new initiative of Precision Medicine that was declared by the U.S. government in January 2015 [12]. Precision Medicine, as defined by Schork, requires the use of a different type of clinical trial that focuses on individual, not average, responses to therapy [13]. Classical clinical trials harvest a handful of measurements from thousands of people. However, because we belong to different ethnic groups and carry different genes, treatments and medications might not have the same effects across the board. The knowledge from various cases can be refined and filtered in a patient's electronic health record (EHR) to personalize her/his diagnosis or treatment plan. Researchers need to probe the myriad genetic, environmental, and other factors [13]. Our research takes a good step in the direction of precision and preventive medicine by personalizing each VTE case. We envisage that input from both EHRs and wearable body sensors can be used for precision and evidence-based medicine for VTEs [14]. Personal and familial data can be obtained from the patient's EHR, while wearable body sensors can collect lifestyle data such as sleep and activity data. Our lives revolve around multiple personal smart devices, which can be interconnected to create the Internet of Things (IoT). This allows data from different sources to be gathered and integrated. However, in this work, we focus on semantic and sentiment analyses of clinical narratives of EHRs for early diagnosis of the first occurrence of VTE.

EHRs are valuable for medical research, but much of the information is recorded as unstructured free text, which is time-consuming to review manually during the diagnostic process. This unstructured data may contain valuable personal information about the lifestyle of the patient and the familial health history, which might be critical for evaluating the patient's potential for developing certain diseases. In the clinical domain, much of the available clinical data is recorded as free text [15]. Usually, primary care providers are the first to record patients' medical complaints in an unstructured format, which can be critical in the diagnostic process. There is a strong motivation for researchers to consider improving the overall quality of care by making such narrative blocks of information usable for searching, health analysis, and decision-making. Afzal et al. developed the Smart Extraction and Analysis System (SEAS) for clinical text extraction, which uses an automated approach that involves Natural Language Preprocessing (NLPreP), named entities, and Pattern Recognition (PR). The SEAS reduced the time and energy of human resources that were spent on manual tasks [16]. Many efforts have been made in this area to transform unstructured data into structured formats, with the aim of saving time by reducing the tedious manual effort that is required [17–19]. However, this transformation process can result in the loss of some essential content, such as hidden risk factors. Therefore, we use the unstructured form, without transforming it, to prevent loss of any critical information from the clinical narrative.

Another important strategy for maximizing the amount of knowledge that is extracted from clinical narratives is to use sentiment analysis on the adjectives and adverbs in each sentence, to fulfill a specific purpose, which depends on the field of research or study. With the emergence of social media and automatic opinion mining, it has become essential to measure the polarity of sentiments in users' inputs to better understand the audience. Adverbs in sentences intensify, affirm, or modify of meanings of the words that are associated with them. In one study, a seed set of adverbs was extracted into a graph to determine the semantic orientations of the adverbs and measure their proximities [20]. Another study used a Part Of Speech (POS) tagger to identify phrases in the input text that contain adjectives, adverbs, verbs or nouns as opinion phrases [21,22]. Hartung proposed a new approach for adjective classification, which is based on the concept of semantic classification [23]. Other researchers defined a set of general axioms based on a classification of adverbs of degree into five categories, which we describe later in Study Design [24]. Their Adjective-Adverb Combination (AAC)-based sentiment analysis technique uses a linguistic analysis of adverbs of degree to determine their relevance scores [24]. We found only two studies that

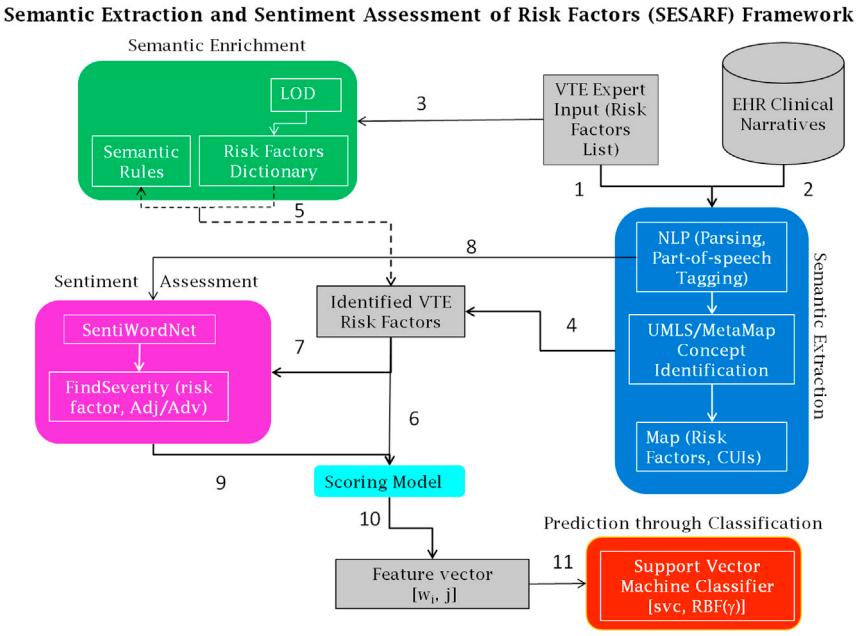


Fig. 2. SESARF framework for predicting first occurrence of VTE.

focused on linguistic features from clinical narratives. They demonstrated state-of-the-art accuracy in automatically identifying Alzheimer's disease from short narrative samples that were elicited with a picture description task, and uncovered the salient linguistic factors with a statistical factor analysis [25]. The second study on linguistic analysis showed that clinical narratives contain moderate numbers of sentiment terms [22].

The goal of our study is to evaluate the hypothesis that early diagnosis of VTE using a Natural Language Processing (NLP) approach on the clinical narratives in an EHR can be predicted with high precision and recall. This evaluation is performed by using semantic web technologies and machine learning to identify the risk factors that are essential for the diagnosis and their severity levels. To the best of our knowledge, no semantic-based data mining study has been done to predict the first occurrence of VTE. Our proposed model consists of semantic analysis of unstructured data from an EHR to properly identify the risk factor keywords and assessment of their severity levels using UMLS/MetaMap and Linked Open Data (LOD) [26,27]. The impact of the risk factor severity level on the prediction is crucial, as it provides a more accurate measure for assigning weights to the identified VTE risk factors for VTE diagnosis.

Our main contributions are summarized as follows:

i. A novel framework, namely, Semantic Extraction and Sentiment Assessment of VTE Risk Factors (SESARF), is developed. It not only matches and maps the relevant concepts but also finds adjectives and adverbs that reflect the level of severity.
ii. A new method is proposed for semantic enrichment on VTE risk factors. It uses LOD and medical semantic rules to assess the impacts of risk factors on one another in a patient's record.
iii. A new perspective is offered on selecting the feature set by combining semantic and sentiment analyses of clinical notes from EHRs to better predict VTE using a support vector machine classifier.

We aim to 1) accurately and correctly extract hidden risk factors from a clinical narrative and map them to UMLS and LOD concepts using a semantic-based approach, 2) accurately and correctly assess and score adjectives and adverbs that are associated with the identified risk concepts, 3) build a feature vector for machine-learning-based classification for predicting the likelihood of a patient having a VTE.

Our work is characterized by a) the originality of the proposed approach, b) the area of focus in preventive medicine, specifically cardiovascular diseases, and c) the pioneering attempt to prevent deaths due to first occurrences of symptomless, unprovoked VTE, which have been very rarely studied.

2. Methods

2.1. Study data

We use datasets from two main sources: Text Retrieval Conference – Clinical Decision Support (TREC CDS) and the i2b2 heart failure challenge [28,29]. TREC CDS Track 2014 is a medical external resource, which comprises biomedical documents and clinical narratives from PubMed. It is focused on the retrieval of biomedical articles that are relevant for answering generic clinical questions about medical records. This dataset is the most relevant unstructured dataset that we use to analyze clinical narrative portions from EHRs. The second dataset, which is from I2B2, is a cardiovascular-enriched cohort. It records occurrences of VTE for some patients, which will be used to confirm and validate a diagnosis. We use both datasets with minimal preprocessing: segmentation to separate sentences and filtering for pre-diagnosis.

We use a dataset of 150 clinical narratives from both sources. We train our Support Vector Machine (SVM) classifier using 5-fold cross-validation with 120 clinical narratives from patients' records, where a total of 62 patients have VTE. The remaining 30 clinical narratives are used for testing.

2.2. Study design

2.2.1. Semantic extraction and sentiment assessment of risk factors (SESARF) framework

Clinical narratives are a rich source of knowledge that need to be analyzed in the process of diagnosis or treatment, as they may contain critical information on risk factors. This free-style data format is convenient for expressing clinical activities in an EHR, but difficult to use for searching, statistical analysis, and decision support. It is very important to make such narrative blocks of information usable for searching, health analysis, and decision-making. One of the challenges to making such clinical narratives usable is the process of identifying medical terms or terms for signs and symptoms, as they appear as different synonyms and/or hyponyms. Moreover, most ontologies do not provide any relationship property among these concepts to facilitate the process of identification. More importantly, data from registry analyses and clinical trials suggest that clinicians abandon ‘silo thinking’ when diagnosing VTE by considering the integration of cardiovascular risk factors and coronary artery diseases to prevent the formation of DVT and PE [30].

To overcome these challenges, we propose a Semantic Extraction and Sentiment Assessment of VTE Risk Factors (SESARF) framework, which is illustrated in Fig. 2. It uses clinical narratives and an expert-approved list (See Appendix 1 for a complete list) of VTE known risk factors as its essential input. It consists of five tasks: 1) Semantic Extraction, 2) Semantic Enrichment, 3) Sentiment Assessment, 4) Scoring Model, and 5) Prediction through Classification.

1) Semantic Extraction is performed by our proposed algorithm *ExtractRiskFactor,* which consists of identifying and extracting VTE Risk Factors by using NLP through UMLS/MetaMap to identify UMLS concept unique identifiers (CUIs), and mapping them to the expert list of risk factors. To extract VTE risk factors from a clinical narrative, we first preprocess it by segmenting it into sentences. We use MetaMap, which is an interface for a UMLS NLP tool that recognizes terms and maps them to UMLS concept unique identifiers (CUIs) [26]. We select the options and parameters of MetaMap, as described in detail later. The MetaMap threshold is set to maximize the number of identified CUIs. The segmented clinical narrative is passed to MetaMap which in its turn takes each sentence in the narrative to find CUIs. MetaMap returns the list of all identified CUIs that match the risk factors identified in the clinical narrative omitting only the negated ones.

2) Semantic Enrichment consists of two parts: the dictionary and the semantic rules. We store every identified risk factor synonym or alternative concept in a dictionary for future use. LOD is a main source for our dictionary vocabulary, from which we extract concept synonyms from properties such as Alt, Label and Preferred Label. These are defined properties of the Simple Knowledge Organization System (SKOS) for describing the multiple definitions of similar signs and symptoms [31].

We store all of the identified risk factor synonyms and alternative concepts that are obtained from multiple sources of Linked Open Data (LOD) in a dictionary. LOD extends the Web by exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF. Examples of LOD sources that we used in this study are Diseasome, DBpedia, and Bio2RDF [27]. The dictionary that we create is updated every time the list of risk factors is updated.

Other concept properties, such as a relationship between concepts, cannot be found in LOD. For this reason and based on the literature and an expert's opinion, we compile a map of the causes and effects on each other of VTE risk factors. The likelihoods of many VTE risk factors increase with the coexistence of other risk factors. In addition, we studied each risk factor's influences, such as which other factors increase or decrease its severity level. For example, an existing study found a strong positive association between overall obesity, as measured by body mass index (BMI), and risk of diabetes [7,8]. Other studies found a relationship between hypertension and each of glucose metabolism and sleep apnea

[32,33]. We translate this cause-and-effect map into a set of semantic enrichment rules, which we apply once we have identified all VTE risk factors in a patient's record. Our semantic rules reflect the cause-and-effect map of risk factors, in addition to taking into consideration the likelihood weight of each risk factor in contributing to the diagnosis. Semantic rules also affect the severity score for each risk factor. Our semantic rules are constructed by following this structure:

Risk factor #+list of all increasing-effect risk factors #-list of all decreasing-effect risk factors #likelihood score.

The rule consists of: 1) each risk factor followed by 2) the list of other risk factors that increase its weight preceded with a "plus" sign, then followed by 3) the list of other risk factors that decrease its weight preceded with a "minus" sign, and last 4) the risk factor VTE likelihood rate assessed to affect the risk factor weight score.

Algorithm *ExtractRiskFactor* covers Semantic Extraction and Semantic Enrichment, and produces a preliminary set of scores. It identifies VTE risk factor concepts in each sentence of the clinical narrative. The last step in Algorithm I *ExtractRiskFactor* is to assign a weight to each risk factor that reflects its criticality among all VTE risk factors. We perform this step based on a preliminary set of VTE risk factors and their likelihoods, and statistical data that we gathered from the most recent studies in the literature [10,11,30,34,35].

We calculate the risk factor weight as follows:

$$\text{RiskFactorsweight} = \text{RiskFactorsweight} + \text{Likelihood\_Factor} + \text{Risk\_Factors\_Score} \quad (1)$$

Eq. (1) is an incremental calculation of each risk factor weight by considering its impact likelihood, which is based on its co-existence with other risk factors and its original score when it was identified in a sentence. This weight value is mainly used in the feature vector for prediction.

3) Sentiment Assessment or Sentiment Polarity of Adjectives and Adverbs is used to assess the severity level. Inspired by the use of sentiment analysis for automatic opinion mining, we aim to measure the severity of risk factors by quantifying the polarity of sentiments in clinical narratives, since adverbs and adjectives in sentences convey the intensity of the meaning of words that are associated with them. A Part Of Speech (POS) tagger is usually used to identify phrases in the input text that contains adjectives, adverbs, verbs, or nouns as opinion phrases, so we use a POS tagger in NLP to tokenize the adjectives and adverbs in a sentence. Our sentiment analysis algorithm *FindSeverity* focuses on evaluating adjectives and adverbs only to assess the severity levels that are associated with the identified VTE risk factors in the text.

Algorithm I: ExtractRiskFactor

```
Get_VTE_RiskFactors()
1. Input : Clinical_narrative, Riskfactors
2. Output: Risk_Factors_Score
3. Begin
4.  set Riskfactors = getRiskFactors(Riskfactors)
5.  set Dictionary = getDictionary(Riskfactors) // Call LOD endpoint to collect synonyms
6.  set narrative  =  preprocessing_text(Clinical_narrative)
7.  MetaMapApi api = new MetaMapApiImpl();
8.  api.setOptions("-y -u --negex -v -c")
9.  for sen in narrative
10. set concepts = getMetaMapConcepts(sen)
11. for concept in concepts
12. if(Riskfactors.contains(concept) == TRUE || Dictionary.contains(concept) == TRUE)
13.     ApplySemanticRules(concept)
14.     RiskFactorsweight = getRiskFactorWeight(concept)
15.     Risk_Factors_Score.insert(concept, RiskFactorsweight)
16.     End
17.   End
18. Return Risk_Factors_Score
19. End
```

To measure a severity level, we design an API call to SentiWordNet for the assessment of the positivity or negativity of an adjective or adverb that is associated with a risk factor in a sentence [36]. Once we identify a risk factor in a sentence, POS tagging allows *FindSeverity* to find the closest term (within two words around the term), whether an adverb or an adjective, for assessing the risk criticality. Through the API call to SentiWordNet, *FindSeverity* finds the correct sense of the term and labels it as either increasing or decreasing the criticality. For example, the term "significantly" increases the criticality of the risk, while the term "rarely" decreases it. In our study, we use the categories of adverbs that were defined by Benamara et al. to assess severity levels when they are modified by adverbs of degree, as follows [24]:

i. Adverbs of affirmation: these include adverbs such as absolutely, certainly, exactly, and totally.
ii. Adverbs of doubt: these include adverbs such as possibly, roughly, apparently, **and** seemingly.
iii. Strong intensifying adverbs: these include adverbs such as astronomically, exceedingly, extremely, and immensely.
iv. Weak intensifying adverbs: these include adverbs such as barely, scarcely, weakly, and slightly.
v. Negation and Minimizers: these include adverbs such as hardly.

In addition, we took into consideration the appropriateness of the increase or decrease for specific risk factors, as the adverb might work in an opposite sense. For example, lacking sleep has a negative polarity, but it obviously increases the risk. In addition, negation is handled differently, as it implies omission of the risk factor. Negation is also considered during risk factor extraction, in cases where it occurs with negation triggers (e.g., not and no). This is reflected by a score of zero.

Algorithm II: FindSeverity
Get_RiskFactors_Severity

```
1. Input: narrative, k = 4
2. Output: RiskFactor_Severity
3. Begin
4. Set Windowsize = k
5. Set RiskFactors = Get_VTE_RiskFactors(narrative)
6. For sen in narrative
7.   Set senRiskfactors = getSenRiskFactors(sen,RiskFactors)
8.   For RiskFactor in senRiskFactors
9.     Set Severity_Score = RiskFactor.weight
10.  Set AdjAdvList = GetAdjAdvs(sen, RiskFactor, windowsize)
11.  For adjadv in AdjAdvList
12.    Set Polarityscore = GetSentiWordNet(adjadv)
13.    If(((Polarityscore > 0 && RiskFactor.Polarity == Positive) ||
14.      (Polarityscore < 0 && RiskFactor.Polarity == Negative)) &&(!Severity Negative
         Rules))
15.      Set Severity_Score++;
16.    End
17.  End
18. RiskFactor_Severity.insert(RiskFactor, Severity_Score)
19. End
20.Return RiskFactor_Severity
21.End
```

FindSeverity performs Sentiment Assessment and produces a partial score to be used in the scoring model. It directly retrieves the polarity scores of adjectives and adverbs from SentiWordNet. Then, it translates them according to a severity score model and applies the rules for reflecting an increase in severity. The severity score ranges from 1 to 4, where 1 is the lowest. For example, the adverb "significantly" can be assessed with positive polarity, which means an increasing severity level for a certain risk factor. For example, suppose diabetes A1C hemoglobin is described as "significantly high", where the SentiWordNet score of "significantly" is 0.136 and that of "high" is 0.045. We score "significantly high" with the average 0.0905, which is above the highest threshold t_h=0.075 of our score range. This leads to an increase in the severity score by 2 points, as it is above the highest threshold. The score reflects a level of severity that represents the criticality of its contribution to the diagnosis. A "significantly low" expression has an average score of (– 0.059), which does not yield any increase in the severity score since it is below threshold t_l=0.04. This important step directly affects the weight of the risk factor, which is used in the later preparation of the feature vector for SVM.

4) Scoring Model: As described earlier, the scoring model is an incremental module and is split in execution between semantic extraction and sentiment assessment tasks as scores are cumulatively being calculated. Once risk factors have been identified and associated with the relevant adjectives and/or adverbs, they are assigned scores according to our scoring model. This scoring mechanism consists of two steps: First, a score is assigned to each identified risk factor and a weight is assigned based on the literature. Second, the severity level is scored by assessing the sentiment impact of an adjective and/or adverb that is associated with each risk factor. A minimum score of +1 implies that the risk factor exists, but with very low severity. The minimum score is increased by a constant k when an adverb completely affirms an adjective to increase the severity level. The value of k is calculated on the scale of scores, relative to two preset thresholds t_h and t_l, which determine the value of the increase and assign it to k. The final risk factor weights reflect the adjustments that are made to the initial weights after incorporating the severity scores.

***5) Prediction through Classification*:** We use support vector machine as a classifier to distinguish patients with VTE from those who do not present any signs or symptoms of VTE. SVM requires a vector for each patient's record. We create this vector, which consists of identified risk factors and their associated weights, including their severity levels. These components are the features for each patient's record that SVM analyzes to predict the possibility of that patient developing a blood clot and make an early diagnosis.

Our choice of classifier was based on many factors, including studies in the literature that provide evidence of the advantage of SVM over many other approaches:

a) The results demonstrated that the accuracy and generalization performance of SVM exceed those of back-propagation neural networks as the training set size decreases [37–39].
b) Kernel methods are a class of learning machines that have become increasingly popular tools in many applications, as they provide a bridge from linearity to non-linearity [40].

Risk factor weights are used for prioritization and optimization purposes, as they enhance the sparsity of data points/vectors. Similarly, we use the radial basis function (RBF) as the kernel type for SVM because it is a popular function for support vector machine classification. RBF supports the data sparsity property.

Moreover, feature selection with multivariate performance measures can be critical to successful prediction through classification of SVM accuracy and precision [41]. F1-measure is a simple technique that measures the discrimination of two sets of real numbers [42]. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates that within each of the two sets. The larger the F1-measure is, the more discriminative the feature is [42].

Table 1
SESARF experiments and parameters.

Variable Name	Optimal Value	Other Values
Adj./Adv. window size	4	[1..6]
MetaMap Threshold φ	−300	−800, −500, −200, −400
SVM c parameter	1	[0.1 … 100]
SVM RBF γ	0.6	[0.03, 1]
Iterations	92	124, 85

3. Experiments and results

3.1. Parameter settings of algorithms I & II

Before initiating our experiments, we studied many parameters. The MetaMap parameters, such as options "y" for Word Sense Disambiguation, "u" for unique abbreviation, "negex" for negated expressions, and "c" for conjunction, were set to true to evaluate a concept or term [26]. In addition, we experimented with the threshold value for the MetaMap internal scoring mechanism of sign and symptom concepts, to select the concepts that will appear in the results. Some threshold values eliminated some of the concepts from the results. We found that φ = −300 is the best threshold, as it maximizes the identification of sign and symptom concepts based on our set of other parameters. In each sentence of the clinical narrative, algorithm *FindSeverity* uses a window size of 4 to detect adverbs and adjectives that are within 2 words of an identified risk factor. As shown in Table 1, we experimented with different window sizes, which ranged from 1 to 6. We found that size 4 was optimal, as the other values yielded incorrect results, mainly because of the incorrect association of some adjective and/or adverbs with a risk factor in the sentence. Table 1 displays the different variable settings for the different experiments that we ran.

We chose optimal values of gamma and c based on results of different k-folds of cross-validation for k = [2,4,5,6]. Table 1 shows the optimal values of parameter c in range [0.1 … 100] and gamma in range [0.03 … 1]. The second column in Table 1 shows the different values of selected initial parameters that were considered, which resulted in very poor precision and recall outcomes.

3.2. Evaluation of algorithms I & II

Now that we have set the parameters for MetaMap, *ExtractRiskFactor* and *FindSeverity,* we can evaluate both algorithms with three performance metrics, namely Precision, Recall and F1-measure, on a training dataset of 120 clinical narratives. Let TP denote true positive, FP false positive, FN false negative, TN true negative, P Precision, and R Recall. We calculate the F1-measure [43]:

$$P = TP / (TP + FP) \quad (2)$$

$$R = TP / (TP + FN) \quad (3)$$

$$F1 - measure = 2PR/(P + R) \quad (4)$$

Both *ExtractRiskFactor* and *FindSeverity* cumulatively produce a feature vector for each clinical note. These vectors compose a matrix of risk factor weights, which are used to training the SVM for prediction. We use this matrix to train our classifier until all parameters have been obtained and all risk factor weights have been optimized. The evaluation results for Algorithms I & II are illustrated in Table 2.

Table 2
Evaluation of semantic and sentiment algorithms.

	Precision	Recall	F1-measure
Algorithm I: *ExtractRiskFactor* Extraction of Risk Factors	81%	62%	70%
Algorithm II: *FindSeverity* Scoring of severity levels	71%	64%	67%

ExtractRiskFactor yielded a precision of 81% in identifying and extracting VTE risk factors from clinical narratives, while *FindSeverity* yielded a precision of 71% in identifying and assessing the severity levels of risk factors.

Table 3
Parameter optimization and evaluation of SESARF with SVM prediction accuracy in %.

γ	c					
	Training Phase Prediction Accuracy in %					
	0.1	0.5	1	10	50	100
0.03	52	60	59	61	60	63
0.3	54	61	61	61	66	66
0.6	56	68	70	69	69	69
0.9	59	63	66	66	66	66
1	57	67	68	69	69	69
10	57	61	68	68	68	68

3.3. SVM parameter settings

We use the LIBSVM library to implement our SVM classifier, with options that are described later, such as "svc" for support vector classification as the type, and options that specify the kernel-type function and the degree [44]. Considering the size of our training data and to condense the information properly, SVM provides a sparse representation. We use the Radial Basis Function (RBF) exp (-γ*|u-v|2) as a kernel function. Gamma is the parameter of a Gaussian Kernel for handling non-linear classification. We express RBF as $f(x) = w_i h_j(x)$, where i is the number of risk factors [1 to 31] and j is the number of training cases [1 to 120]. We set weight parameter w_i to the vector of risk factor weights that was calculated by Algorithms *ExtractRiskFactor* and *FindSeverity*.

We trained our SVM with a dataset of 120 clinical notes, where 62 notes are labeled positive for VTE. Table 3 illustrates the prediction accuracy of SVM with SESARF during the different experiments in the training phase, in which the SVM parameters are optimized.

We use the measure of accuracy to evaluate the performance of our SVM during parameter optimization only. Accuracy is the total proportion of correctly identified labels and is calculated as follows [45]:

$$Accuracy = (TP + TN)/ TP + TN + FP + FN \quad (5)$$

After multiple experiments during the training phase, c = 1 yielded 70% as the best prediction accuracy of our SVM with 5-fold cross-validation and an optimal value of γ = 0.6 after 92 iterations. Table 4 shows the precision, recall, and F1-measure for SVM performance during the training phase, as described above.

We use 30 clinical narratives that were not in the training dataset for testing. To maintain consistency in the evaluation measures, we evaluate our SVM with the same three measures: Precision, Recall, and F1-measure, as shown in Table 5. SVM prediction through classification testing yielded a precision of 54.5% and a recall of 85.7% with c = 1 and γ = 0.6.

We ran multiple experiments to find the optimal combination of parameters for each phase of this research. MetaMap threshold and window size were very carefully set, since they form an initial

Table 4
SVM Performance Evaluation (Training Phase). With c = 1, γ = 0.6

	Precision	Recall	F1-Measure	Accuracy
SVM Prediction + SESARF	75.5%	61.5%	67.8%	69.8%

SVM Evaluation.

Table 5
SVM Evaluation Experiments (Testing Phase). c = 1, γ = 0.6.

	Precision	Recall	F1-measure	Accuracy
SVM Prediction	50.0%	28.6%	36.4%	70.8%
SVM Prediction + SESARF	54.5%	85.7%	66.7%	75.0%
SVM Prediction + Dimensionality Reduction	26.7%	57.1%	36.4%	41.7%
SVM Prediction + SESARF + Dimensionality Reduction	28.6%	57.1%	38.1%	45.8%

parameter set, which is necessary for the successful execution of the next phase. Low precision and recall were obtained initially due to 1) the MetaMap threshold and 2) the window size value in the FindSeverity algorithm. First, the MetaMap threshold, when not set appropriately, resulted in many risk factors in the clinical narratives being missed, which affected the precision and recall for detecting adjectives and adverbs. When a risk factor is not detected in a sentence, no associated adjectives or adverbs are detected. Second, the window size for detecting an adjective or adverb that is near the identified concept of a risk factor in a sentence is critical. When set to a small value, for example 1 or 2, it might miss an adjective or adverb that is near the risk factor concept, especially if there are articles or pronouns in between them. When set to a large value, in the range 3–6 inclusive, adjectives and adverbs in the sentence are detected, but may not be associated with the identified risk factor concept in that sentence. In some cases, disassociation took place when there was more than one risk factor concept in the same sentence. After multiple experiments were conducted to determine the optimal values of c and γ, we extended our evaluation of SVM performance by applying an F-score dimensionality reduction technique to the features vector. The F-score dimensionality reduction technique selected 26 out of 31 features based on their high F-scores, which indicate they are the more discriminative features [42]. We also used LIBSVM to implement the F-score dimensionality reduction technique as a comparative baseline for our SVM. Table 5 lists the evaluation results for each of the following experiments: 1) SVM prediction without SESARF or dimensionality reduction, 2) SVM prediction with SESARF, 3) SVM prediction with dimensionality reduction, and 4) SVM prediction with SESARF and dimensionality reduction. We ran the experiments on 30 clinical notes; c and γ were set to optimal values 1 and 0.6, respectively, as described in Table 3.

The dimensionality reduction technique, namely F-score, was applied within the SVM library for feature selection [42]. It yielded a testing accuracy value of 72% with 26 selected features.

The result indicates that feature reduction did not deliver a more precise prediction, compared to using all features. Every feature plays an important role in the evaluation of each patient's record for a more accurate diagnosis. Due to the absence of a comparative baseline in the literature for assessing our framework, we also experimented with running a simple experiment of SVM prediction in which only risk factor identification was used, without any semantic enrichment or sentiment analysis. This experiment yielded an SVM prediction recall of 28.6%, in comparison to 85.7% recall for our SESARF framework, while precision improved from 50% to 54.5% using SESARF. Another SVM prediction experiment without the SESARF framework but with F-score dimensionality reduction yielded a precision of 26.7% and a recall of 57.1%. This result indicates that dimensionality reduction has a negative impact on the results of SVM prediction. In addition, when combining F-score dimensionality reduction with SVM and SESARF, the results were constant in terms of recall but slightly improved in terms of precision. These experimental results show that the combination of our SESARF framework with SVM performed better than all other combinations: it achieved 30% improvement in the prediction of VTE patients, with an F1-measure of 66.7%. This shows that semantic and sentiment analyses can be very valuable in predicting a diagnosis.

4. Discussion

Venous thromboembolism is the third most common and fatal cardiovascular condition. It can be prevented with the appropriate use of effective state-of-the-art technologies in high risk individuals. The purpose of this study is to identify these high risk individuals before a formal diagnosis to prevent fatalities. Our novel framework, namely, Semantic Extraction and Sentiment Assessment of VTE Risk Factors (SESARF), combines semantic web technologies and machine learning to predict such occurrence. It matches and maps the relevant concepts of VTE risk factors. It also finds adjectives and adverbs that reflect the level of severity. The results in Tables 2 and 5, show that SESARF effectively extracts the VTE risk factors from clinical narratives and assesses their severity levels. The second row in Table 5 shows the results of our proposed new method for semantic enrichment on VTE risk factors using LOD and medical semantic rules. This method assesses the impact of risk factors on one another in a patient's record. Table 2 displays the precision and recall of both algorithms: ExtractRiskFactor and FindSeverity. These two algorithms incrementally build the feature vector for the predictor. In Table 5 when comparing the second row to the first row, results show that the combination of semantic and sentiment analyses on clinical notes from EHRs is better to predict VTE using a support vector machine classifier. In Ref. [46], they devised a new semantic enrichment algorithm for automatically generalizing the lexical patterns found in the encyclopedia entries. They extracted more than 2600 new relationships that did not appear in WordNet [47] originally to semantically enrich the general patterns extraction. The precision to find such patterns and relationships yielded 60–70% which compared to our semantic extraction and enrichment precision of 81% indicates that our algorithm yields a minimum of 10% better precision.

Our results demonstrate that such prediction is feasible and accurate. In this study, we carry out three different analyses on the data: 1) general, 2) gender-based, and 3) age-based. Our test dataset consists of 57% female patients, and 52% of all patients are aged 40 or above. We make the following observations:

1. Our algorithms and SVM classifier were able to correctly predict the likelihood of a patient developing a VTE with high accuracy and confidence.
2. Age group results did not contribute to any meaningful observations.
3. We identified a combination of predictors for VTE. We found there is a strong association between the occurrence of VTE and the combination of the following three risk factors: diabetes, obesity, and smoking. This combination seems to be a significant predictor for the development of VTE when co-occurring with lack of sleep.

Based on the valuable input from our expert, we were able to determine many VTE risk factors that some are rarely studied, specifically the lifestyle related ones. In comparison with the study in Ref. [48], we found that we shared 14 VTE risk factors. Their complete list consisted of only 21 risk factors while we enlisted a total of 31. They

developed a new risk prediction model to quantify the absolute risk of thrombosis at 1 and 5 years. The algorithm is based on simple clinical variables which are routinely recorded in general practice records. However, they completely omitted the unstructured data part of the EHR which may have more risk factors than the ones they enlisted. Our aim is to predict the occurrence of VTE before it happens. This is why we rely more on primary care electronic health records. Upon examination and analysis of the study in Ref. [49], we found that our study was similar in terms of risk factors data such as age, gender, and lifestyle information; however their prediction of mortality rate was based on analysis of secondary care data yielding two times higher rate than primary care. That study based on coded data may return a differing result to one based on laboratory results [49]. Similarly, this implication applies when comparing patients diagnosed in primary versus secondary care. This indicates that since the mortality rate is higher at secondary care, then prediction should take place at an earlier phase. Our prediction model uses primary care data to prevent such fatalities.

The results of our different experiments show that using our SESARF framework, patients can be correctly diagnosed by relying only on text of clinical notes as they may contain most of the structured data as well. These results signify reinforcement for any other method of professional diagnosis. Hence, our tool can be EHR-embedded to provide clinical decision support and alert primary care providers of any potential VTE risks for their patients.

Extensive testing can be done to study the different combinations of risk factors, to obtain a more rigorous assessment of predictors. However, to study all possible combinations of risks, a more comprehensive dataset is needed, which can be obtained as part of our future work.

5. Conclusions

This study highlights the possibility of preventing fatalities due to VTE and reducing diagnostic errors that are caused by disregarding the clinical narrative portion of a patient's EHR. Our proposed model contributes to the field of preventive medicine by predicting the occurrence of first-time symptomless VTE with 54.5% precision and 85.7% recall. This prediction may prevent many patients from having a heart or brain stroke, or a pulmonary embolism, which can be fatal. Our predictive framework uses a combination of the latest technologies for semantic and sentiment analysis, medical ontologies through UMLS and a support vector machine for prediction through classification. In addition to applying the same approach to other diseases, our future work will include the prediction of VTE location and type, such as DVT or PE, through a more extensive analysis of risk factors and the use of wearable body sensors to collect lifestyle data. Moreover, as scientists and physicians make additional discoveries on linking risk factors, we plan to extend our semantic enrichment rules according to additional evidence-based knowledge on the causes and effects of risk factors and the combinations of VTE risk factors.

Summary

Venous thromboembolism (VTE) is the third most common cardiovascular disorder. It affects people of both genders at ages as young as 20 years. The increased number of VTE cases with a high fatality rate of 25% at first occurrence makes preventive measures essential. Mining clinical narratives in patients' EHRs can improve the accuracy of diagnosis by finding hidden risk factors that are not easily measured in routine clinical visits. We propose the Semantic Extraction and Sentiment Assessment of Risk Factors (SESARF) framework, which not only matches and maps the concepts of VTE risk factors but also finds adjectives and adverbs that reflect the level of severity using a semantic- and sentiment-based approach for analyzing the clinical narratives of electronic health records (EHRs) to predict a diagnosis of VTE. The SESARF framework first detects VTE risk factors through UMLS/MetaMap to extract signs and symptoms using semantic analysis and rules. Then, it assesses the severity level that is associated with each risk factor by sentiment analysis of the clinical narratives in the patient's record. The prediction classifier support vector machine is trained with 120 clinical narratives from patient EHRs, of which 62 are labeled as positive for VTE. Precision evaluation of our proposed algorithms for semantic and sentiment analysis yielded precisions of 81% and 70%, respectively. SVM prediction through classification of patients in terms of VTE yielded a precision of 54.5% and a recall of 85.7%.

The NLP approach, which uses semantic and sentiment analyses, can identify a patient's likelihood of developing VTE early in the process of diagnosis, to initiate treatment as soon as possible to prevent fatality. Our proposed model contributes to the field of preventive medicine by predicting the occurrence of first-time symptomless VTE with high precision. Early prediction will prevent many patients from suffering from a heart or brain stroke or a pulmonary embolism, which can be fatal. This study highlights the possibility of preventing fatalities due to VTE and reducing diagnostic errors that are caused mainly by disregarding the clinical narrative portion of a patient's EHR. Our predictive framework uses a combination of the latest technologies for semantic and sentiment analysis, medical ontologies through UMLS and a support vector machine for prediction through classification. Our approach can also be applied to many other diseases and contribute to clinical decision support systems by helping primary care physicians provide a diagnosis. In addition to applying the same approach to other diseases, our future work will include the prediction of VTE location and type, such as DVT or PE, through a more extensive analysis of risk factors and utilization of wearable body sensors for collecting lifestyle data. Moreover, as scientists and physicians make additional discoveries on linking risk factors, we plan to extend our semantic enrichment rules according to additional evidence-based knowledge on the causes and effects of risk factors and the combinations of VTE risk factors.

Acknowledgements

The authors would like to thank Dr. Hamid Sattar from Detroit Mercy Hospital for his collaboration and for giving us his constant feedback as an expert in the cardiovascular field.

We would also like to acknowledge NIH for the data access through Informatics for Integrating Biology and the Bedside (i2b2), which is an NIH-funded National Center for Biomedical Computing that is based at Partners HealthCare System. The i2b2 Heart Failure Challenge Dataset is very valuable for our research and is cited as follows: Stubbs, A., Kotfila, C., Xu, H., and Uzuner, O. (2015). "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2". J Biomed Inform. 2015 Jul 22. pii: S1532-0464(15)00140-9. https://doi.org/10.1016/j.jbi.2015.07.001.

Appendix A

List of VTE main risk factors:

Obesity
Diabetes
Varicose veins
Hypertension
Long flights/rides
Cancer
Age
Hormonal therapy
Pregnancy

Hyperthyroidism
Surgery
Trauma/Hospitalization
Immobility (bed rest)
Air Pollution
Heart disease
Hypercholesterolemia
Height
Postpartum
Contraceptive
Kidney disease
Alcohol abuse
Drug abuse
Smoking
Insomnia
Stress
Liver disease
Lymphoma
Anemia
COPD (Compulsive Obstruction Pulmonary Disease)
Thrombophilia (genetic disorder factor V Leiden, prothrombin gene mutation G20210A, protein C and S deficiency, and anti-thrombin deficiency)
Diet

References

[1] A. Elias, L. Mallard, M. Elias, C. Alquier, F. Guidolin, B. Gauthier, H. Boccalon, A single complete ultrasound investigation of the venous network for the diagnostic management of patients with a clinically suspected first episode of deep venous thrombosis of the lower limbs, Management 1 (2003) 5–10.

[2] J.A. Heit, M.D. Silverstein, D.N. Mohr, T.M. Petterson, W. O'Fallon, L. Melton III, Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case-control study, Arch. Intern. Med. 160 (6) (2000) 809–815. https://doi.org/10.1001/archinte.160.6.809.

[3] F.A. Anderson, F.A. Spencer, Risk factors for venous thromboembolism, Circulation 107 (2003). I–9–I–16, https://doi.org/10.1161/01.CIR.0000078469.07362.E6.

[4] F.A. Spencer, J.M. Gore, D. Lessard, J.D. Douketis, C. Emery, R.J. Goldberg, Outcomes after deep vein thrombosis and pulmonary embolism in the community: the worcester venous thromboembolism study, Arch. Intern. Med. 168 (2008) 425–430. https://doi.org/10.1001/archinternmed.2007.69.

[5] W. Ageno, C. Becattini, T. Brighton, R. Selby, P.W. Kamphuisen, Cardiovascular risk factors and venous thromboembolism a meta-analysis, Circulation 117 (1) (2008) 93–102.

[6] J. Douketis, A. Tosetto, M. Marcucci, T. Baglin, B. Cosmi, M. Cushman, et al., Risk of recurrence after venous thromboembolism in men and women: patient level meta-analysis, BMJ 342 (2011). Retrieved from, http://www.bmj.com/content/342/bmj.d813.abstract, https://doi.org/10.1161/CIRCULATIONAHA.107.709204.

[7] J.M. Chan, E.B. Rimm, G.A. Colditz, M.J. Stampfer, W.C. Willett, Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men, Diabetes Care 17 (9) (1994) 961–969.

[8] C.M.Y. Lee, R.R. Huxley, R.P. Wildman, M. Woodward, Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis, J. Clin. Epidemiol. 61 (7) (2008) 646–653. https://doi.org/10.1016/j.jclinepi.2007.08.012.

[9] J. Tsai, A.M. Grant, M.G. Beckman, S.D. Grosse, H.R. Yusuf, L.C. Richardson, Determinants of venous thromboembolism among hospitalizations of US adults: a multilevel analysis, PLoS One 10 (4) (2015), e0123842. https://doi.org/10.1371/journal.pone.0123842.

[10] J.H. Scurr, S.J. Machin, S. Bailey-King, I.J. Mackie, S. McDonald, P.D.C. Smith, Frequency and prevention of symptomless deep-vein thrombosis in long-haul flights: a randomised trial, Lancet 357 (9267) (2001) 1485–1489. https://doi.org/10.1016/S0140-6736(00)04645-6.

[11] M. Franchini, A. Guida, A. Tufano, A. Coppola, Air pollution, vascular disease and thrombosis: linking clinical data and pathogenic mechanisms, J. Thromb. Haemostasis 10 (12) (2012) 2438–2451. https://doi.org/10.1111/jth.12006.

[12] Precision Medicine Initiative. https://obamawhitehouse.archives.gov/precision-medicine. Accessed on 04 January 2017.

[13] N.J. Schork, Personalized medicine: time for one-person trials, Nature 520 (7549) (2015) 609–611.

[14] S. Sabra, K. Mahmood, M. Alobaidi, Semantic-based approach for predicting venous thromboembolism using Kohonen self organized map neural network, in: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2016, pp. 106–109. https://doi.org/10.1109/BHI.2016.7455846.

[15] X. Zhou, H. Han, I. Chankai, A. Prestrud, A. Brooks, Approaches Applied to text mining for clinical medical records, Proc. 2006 ACM Symp. Comput. ACM (2006) 235–239.

[16] M. Afzal, M. Hussain, W.A. Khan, T. Ali, A. Jamshed, S. Lee, Smart extraction and analysis system for clinical research, Telemed. J. e Health 23 (2017) 404–420. https://doi.org/10.1089/tmj.2016.0157.

[17] S. Ananiadou, J. McNaught, Text Mining for Biology and Biomedicine, Artech House Boston, London, 2006.

[18] H.D. Tolentino, M.D. Matters, W. Walop, B. Law, W. Tong, F. Liu, P. Fontelo, K. Kohl, D.C. Payne, A UMLS-based spell checker for natural language processing in vaccine safety, BMC Med. Inf. Decis. Making 7 (3) (2007).

[19] K. Tomanek, J. Wermter, U. Hahn, A reappraisal of sentence and token splitting for life sciences documents, Stud. Health Technol. Inf. 129 (2007) 524.

[20] S.B. Potemkin, G.E. Kedrova, Exploring Semantic Orientation of Adverbs, 2011. CDUD'11–Concept Discovery in Unstructured Data 71.

[21] S.S. Htay, K.T. Lynn, Extracting product features and opinion words using pattern knowledge in customer reviews, Sci. World J. (2013), e394758. https://doi.org/10.1155/2013/394758.

[22] Y. Deng, M. Stoehr, K. Denecke, (n.d.). Retrieving Attitudes: Sentiment Analysis from Clinical Narratives.

[23] M. Hartung, Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns, 2016 [WWW Document], http://archiv.ub.uni-heidelberg.de/volltextserver/20013/ (accessed 10.2.16).

[24] F. Benamara, C. Cesarano, A. Picariello, D.R. Recupero, V.S. Subrahmanian, Sentiment analysis adjectives and adverbs are better than adjectives alone, in: ICWSM. Citeseer, 2007. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.1338&rep=rep1&type=pdf.

[25] K.C. Fraser, J.A. Meltzer, F. Rudzicz, Linguistic features identify Alzheimer's disease in narrative speech, J. Alzheim. Dis. 49 (2015) 407–422.

[26] MetaMap. https://metamap.nlm.nih.gov/. Accessed on 30 March 2017.

[27] Linked Open Data. http://linkeddata.org/. Accessed on 17 January 2017.

[28] S. Sungbin, Choi, J. Choi, SNUMedinfo at TREC CDS Track 2014: Medical Case Based Retrieval Task, Seoul National Univ (Republic of Korea), 2014.

[29] A. Stubbs, C. Kotfila, H. Xu, O. Uzuner, Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inform (2015), https://doi.org/10.1016/j.jbi.2015.07.001, 2015 Jul 22. pii: S1532–0464(15)00140-9.

[30] G. Piazza, S.Z. Goldhaber, Venous thromboembolism and atherothrombosis an integrated approach, Circulation 121 (19) (2010) 2146–2150. https://doi.org/10.1161/CIRCULATIONAHA.110.951236.

[31] The SKOS Content. https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_altLabel.html Accessed on 17 January 2017.

[32] K.-T. Chou, C.-C. Huang, Y.-M. Chen, K.-C. Su, G.-M. Shiao, Y.-C. Lee, H.-B. Leu, Sleep apnea and risk of deep vein thrombosis: a non-randomized, pair-matched cohort study, Am. J. Med. 125 (4) (2012) 374–380.

[33] R.P. Millman, S. Redline, C.C. Carlisle, A.R. Assaf, P.D. Levinson, Daytime hypertension in obstructive sleep apnea: prevalence and contributing risk factors, Chest 99 (4) (1991) 861–866.

[34] S.Z. Goldhaber, H. Bounameaux, Pulmonary embolism and deep vein thrombosis, Lancet 379 (9828) (2012) 1835–1846.

[35] J.T. Lee, K.D. Lawson, Y. Wan, A. Majeed, S. Morris, M. Soljak, C. Millett, Are cardiovascular disease risk assessment and management programmes cost effective? A systematic review of the evidence, Prev. Med. 99 (2017) 49–57. https://doi.org/10.1016/j.ypmed.2017.01.005.

[36] SentiWordNet tool. http://sentiwordnet.isti.cnr.it/. Accessed on 30 March 2017.

[37] M.A. Mohandes, T.O. Halawani, S. Rehman, A.A. Hussain, Support vector machines for wind speed prediction, Renew. Energy 29 (6) (2004) 939–947.

[38] K.-S. Shin, T.S. Lee, H. Kim, An application of support vector machines in bankruptcy prediction model, Expert Syst. Appl. 28 (1) (2005) 127–135.

[39] S. Ertekin, L. Bottou, C. Lee Giles, Fast Classification with Online Support Vector Machines, 2016 web.mit.edu/seyda/www/Papers/GHC06_ACMSRC_abstract.pdf N.p., n.d. Web.

[40] G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (Jul) (2010) 2079–2107.

[41] Q. Mao, I.W.H. Tsang, A Feature Selection Method for Multivariate Performance Measures, IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (Sep. 2013) 2051–2063.

[42] Chen, C.-J. Lin, Combining SVMs with Various Feature Selection Strategies, SpringerLink, 2006, pp. 315–324.

[43] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 233–240. Retrieved from, http://dl.acm.org/citation.cfm?id=1143874.

[44] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011), 27:1–27:27, https://doi.org/10.1145/1961189.1961199.

[45] Y. Tang, Y.Q. Zhang, N.V. Chawla, S. Krasser, SVMs Modeling for Highly Imbalanced Classification, IEEE Trans. Sys. Man Cybern. Part B (Cybern.) 39 (1) (2009) 281–288. https://doi.org/10.1109/TSMCB.2008.200290.

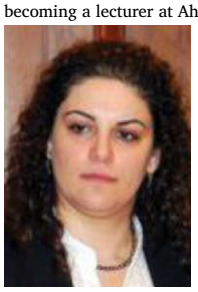
[46] M.R. Casado, E. Alfonseca, P. Castells, Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from wikipedia, Data Knowl. Eng. 61 (3) (2007) 484–499. ISSN: 0169–023X.

[47] https://wordnet.princeton.edu/.

[48] J. Hippisley-Cox, C. Coupland, Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study, BMJ 343 (Aug. 2011) p. d4656.

[49] A.M. Gallagher, T. Williams, H.G.M. Leufkens, F. de Vries, The Impact of the Choice of Data Source in Record Linkage Studies Estimating Mortality in Venous Thromboembolism, PLoS One 11 (2) (Feb. 2016), e0148349.

Susan Sabra is a PhD candidate at Oakland University, MI. Her research area is in health informatics for preventive medicine using Semantic Web technologies. She held a part-time lecturer position at Oakland University and Lawrence Technological University in Michigan. Also, Susan worked as an IT system analyst for Saudi Aramco in Saudi Arabia before becoming a lecturer at Ahlia University in the kingdom of Bahrain.



Dr. Malik is currently working as an assistant professor at School of Engineering and Computer Science, Oakland University, MI, USA. His research interests include integrated area of Health Informatics, Cognitive Mobile Computing, Autonomous Decentralized Systems (ADS), Internet of Things and Semantic Web. Research topics comprise of Ontology based Information Extraction for Health Informatics, Semantic based Information Security & Data Loss Prevention, ADS based architecture and its applications, Big Data analytics using Linked Data, Semantic based Web Filtering and Semantic based Cloud Robotics.

Mazen Alobaidi is software engineer lead at Micro Focus International cooperation and a Ph.D. student at department of Computer Science Engineering in Oakland University, USA. His research interests include Semantic Information Extraction, Learning Ontologies and Sematic Web technologies.