Sādhanā (2021)46:77
https://doi.org/10.1007/s12046-021-01593-5

Ensemble approach for identifying medical concepts with special attention to lexical scope

ANUPAM MONDAL* and DIPANKAR DAS

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
e-mail: link.anupam@gmail.com

MS received 1 November 2019; revised 31 December 2020; accepted 24 February 2021

Abstract. Health-care services are implanted by deploying roots of information extraction techniques. This extraction process is laborious and time-consuming due to unavailability of medical experts. Thus, in the present task, we were motivated to develop an automated extraction system for identifying medical and non-medical concepts. These concepts help to extract the key information from medical corpora. Not only medical concepts but also their non-medical counterparts are equally important for diagnosis purposes. Hence, we have employed three different approaches such as unsupervised, supervised, and their combined ensemble version to identify both medical and non-medical terms (words/phrases). The unsupervised module consists of two phases: parts-of-speech (POS) tagging followed by searching in a domain-specific lexicon, namely WordNet of Medical Event (WME 3.0). On the other hand the supervised module is designed by two machine learning classifiers, namely Naïve Bayes and Conditional Random Field (CRF) along with various features like category, POS, sentiment, etc. Finally, we have combined the important outcomes of unsupervised and supervised modules and developed two versions of ensemble module (Ensemble-I and Ensemble-II). All the modules identify uni-gram, bi-gram, tri-gram, and more than tri-gram medical concepts and separate non-medical words or phrases in a context. In order to evaluate all modules of concept identification system, we have prepared an experimental dataset. It has been split into three parts, namely training, development, and test. We observed that ensemble module provides better output in contrast with individual modules and Ensemble-I outperforms Ensemble-II in identifying medical concepts consisting of all possible *n*-grams. The result analysis shows that the *F*-measures of 0.91 and 0.94 have been obtained for identifying medical concepts and non-medical words/phrases using both of the ensemble modules, respectively. The present research reports the initial steps to build an automated concept identification framework in health-care. This system assists in designing various domain-specific applications like annotation, categorization, recommendation system, etc.

Keywords. Health-care; medical concepts; lexicon; lexical scope; wordNet of Medical Events; ensemble.

1. Introduction

During last few decades, retrieval and extraction of medical information are largely being observed in the web in textual form.[1,2] Such important information are primarily scattered in the form of medical reports, discharge summaries, prescriptions, etc. Conversion of unstructured information to a structured representation is considered as an emerging research in health-care [1]. Structured information helps the doctors as well as patients in their decision-making processes related to treatment. It also assists in recognizing the medical concepts (MC) and their associated knowledge [2, 3]. Primarily, the information appears in the form

[1] http://www.prmoment.in/category/pr-news/survey-shows-that-49-of-indians-use-the-internet-for-health-information.

[2] http://www.nbcnews.com/id/3077086/t/more-people-search-health-online/#.XVUZl599LqM.

*For correspondence

Published online: 12 April 2021

of MC and their categories and relations. In the present work, we have focused on identifying MC using an ensemble approach. However, the identification of MC requires attention to solve the following challenges: (A) how to identify the lexical scope (text span) of an MC from a context [4]? (e.g., multi-word MC, “A [chronic^{MC1} cough^{2}]^{MC} is more than just an annoyance.”) (B) How to recognize MC and isolate them from rest of the context? (e.g., “Abdominal_pain^{MC} is a [sign$^{non-medicalconcept}$ of early pregnancy^{MC}]^{MC}”. In general, context is referred to as an individual sentence of a medical corpus.

In order to address the mentioned challenges, we have developed an ensemble module by combining unsupervised and supervised modules. The unsupervised module deals with a domain-specific lexicon (WME 3.0) [5] and several knowledge-bases like SentiWordNet, SenticNet, etc. [6]. On the other hand, the supervised module extracts various features (e.g. *n*-gram, term frequency (TF), parts-of-speech (POS), similar sentiment words (SSW), polarity score, sentiment tag, etc.) related to MC and employ them into two well-known machine learning classifiers: Naïve Bayes and Conditional Random Field (CRF) [7]. In the present work, we primarily considered the MC of lexical size up to tri-grams. The unsupervised module fails to capture MC containing lexical scopes more than bi-gram whereas CRF outperforms Naïve Bayes in case of MC containing lexical scopes of all sizes. We have noticed that *n*-gram, TF, polarity score, and sentiment tag features play the crucial roles. Therefore, instead of individual modules, in order to enhance the results, we have developed two versions of ensemble module by combining the output of both the classifiers of supervised module along with unsupervised module. First, ensemble module ensures priority to higher-order *n*-grams rather than the performance of individual classifiers while dealing with lexical scopes. The ensemble module depends on the majority voting technique, which deals with the outputs extracted from unsupervised and supervised classifiers. Thus, priority is given more to the extraction of higher-order *n*-grams than the approach or output of the classifiers because our focus is to trace lexical scope that indeed deals with *n*-grams. Second, ensemble module is designed by employing stacking technique on unsupervised and supervised modules in the presence of an artificial neural network (ANN) classifier [8, 9]. It has been observed that both of the ensemble modules gain performance while utilizing the advantage of individual modules. In order to validate all of the modules, we have prepared an experimental dataset collected from three different resources: SemEval 2015 Task-6,[3] MedicineNet,[4] and Ohsumed dataset.[5] The dataset was annotated by a group of medical experts and linguists and we achieved an agreement score of 0.88 and 0.90 (kappa) for different classes of MC and distribution of medical and non-medical terms, respectively. *F*-measure of ensemble modules for various lexical scopes of MC provided averages 0.91. In addition, *F*-measure score of 0.97 has been offered by ensemble modules for identifying MC from the test dataset.

In order to address our second challenge, all three modules help to distinguish MC from non-medical counterparts within a context only at lexical level. It has been observed that WME 3.0 lexicon of unsupervised module and TF feature of supervised module play the specialized roles in order to deal with this challenge. Finally, ensemble modules were also able to separate the lexical scopes of MC from non-medical counterparts within a context.

The rest of the paper is as follows. Section 2 presents the background study carried out in this field. Section 3 describes how we have prepared our experimental dataset for the current research. Thereafter, various concept identification frameworks for MC have been discussed in Section 4. Section 5 describes the validation steps in the form of result analysis for the proposed frameworks in detail. Finally, the conclusion and final remarks appear in Section 6.

2. Background

Information extraction from medical corpus has been considered as a challenging research in health-care from the past few decades [10, 11]. Primarily the challenge is due to the unavailability of the structured corpus and a small number of involvements of doctors and medical practitioners [12]. The researchers have realized the importance of designing an automated identification system of MC to overcome this challenge.

Pakhomov *et al* [13] designed an automated system to annotate concepts from a publicly available clinical corpus with the label of POS and anaphoric relations [13, 14]. Roberts *et al* [15] prepared a well-known annotated corpus, namely Clinical E-Science Framework (CLEF), with medical knowledge like named entities and their relations, modifiers, and co-references [15, 16]. They were motivated to design this corpus for the capture, integration, and presentation of clinical information from the corpus, which is used in clinical research. In this regard, we have developed an unsupervised module for identifying MC from the corpus. The unsupervised module has been designed by employing a POS tagger along with a domain-specific lexicon, namely WordNet of Medical Events (WME) [5].

Tsuruoka and Tsujii [17] developed a GENIA tagger for biomedical text such as MEDLINE abstracts to annotate the concept as their base forms, POS, chunk, and named entity. Moreover Mandel [18] has developed an oncology corpus, viz. PeenBioIE, which consists of 1414 PubMed abstracts on cancer. The corpus primarily focuses on molecular

[3] http://alt.qcri.org/semeval2015/task6/.

[4] http://www.medicinenet.com/script/main/hp.asp.

[5] http://davis.wpi.edu/xmdv/datasets/ohsumed.html.

genetics, comprising approximately 327,000 words of biomedical text, tokenized and annotated for paragraph, sentence, and POS, and 24 types of biomedical named entities in five categories of interest. In contrast with this contribution, we have designed a supervised module using various concepts as well as contextual features with machine learning classifiers.

In addition, researchers have focused on various ensemble approaches to identify MC. Kang *et al* [19] combined two dictionary-based and five statistical-based systems for recognizing MC from clinical records. They have used 2010 i2b2/VA challenge dataset for their experiment. Their ensemble system produced achieved 0.82 *F*-measure score for identifying MC from test dataset [19]. Kim and Riloff [20] designed a stacked generalization-oriented ensemble system for extracting MC from clinical notes that also collected from 2010 i2b2 dataset [20]. In this research, we have developed two different versions of ensemble modules on our designed unsupervised and supervised modules. The first ensemble module looks for covering maximum lexical scope of initial modules of concept identification system. The second ensemble module is presented by stacking generalization where an ANN is used as a meta classifier [21]. The proposed ensemble module may assist in developing several domain-specific applications as future MC network and recommendation system, etc. in health-care [22–25].

The background work on concept identification system shows a clear idea on how a domain-specific lexicon and machine learning classifiers help to identify MC from the unstructured corpus. It also provides an overview related to the applied techniques for building an concept identification system in health-care. These observations motivated developing an automated system for identifying MC and non-MC in a context.

3. Dataset preparation

Initially, the dataset was collected from three different sources: SemEval 2015 Task-6, MedicineNet, and Ohsumed dataset. SemEval 2015 provides annotated MC with span, type, polarity, and relational information between a concept (event) and the document creation time. On the other hand, we have collected a set of diseases, symptoms, and drugs from MedicineNet website to enhance our experimental dataset. In addition, we have included Ohsumed dataset to represent the versatility of our experimental dataset. The terms of the Ohsumed dataset state that “Any human user of the data will explicitly be told that the data is incomplete and out-of-date.” Thereafter, We have collected 6774, 6432, and 2756 unique number of contexts from SemEval, MedicineNet, and Ohsumed resources, respectively. MC present in the collected data have been annotated by a group of medical practitioners. However, the data is split into three different parts: training (50%), development (30%), and test (rest of 20%) datasets to accomplish our experiments. In order to avoid the overfitting of the models and reducing the excessive training time of all the modules, we have not used the cross-validation approach. Distributions of the experimental dataset are shown in table 1.

It was observed that the patterns combining adjective (JJ) and noun (NN) appeared frequently in our experimental dataset. Also, we have noted that the POS tags such as the noun (NN), verb (VB), and adjective (JJ) play the crucial roles to identify MC from the experimental dataset. Hence, based on such important observations, a group of medical practitioners helped to design the following annotation policies. Finally, we conducted an agreement analysis for the annotated dataset.

Annotation policy: Collaborating with a group of medical practitioners, we have designed a set of manual annotation policies that helps to annotate the lexical scopes of MC and their non-medical counter parts. Annotation policies are considered to label MC at unigram <B-MC>, bi-gram <B-MC, I-MC>, tri-gram <B-MC, I-MC, I-MC>, and more than tri-gram <B-MC, I-MC, I-MC, ..> levels. We have observed that more than tri-gram MC occur very less in contrast with other MC. On the other hand, non-medical words or phrases are tagged as an N-MC in the context. In general, our medical experts observed two types of difficulties during annotation: (A) presence of non-medical word beginning, inside, or end in the lexical scope of MC (e.g., increased intracranial pressure) and (B) lexical boundary identification of overlapped MC (e.g. abnormal cell growth presented as MC whereas abnormal and cell growth refer to MC, individually). In order to assist this process and visualization, we have employed a tool named Brat.[6] Figure 1 shows a sample annotated output of the concept identification system.

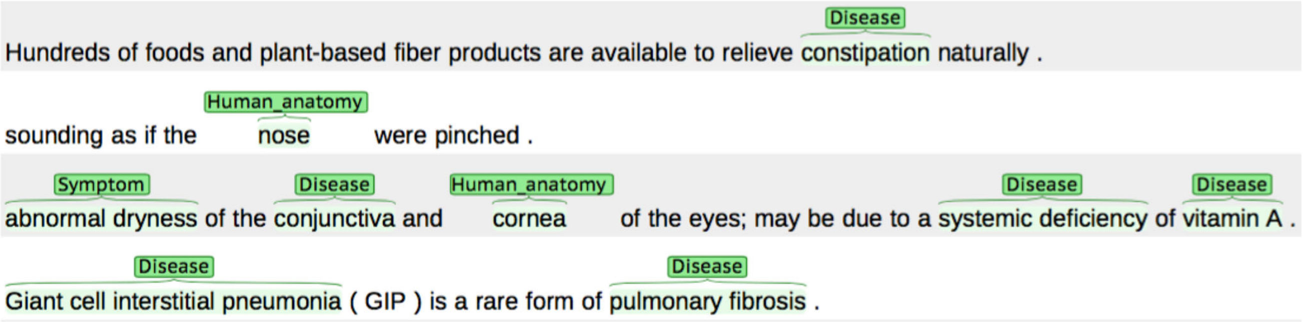
Agreement analysis We engaged another two domain experts to validate the annotated dataset. The agreement between two experts was calculated using Cohen’s kappa coefficient score (κ) [26]. The agreement analysis was conducted for MC as well as for non-medical words or phrases. In case of analysis, the annotators provided the agreement in terms of options, either “yes” (agreed) or “no” (disagreed). Table 2 shows the overall agreement scores that are provided by both of the annotators. The options produce 0.88 κ score, which implies almost perfect annotation for a manually annotated dataset. It was also noticed that the number of occurrences of the afore-mentioned difficulties (A, B) during the annotation was very small. Thus, such difficulties were resolved by mutual consensus among the annotators.

The following section describes how we have utilized the annotated dataset to build the proposed system. The system

[6] http://brat.nlplab.org.

Table 1. A detailed statistics of the experimental dataset.

Dataset statistics		Experimental	Training (50%)	Development (30%)	Test (20%)
No. of sentences		21786	10913	6105	4768
Unique no. of medical contexts$_{\text{sentences}}$		15962	7981	4789	3192
No. of words/phrases		56436	28211	16925	11300
Unique no. of medical concepts$_{\text{words/phrases}}$		24293	12143	7288	4862
Unique no. of *n*-gram medical concepts	$N = 1$ (uni-gram)	10391	5094	2915	2382
	$N = 2$ (bi-gram)	9732	4768	2892	2072
	$N = 3$ (tri-gram)	3223	1605	1312	306
	$N > 3$ (more than tri-gram)	947	676	169	102
Unique no. of major POS tags (all *n*-gram medical concepts)	Noun (NN, NNP, NNS ..)	14867	7385	4459	3023
	Verb (VB, VBP ..)	7891	3842	2294	1755
	Adjective (JJ, JJS ..)	1047	657	336	54
Unique no. of major POS tags (uni-gram medical concepts)	Noun (NN, NNP, NNS ..)	5919	3042	1693	1179
	e.g., consciousness, disorder, cancer, and malnutrition etc.				
	Verb (VB, VBP ..)	3679	1829	997	853
	e.g., transmitted, fissured, infected, and swelling etc.				
	Adjective (JJ, JJS ..)	358	165	102	91
	e.g., abnormal, structural, dry, and humorous etc.				
Unique no. of major POS patterns (bi-gram medical concepts)	JJ-NN	6345	3172	1903	1270
	e.g., painful inflammation, nutritional deficiencies, and fecal impaction etc.				
	NN-NN	2513	1256	744	513
	e.g., nail discoloration, memory loss, and neurocardiogenic syncope etc.				
	JJ-VB	834	416	228	190
	e.g., metabolic disorder, life-threatening condition, and acute spread etc.				
Unique no. of POS patterns (tri-gram medical concepts)	JJ-NN-NN	1513	765	623	125
	e.g., high blood pressure, anterior pituitary gland, and bacterial family Chlamydiaceae etc.				
	NN-NN-NN	941	468	364	109
	e.g., phosphoglycerate kinase deficiency, trade name Zocor, and family health history etc.				
	JJ-JJ-NN	738	367	309	62
	e.g., sympathetic nervous system, contagious viral disease, and abnormal cardiac rate etc.				



Hundreds of foods and plant-based fiber products are available to relieve constipation naturally .

sounding as if the nose were pinched .

abnormal dryness of the conjunctiva and cornea of the eyes; may be due to a systemic deficiency of vitamin A .

Giant cell interstitial pneumonia (GIP) is a rare form of pulmonary fibrosis .

Figure 1. A sample annotated output using Brat annotation tool.

helps to identify MC in three primary levels: uni-gram, bi-gram, and tri-gram, and a general level namely more than tri-gram. In order to identify the MC we have proposed three isolated modules, viz. unsupervised, supervised, and their combined ensemble module, which are discussed in the following subsections.

Table 2. An agreement analysis between two annotators to validate the dataset individually with respect to medical concepts (MC [24293]) as A# and non-medical concepts (N-MC [16548]) as B#.

	Annotator-1										
	Yes					No					*K Score*
Annotator-2											
A#											
Yes	20946	*N*=1	*N*=2	*N*=3	*N*>3	296	*N*=1	*N*=2	*N*=3	*N*>3	0.88
	N=1	8414	346	189	0	*N*=1	121	5	2	0	
	N=2	303	7907	191	0	*N*=2	4	111	2	0	
	N=3	0	201	2382	197	*N*=3	0	4	33	2	
	N>3	0	38	91	687	*N*>3	0	1	2	9	
No	362	*N*=1	*N*=2	*N*=3	*N*>3	2689	*N*=1	*N*=2	*N*=3	*N*>3	0.88
	N=1	151	6	2	0	*N*=1	1092	38	25	0	
	N=2	5	133	3	0	*N*=2	36	1014	23	0	
	N=3	0	4	41	3	*N*=3	0	45	292	19	
	N>3	0	1	3	10	*N*>3	0	9	14	82	
B#											
Yes	*15542*					76					0.90
No	91					839					

A#: MC (24293) and B#: N-MC (16548)

4. Concept identification framework

The concept extraction refers to identifying the textual span(s) of a single, multiple, or overlapped concept(s) within a context, whereas the task of concept recognition focuses on separately recognizing the MC from other non-medical entities. It has been found that our target MC appear in forms of words or phrases in our experimental dataset. Moreover, we identified them at three levels, namely uni-gram, bi-gram, and tri-gram. Thus, in order to identify such MC from their contexts, we have adopted a two-stage policy: first, concept extraction, and second, concept recognition. In order to achieve these two goals together, we employed a domain-specific lexicon into our unsupervised module, and concept- and context-related features into supervised module. Finally, we have combined both the modules and developed two versions of an ensemble module to enhance the performance of the proposed system. Figure 2 shows the overall flow-diagram of this framework.

4.1 *Unsupervised module*

Unsupervised module is designed by utilizing the enriched version of WME 3.0 [5] and phrase identification system using POS tagss [27, 28]. The WME 3.0 lexicon contains 10186 number of MC along with their POS, gloss, category, affinity score, gravity score, sentiment, polarity score, and semantics features.

Initially, we have assigned various POS tags like noun, verb, adverb, adjective, pronoun, preposition, etc. to all words using NLTK POS tagger.[7] Training dataset was used to learn the module whereas development dataset was used for enhancing the efficiency of this module. We have noticed that POS tags like nouns, verbs, and adjectives play the crucial roles to recognize both MC as well as non-medical words or phrases from the contexts. Thereafter, to identify *n*-gram MC, we have designed a few regular expressions based on POS tags and their patterns. Finally, we have employed WME 3.0 lexicon to identify MC and isolate them from non-medical words or phrases in the context. These identified MC are primarily labeled with four different tags: uni-gram (e.g., “cancer”), bi-gram (e.g., “adrenal_gland”), tri-gram (e.g., “central_nervous_system”), and more than tri-gram (e.g., “bounced_off_internal_tissues”). The lexical scopes of MC were identified either based on exact or partial match. The words or phrases in rest of the contexts are tagged as non-medical (N-MC) counter parts.

For example, unsupervised module is able to identify MC and non-medical (N-MC) words or phrases from the following context.

“Balance$^{\text{MC}}$ can$^{\text{N-MC}}$ do a body$^{\text{MC}}$ good$^{\text{N-MC}}$, beginning$^{\text{N-MC}}$ with the breath$^{\text{MC}}$.”

We have observed that due to the limitation of WME 3.0 lexicon and versatile nature of MC in terms of named entities of new domain, the unsupervised module often fails to capture MC containing lexical scopes more than bi-gram. Hence, we have engaged various conceptual and contextual features along with machine learning classifiers to build a supervised module in order to capture the concepts.

[7] https://www.nltk.org/book/ch05.html.

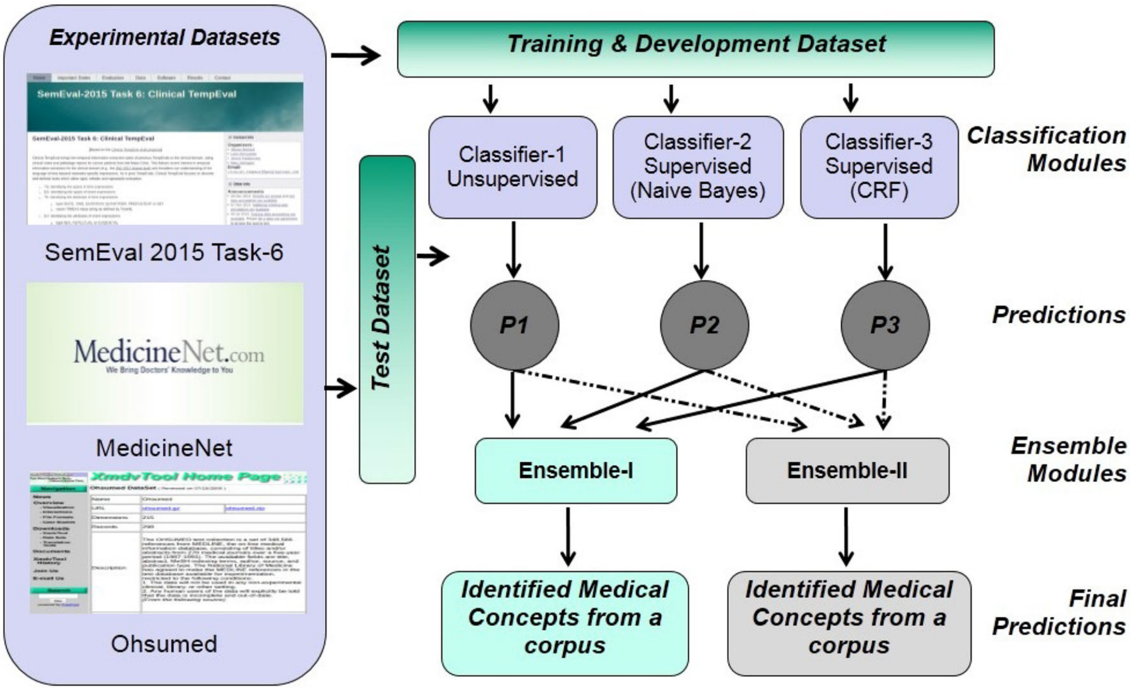

Experimental Datasets
SemEval-2015 Task 6: Clinical TempEval
SemEval 2015 Task-6
MedicineNet.com
We Bring Doctors' Knowledge to You
MedicineNet
XmdvTool Home Page
Ohsumed
Training & Development Dataset
Test Dataset
Classifier-1 Unsupervised
Classifier-2 Supervised (Naive Bayes)
Classifier-3 Supervised (CRF)
Classification Modules
P1
P2
P3
Predictions
Ensemble-I
Ensemble-II
Ensemble Modules
Identified Medical Concepts from a corpus
Identified Medical Concepts from a corpus
Final Predictions

Figure 2. A flow-diagram of the developed medical concept identification framework.

4.2 *Supervised module*

In order to build the supervised module, the concepts label features like category, POS, SSW, polarity score, and sentiment were used. On the other hand, context label features such as TF, capital letter, punctuation, and surrounding words help to realize the overall importance of the concept within a context. Specifically, category, POS, and SSW refer to extraction of lexical and syntactic knowledge of the concept whereas sentiment and polarity score assist in realizing the opinion of the concept. Similarly, affinity and gravity scores were used to capture the semantic information of the MC within its context. Alongside, we have attempted to use other Named Entity Recognition (NER) features such as bag of words (BOW). Unfortunately, we were unable to apply BOW feature in this module due to the versatile nature of MC. The definitions and examples of all the features are summarized in table 3.

These extracted features were applied on Naïve Bayes and CRF classifiers. In the process of learning, we have used three different tags: beginning <B-MC> and intermediate <I-MC> for MC and <N-MC> tag for non-medical words or phrases.

Naïve Bayes and CRF classifiers have been selected due to their probabilistic nature for labeling as well as dealing with segmented data. Naïve Bayes classifier helps to enhance the training dataset and handles non-linear effects of the dataset whereas CRF assists in handling conditional probability distribution over labeled sequences. In order to distinguish medical and non-medical words independently, we used Naïve Bayes as it supports feature independence. On the other hand, to capture the *n*-gram sequence of MC, we employed CRF as it deals with sequence labeling problem.

In order to implement both of the classifiers, we have used various standard python libraries such as pandas, numpy, scikit-learn, PyTorch, and matplotlib. Pandas, numpy, and matplotlib libraries help to summarize the dataset with output classes. On the other hand, scikit-learn and PyTprch libraries assist in building and validating Naïve Bayes and CRF models individually. Besides, we have used a dedicated CPU-based system such as i7-7500U to conduct the experiments. This system is able to train an average of 75 samples per second. In addition, we have noted the quadratic complexity to classify the MC from non-MC and identify the lexical scope for MC. This run-time complexity has been achieved while considering all features for all samples under supervised module.

For example, supervised module is able to identify MC and non-medical (N-MC) words or phrases from the following context using (a) Naïve Bayes and (b) CRF classifiers.

(a) [“In migraine^{MC}, pain^{MC} may occur$^{N\text{-}MC}$ on one side$^{N\text{-}MC}$ of the face^{MC}.”]^{NaiveBayes}
(b) [“In migraine^{MC}, pain^{MC} may occur$^{N\text{-}MC}$ on one side_of_the_face^{MC}.”]^{CRF}

We have noticed that CRF outperforms Naïve Bayes in case of identifying lexical scopes of MC up to tri-gram

Table 3. A detailed description and example of used features in supervised module. Example context: “Fever is usually defined as an oral temperature above 37.4 degrees.”s

Feature	Description	Example
Category	A class of words or concepts with similar characteristics	Fever: disease
POS	A syntactic representation of a word or concept	Fever: noun
SSW	Similar sentiment words of the target word	Pyrexia is SSW of fever
Sentiment	A view or opinion related to the target word in the form of positive, negative, and neutral	Neutral
Polarity score	A score that expresses the emotion of the word or sentence or corpus, which ranges from –1 to +1	Fever: – 0.56
Term frequency (TF)	How often a term occurs in the corpus	Fever: 1
Capital letter	Abbreviation and capitalization of word in a sentence	Fever
Punctuation	Target word closed punctuation marks in English grammar	Period
Surrounding words	Previous and next word of the target word in the sentence (context)	Oral and temperature

under supervised module. However, both the classifiers perform equally well in case of distinguishing medical and non-medical terms. In order to avail the facilities from individual modules together for enhancing the overall performance, we combined and developed an ensemble module.

4.3 *Ensemble module*

We have developed two versions of ensemble module to correctly recognize the lexical scopes of MC from non-medical counterparts within a context. Both of the ensemble modules have been designed by combining prior - mentioned unsupervised and supervised modules. The first ensemble module (Ensemble-I) has been designed by combining the important properties of unsupervised and supervised modules. This module aims to capture the concepts with maximum lexical scope from all possible *n*-grams. On the other hand, the second ensemble module (Ensemble-II) has been developed by introducing a stacking classifier [29]. The stacking classifier helps to learn the optimal mappings from the outputs of the individual modules and produces the final output. In the following part we have discussed Ensemble-I and Ensemble-II algorithms in detail, which assist in identifying MC ($MCI_{Ensemble}$) from the context.

4.4 *Ensemble-I*

i = no. of classifiers (1 to N) where N refers to the output of the unsupervised module and two classifiers of supervised module.

i' = $i + 1$

j = no. of classes (1 to L) where L refers to different classes (e.g., uni-gram, bi-gram, tri-gram, and more than tri-gram) of concepts.

j' = $j + 1$

```
if $C_i \neq C_{i'}$ then
    if $M_j == M_{j'}$ then
        $MCI_{Ensemble-I} = C_i M_j / C_{i'} M_{j'}$
    end
    else
        $MCI_{Ensemble-I} = C_{i'} M_{j'} when j' > j$;
        or
        $MCI_{Ensemble-I} = C_i M_j when j > j'$;
    end
end
```

Here, C_i and $C_{i'}$ indicate each class/classifier of the two modules individually. Additionally, M_j and $M_{j'}$ present each assigned class for the two concepts medical and non-medical.

Ensemble-I is able to identify MC and non-medical (N-MC) words or phrases from the following context using (a) unsupervised module, (b) Naïve Bayes and (c) CRF classifier, and (d) ensemble output. Figure 3 shows the flow-diagram of the model in detail. For example, Ensemble-I selects “pain” and “abdomen” as MC instead of “the_pain” or “the_abdomen” based on majority voting obtained from the unsupervised and CRF classifiers.

(a) [“The pain^{MC} is the result$^{N\text{-}MC}$ of an injury^{MC} to the abdomen^{MC}.”]^{unsupervised}

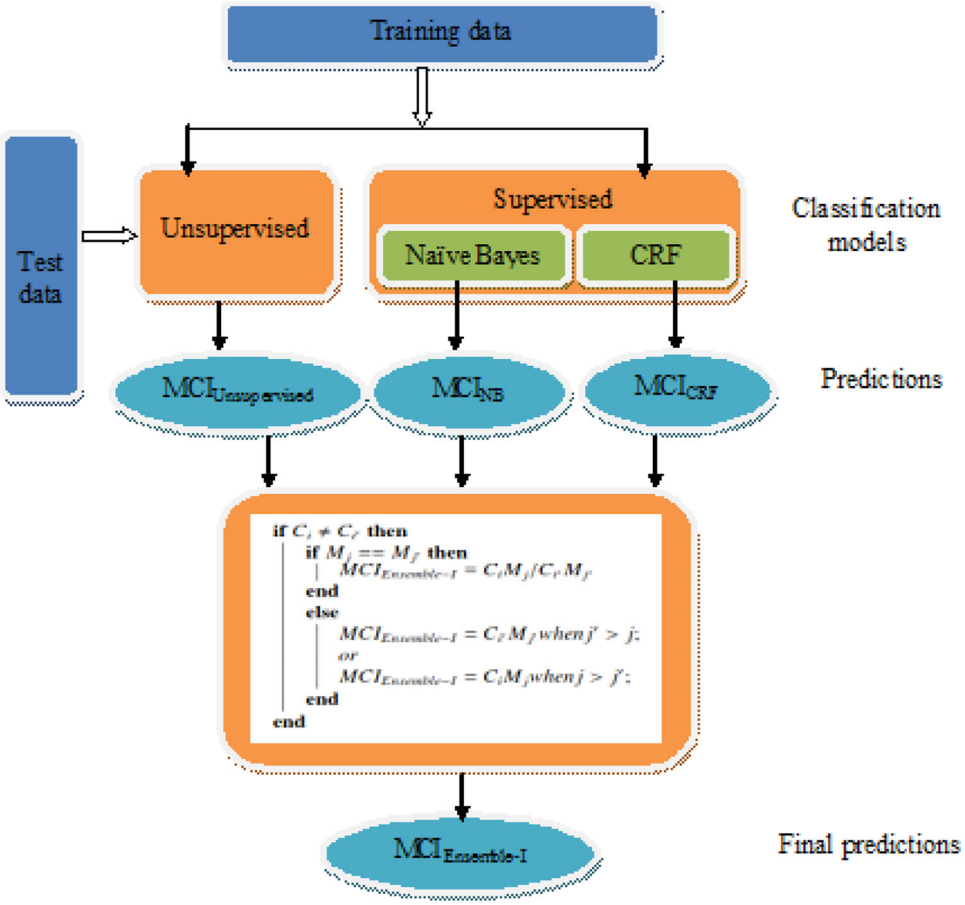
(b) [“The_pain^{MC} is the result$^{N\text{-}MC}$ of an injury^{MC} to the_abdomen^{MC}.”]^{NaiveBayes}

(c) [“The pain^{MC} is the result$^{N\text{-}MC}$ of an injury^{MC} to the abdomen^{MC}.”]^{CRF}

(d) [“The pain^{MC} is the result$^{N\text{-}MC}$ of an injury^{MC} to the abdomen^{MC}.”]$^{Ensemble-I}$

4.5 *Ensemble-II*

In order to develop Ensemble-II module, we have employed stacking-based ensemble learning technique. It



```
if $C_i \neq C_{i'}$ then
    if $M_j == M_{j'}$ then
        $MCI_{Ensemble-I} = C_i M_j / C_{i'} M_{j'}$
    end
    else
        $MCI_{Ensemble-I} = C_{i'} M_{j'}$ when $j' > j$;
        or
        $MCI_{Ensemble-I} = C_i M_j$ when $j > j'$;
    end
end
```

Figure 3. A flow-diagram of Ensemble-II to predict the medical concepts.

helps to combine multiple classifiers via a meta-classifier [8, 9]. In our case, we have used unsupervised module and supervised module consisting of Naïve Bayes and CRF as individual classifiers whereas ANN was applied as a meta-classifier. The meta-classifier was trained on the predicted class labels of the individual classifiers. The following algorithm is discussed in detail. Figure 4 shows the flow-diagram of the model in detail.

i = no. of classifiers (1 to *N*) where *N* refers to the output of the unsupervised module and two classifiers of supervised module.

j = no. of classes (1 to *L*) where *L* refers to different classes (e.g., uni-gram, bi-gram, tri-gram, and more than tri-gram) of concepts.

F = features for meta-classifier <f_i >

L = label for meta-classifier <l_j >

D = training data <f_i : l_j >

$MCI^{Ensemble-II}$ = predicted output of meta-classifier (ANN) using dataset D

For example, Ensemble-II is able to identify MC and non-medical (N-MC) words or phrases from the following context.

[“Cancer^{MC} is a group$^{N\text{-}MC}$ of diseases^{MC} involving$^{N\text{-}MC}$ abnormal_cell_growth^{MC} with the potential$^{N\text{-}MC}$ to invade$^{N\text{-}MC}$ or spread$^{N\text{-}MC}$ to other_parts$^{N\text{-}MC}$ of the body^{MC}.”]$^{Ensemble-II}$

5. Experimental results

The result analysis shows the important contributions of recognizing MC and non-medical words/phrases from a context. Results of all the modules are presented in terms of three different validation matrices: precision, recall, and *F*-measure. Precision measures the number of predicted positive class that belongs to the original positive class. On the other hand, recall measures the number of predicted positive class made out of all positive counts in the dataset. Finally, *F*-measure provides a single value to balance both the concerns of precision and recall. In our present framework, MC are validated with respect to four different lexical scopes uni-gram, bi-gram, tri-gram, and more than tri-gram whereas non-medical words or phrases are evaluated only in terms of their presence and absence

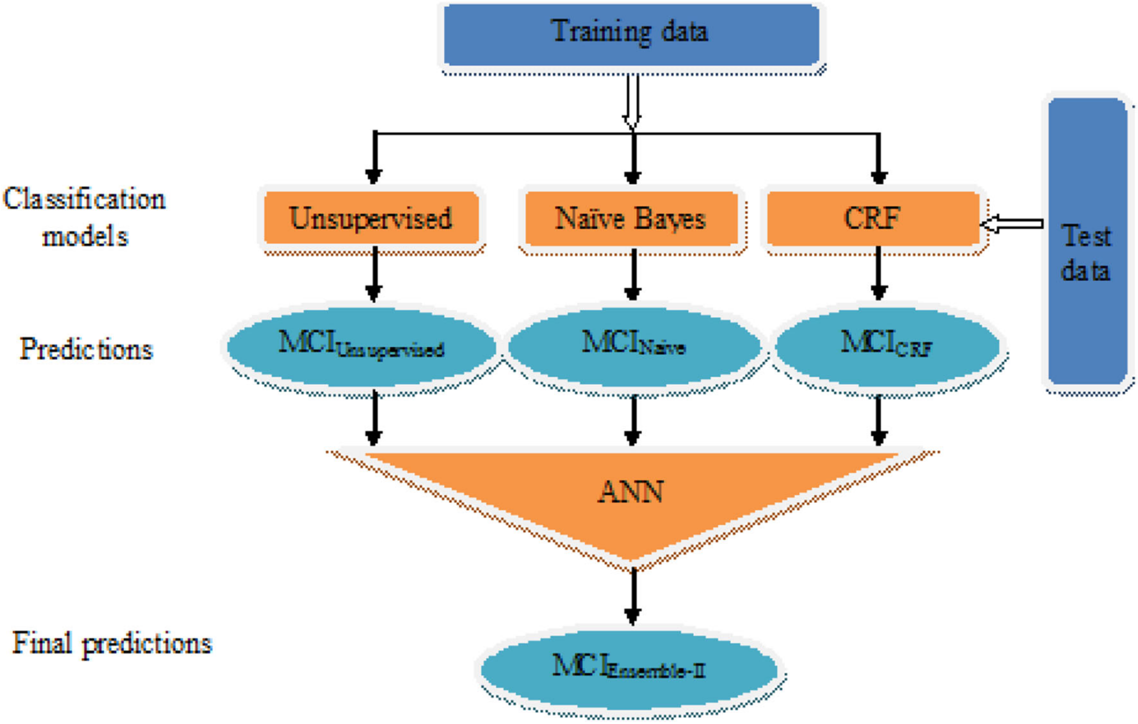


Figure 4. A flow-diagram of Ensemble-II to predict the medical concepts.

Table 4. Ablation study (as *F*-measure score) for identifying medical concepts using Naïve Bayes and CRF classifiers.

	Naïve Bayes				CRF			
Features	$N = 1$	$N = 2$	$N = 3$	$N > 3$	$N = 1$	$N = 2$	$N = 3$	$N > 3$
Category	0.87	0.88	0.80	0.83	0.92	0.90	0.82	0.81
Category + TF	0.89	0.88	0.81	0.83	0.92	0.91	0.82	0.82
POS + SSW	0.88	0.87	0.83	0.83	0.93	0.90	0.84	0.83
Category + POS + SSW + polarity score + sentiment	0.94	0.93	0.83	0.84	0.94	0.92	0.85	0.85
Category + TF + POS + SSW	0.95	0.93	0.84	0.85	0.96	0.94	0.85	0.86
Category + TF + POS + polarity score + sentiment	0.97	0.95	0.85	0.85	0.97	0.95	0.87	0.88

in the context. The complete evaluation process has been conducted in three stages: ablation study, error analysis, and comparative study, which are discussed in the following subsections.

5.1 *Ablation study*

Ablation study is essential for selecting important features in order to build a good machine learning model. Therefore, we have used this approach to recognize important features from all of our conceptual and contextual features under supervised module. It also helps to compare the performances of prior-mentioned two classifiers with respect to different feature combinations. Table 4 shows the results of the ablation study conducted for Naïve Bayes and CRF classifiers.

It was observed that our ablation study helps to identify the common features such as category, TF, POS, polarity score, and sentiment, which are important to improve the performances of the two classifiers. It has been also observed that the CRF outperforms in capturing the MC of lexical scopes more than tri-gram while lexical and syntactic features were added. On the other hand, to validate the overall output of all three modules, we have conducted an error analysis in the form of confusion matrix discussed in the following subsection.

Table 5. Confusion matrix for identifying various classes of medical concepts and confusion between non-medical (Yes) and medical (No) concepts using all modules.

	Predicted												
	Unsupervised				Supervised (Naïve Bayes/CRF)				Ensemble (Ensemble-I/Ensemble-II)				
#	$N = 1$	$N = 2$	$N = 3$	$N > 3$	$N = 1$	$N = 2$	$N = 3$	$N > 3$	$N = 1$	$N = 2$	$N = 3$	$N > 3$	Total
Original													
$N = 1$	*2302*	73	7	0	*2309/2311*	67/65	6/6	0/0	*2314/2309*	63/67	5/6	0/0	2382
$N = 2$	58	*1995*	19	0	55/53	*1998/2001*	19/18	0/0	50/52	*2004/2001*	18/19	0/0	2072
$N = 3$	0	53	*219*	34	0/0	36/27	*252/264*	18/15	0/0	22/24	*273/270*	11/12	306
$N > 3$	0	5	8	*89*	0/0	5/4	7/6	*90/92*	0/0	3/4	4/6	*95/92*	102
Total	2360	2126	253	123	2364/2364	2106/2097	284/294	108/107	2364/2361	2092/2096	300/301	106/104	

	Unsupervised		Supervised (Naïve Bayes/CRF)		Ensemble (Ensemble-I/Ensemble-II)		
$	Yes	No	Yes	No	Yes	No	Total
Original							
Yes	*3241*	72	*3248/3251*	65/62	*3251/3252*	62/61	3313
No	211	*4651*	196/195	*4666/4667*	195/192	*4667/4670*	4862
Total	3452	4723	3444/3446	4731/4729	3446/3444	4729/4731	

Italic values represent the correctly identified concepts under each lexical scope. Lexical scopes are respresented using '*N*' in the table

# Number of medical concepts: 4862 and $ number of concepts: 8175

Table 6. A comparative result analysis for identifying medical concepts and non-medical words or phrases from the test dataset.

	Unsupervised			Supervised (Naïve Bayes/CRF)			Ensemble (Ensemble-I/Ensemble-II)		
	Precision	Recall	*F*-measure	Precision	Recall	*F*-measure	Precision	Recall	*F*-measure
$N = 1$	0.97	0.98	0.97	0.97/0.97	0.98/0.98	0.97/0.97	0.97/0.97	0.98/0.98	0.97/0.97
$N = 2$	0.96	0.94	0.95	0.96/0.96	0.95/0.95	0.95/0.95	0.97/0.97	0.96/0.95	0.96/0.96
$N = 3$	0.71	0.87	0.78	0.82/0.86	0.89/0.89	0.85/0.87	0.89/0.88	0.91/0.90	0.90/0.89
$N > 3$	0.87	0.72	0.79	0.88/0.90	0.83/0.86	0.85/0.88	0.93/0.90	0.90/0.88	0.91/0.89
MC	0.96	0.98	0.97	0.96/0.96	0.98/0.99	0.97/0.97	0.96/0.96	0.99/0.98	0.97/0.97
N-MC	0.98	0.94	0.96	0.98/0.98	0.95/0.95	0.96/0.96	0.99/0.99	0.95/0.95	0.97/0.97

5.2 *Error analysis*

We have conducted an error analysis in the form of confusion matrix with respect to all three modules. The confusions of the identified MC are represented by four different tags, namely uni-gram, bi-gram, tri-gram, and more than tri-gram. On the other hand, the confusions of non-medical words or phrases are represented by yes (non-medical) and no (medical). Table 5 shows the confusion matrices for identifying various tags of MC and distributions of medical and non-MC in test dataset.

In case of unsupervised module, we find that the upper triangular part of the confusion matrix is responsible for insertion-based confusions whereas the lower part indicates the confusions that occur due to deletion of both types of concepts. This confusion helps to understand the primary challenge to get correct concepts, which contains lexical scopes more than bi-grams. It also assists in enhancing the performance of next two modules: supervised and ensemble. Thereafter we validate all three modules using precision, recall, and *F*-measure matrices, which are discussed in detail in the following subsection.

5.3 *Comparative study*

We have evaluated the identified medical and non-MC using *F*-measure score in the presence of the test dataset. Test dataset has been applied on unsupervised, supervised, and ensemble modules to get their *F*-measure scores, which proves the efficiency of ensemble module over rest of the modules. Table 6 shows the comparative analysis of the performance to identify various tags of MC and overall identified medical and non-medical terms. MC are tagged in fine-grained *n*-gram labels whereas non-MC are tagged in a general way in the form of N-MC.

In the following subsection, we discuss our observations to develop a concept identification system in health-care.

5.4 *Observations*

We have observed that the unsupervised module provides better accuracy to identify the uni-gram and bi-gram MC over the rest due to incorporation of lexicon. Two types of confusions were found: one is insertion based where a word is appended either in front or rear of a medical phrase (e.g. identified "treatment option" instead of "treatment") and another is deletion where a word is deleted either from the front or rear of a medical phrase (e.g. identified "retinoic acid" concept instead of "13-cis retinoic acid"). On the other hand, the non-MC are primarily extracted as a noun and verb phrases rather than any other POS tag.

In case of supervised module, we have noticed that it provides an adequate output for identifying MC up to tri-gram. Here, more number of non-MC are identified over unsupervised module which are not only presented by noun and verb phrases. Also, we have observed that the domain-specific features such as SSW, sentiment, polarity score, and category of MC help to enhance the accuracy of the supervised module. These features assist in identifying more unique number of unknown MC over WME 3.0 lexicon-driven unsupervised module.

On the other hand, being a meta-classifier, Ensemble-II comprises the errors of individual classifiers. As a result, the performance of this module is weak in contrast with Ensemble-I.

It is observed that the performances are degraded at the individual level as well as ensemble level with respect to all classifiers while the length of n is increased in case of n-gram MC. We have noted that the presence of MC with $n > 3$ is very less compared with other related MC in this work. Hence, in some cases we find the accuracy of the identified MC with $n > 3$ is higher than that with $n = 3$. Also, we noticed that the seen concepts (present in training dataset) dominated over unseen concepts (absent in training dataset) in test dataset. In this work, we have not focused on these unseen MC due to the limitation of domain-specific features.

In addition, to compare our approach to an existing method on a different dataset, we were unable to get a state-of-the-art dataset that contains five different categories like disease, symptom, drug, human anatomy, and miscellaneous medical terms (MMT) for MC all together in a single corpus. The available corpus tagged with either drugs or diseases/conditions or treatments category for MC, individually. The National Center for Biotechnology Information (NCBI) corpus is a collection of 793 PubMed abstracts that annotates concept levels diseases. On the other hand Ohsumed corpus includes medical abstracts from the MeSH categories, primarily cardiovascular diseases categories. Also, we are unable to find any benchmark dataset that contains human anatomy category for MC. Hence, we are motivated to prepare an experimental dataset that consists of all mentioned five categories for MC for a corpus in this research.

Finally, it has been noticed that both of the ensemble modules gain performance while utilizing the advantage of individual modules. Ensemble modules provide an average 0.91 F-measure score for identifying all tags of MC and 0.97 F-measure score for recognizing non-medical words or phrases from the test dataset. Ensemble modules may assist in designing various applications such as annotation system and recommendation system in health-care in future [30].

6. Conclusions and future scopes

The present task is reporting the development process of an ensemble module to identify MC as well as non-MC from an unstructured corpus in health-care. The primary motivation behind this research is to support the expert (e.g. doctors) as well as non-expert (e.g. patients) groups of people to enhance their understanding of key information from a text corpus. Hence, we have proposed three different modules, namely unsupervised, supervised, and their combined two versions of ensemble modules. The unsupervised module is presented by POS tagging followed by searching in a domain-specific lexicon, namely WME 3.0. In addition, the supervised module is designed by Naïve Bayes and CRF classifiers along with various features like category, POS, sentiment, etc. Thereafter, we have combined the important outcomes of unsupervised and supervised modules and developed two versions of ensemble module, namely Ensemble-I and Ensemble-II. Besides, a group of manual annotators help to design an experimental dataset in this research.

Future effort will be to enhance the accuracy of the proposed system, which may assist in designing various applications such as annotation system, disease and drug recommendation system, and symptom checker in health-care.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Informed consent Informed consent was not required as no human or animals were involved.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

References

[1] Cambria E, Hussain A and Eckl C 2011 Bridging the gap between structured and unstructured healthcare data through semantics and sentics. In: *Proceedings of ACM WebSci'11*, pp. 1–4

[2] Mondal A, Cambria E, Das D, Hussain A and Bandyopadhyay S 2018 Relation extraction of medical concepts using

categorization and sentiment analysis. *Cognitive Computation* 10(1): https://doi.org/10.1007/s12559-018-9567-8

[3] Mondal A, Das D and Bandyopadhyay S 2017 Relationship extraction based on category of medical concepts from lexical contexts. In: *Proceedings of the 14th International Conference on Natural Language Processing (ICON)*, pp. 212–219

[4] Ma Y, Cambria E and Gao S 2016 Label embedding for zero-shot fine-grained named entity typing. In: *Proceedings of COLING*, Osaka, pp. 171–180

[5] Mondal A, Das D, Cambria E and Bandyopadhyay S 2018 WME 3.0: an enhanced and validated lexicon of medical concepts. In: *Proceedings of the Ninth Global WordNet Conference*

[6] Mondal A, Cambria E, Das D and Bandyopadhyay S 2017 Employing sentiment-based affinity and gravity scores to identify relations of medical concepts. In: *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, pp. 1–7

[7] Chaturvedi I, Ragusa E, Gastaldo P, Zunino R and Cambria E 2017 Bayesian network based extreme learning machine for subjectivity detection. *Journal of The Franklin Institute* 355(4): 1780–1797

[8] Aggarwal C C 2014 *Data classification: algorithms and applications*. CRC Press,

[9] Wolpert D H 1992 Stacked generalization. *Neural Networks* 5(2): 241–259

[10] Garten Y, Coulet A and Altman R B 2010 Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics* 11(10): 1467–1489

[11] Kim Y, Riloff E and Hurdle J F 2015 A study of concept extraction across different types of clinical notes. In: *Proceedings of the AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2015, p. 737

[12] Mondal A, Satapathy R, Das D and Bandyopadhyay S 2016 A hybrid approach based sentiment extraction from medical context. In: *Proceedings of SAAIP, IJCAI*, vol. 1619, pp. 35–40

[13] Pakhomov S V, Coden A and Chute C G 2006 Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics* 75(6): 418–429

[14] Savova G K, Chapman W W, Zheng J and Crowley R S 2011 Anaphoric relations in the clinical narrative: corpus creation. *Journal of the American Medical Informatics Association* 18(4): 459–465

[15] Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, Kola J S, Roberts I, Setzer A, Tapuria A, *et al* 2007 The clef corpus: semantic annotation of clinical text. In: *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2007, p. 625

[16] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I and Setzer A 2009 Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics* 42(5): 950–966

[17] Tsuruoka Y and Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 467–474. Association for Computational Linguistics, 2005.

[18] Mandel M A 2006 Integrated annotation of biomedical text: creating the pennbioie corpus. In: *Proceedings of Text Mining Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK

[19] Kang N, Afzal Z, Singh B, Van Mulligen E M and Kors J A 2012 Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics* 45(3): 423–428

[20] Kim Y and Riloff E 2015 Stacked generalization for medical concept extraction from clinical notes. In: *Proceedings of BioNLP 15*, pp. 61–70

[21] Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A and Gonzalez-Hernandez G 2019 Deep neural networks ensemble for detecting medication mentions in tweets. arXiv:1904.05308

[22] Mondal A, Cambria E, Das D and Bandyopadhyay S 2017 Mediconceptnet: an affinity score based medical concept network. In: *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS*, pp. 22–24

[23] Dey M, Mondal A and Das D 2016 Ntcir-12 mobileclick: sense-based ranking and summarization of english queries. In: *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*, pp. 138–142

[24] Cambria E, Hussain A, Durrani T, Havasi C, Eckl C and Munro J 2010 Sentic computing for patient centered applications. In: *Proceedings of the 10th IEEE International Conference on Signal Processing*, IEEE, pp. 1279–1282

[25] Smith B and Fellbaum C 2004 Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In: *Proceedings of the 20th International Conference on Computational Linguistics*, Association for Computational Linguistics, p. 371

[26] Viera A J, Garrett J M, *et al* 2005 Understanding interobserver agreement: the kappa statistic. *Family Medicine* 37(5): 360–363

[27] Mondal A, Chaturvedi I, Das D, Bajpai R and Bandyopadhyay S 2015 Lexical resource for medical events: a polarity based approach. In: *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, IEEE, pp. 1302–1309

[28] Mondal A, Das D, Cambria E and Bandyopadhyay S 2016 WMW: sense, polarity and affinity based concept resource for medical events. In: *Proceedings of the Eighth Global WordNet Conference*, pp. 242–246

[29] Lee E S 2017 Exploring the performance of stacking classifier to predict depression among the elderly. In: *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, pp. 13–20

[30] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L, Winters S and White P 2004 Integrated annotation for biomedical information extraction. In: *Proceedings of the HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, pp. 61–68