Framework for enhancing the Emotions of EMR using Ontology in Sentiment Analysis

Manikandan K [1*], G.Victo Sudha George [2], D.Usha [3], V.R.Niveditha [4], V. Vinoth Kumar [5]

[1] Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu India.

[2] Professor, Department of Computer Science and Engineering,

[3] Associate Professor, Department of Computer Science and Engineering,

[4] Research scholar, Department of Computer Science and Engineering,

[2, 3, 4] Dr.M.G.R. Educational and Research Institute, Chennai, India.

[5] Associate Professor, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore.

Abstract

By analyzing a data using NPL, electronic health records (EMRs) could offer many insights, which have not been exploited yet .Through Sentiment Analysis we were able to present the issue in a way that the physicians could identify with and solve. If the stored records are analyzed using a Natural Language Processing methodology (NLP) it will be very helpful in automating the process of collecting ,analyzing the data. This work aim to predict the International Classification of Diseases, Revision 10(ICD-10) code(s) – or it's (their) derivatives – from the raw text records. Through we can easily diagnose the disease based upon the patient's foretelling symptoms instead of going each and every long written data through nurses or its previous medical history. In this paper we represent the pipeline approach on information extraction, sentiment analysis, creating ontology for unsupervised learning and summarization technologies. Sentiment Analysis is performed through recursive neural deep learning and lexicon analysis. In this paper Ontology has the major concern to provide better prediction of related diseases and helps in more proficient summarization. The feasibility of the approach is evaluated through linguistic analysis and user studies. In the presented work we also summarize the effectiveness of the automated EMR against the traditional EMR.

Keywords: *Natural Processing Language (NPL), lexicon analysis, Electronic health records (EMRs), (ICD-10) code(s), Sentiment Analysis, Word clouds, Ontology, RDF graph, Neural Learning.*

1. Introduction

EMR is the significant division in Healthcare advancing now a days. EMR software is widely used for assisting and managing the procedures in the hospital. EMR programming is generally utilized for helping and dealing with the strategies in the doctor's facility. Initially the patient's record is managed on a standalone application due to the reason not able to access worldwide. The major aim of EMR is available the medical history of an individual anywhere anytime without carrying its previous medical paper records. As shifting to new technology it will improve in accessing the patient's medical history easily and which helps in diagnosis the present scenario. If we analyzed using Natural Processing Language (NPL), electronic health records (EHRs/EMRs) could offer numerous bits of knowledge, which have not been exploited yet. NPL can help by scanning the huge amounts of documents and literature who has hidden correlations. It aid in clustering and categorizing these documents or help to build the fundaments of systems by assisting the doctors with diagnosis and suggesting treatment plans. A few authors have grouped patients in light of their clinical notes and highlighted connections between these formed clusters and gene changes in those patients' tumors [4]. This work means to foresee the International Classification of Diseases, Revision 10(ICD-10) code(s) – or it's (their) derivatives – from the raw text records. In this fast pace world we need everything result fast and precise. All information is maintained over the huge database in the cloud. Through Sentiment Analysis we can improve the productivity through both ends patients as well as Physician ends. Using the Sentiment Analysis with Ontology will act as a Cherry on a cake. With the help of Ontology we not only get the better prediction of the diseases but it also helps us in performing unsupervised learning. The ability to reliably predict outcomes to improve quality of care. A physician without undergo the previous medical history can determine and predict the possibility of the present scenarios of the patient. This way of summarizing the data can be also considered as prediction analysis.

Sentiment Analysis: - Sentiment Analysis (SA) is widely known as opinion mining that aims to determine the attitude of the writer or a speaker related to the topic or overall contextual polarity for a document [12]. This task has created a considerable interest due to its wide applications. However, in the classic Sentiment Analysis irrespective of the domain the polarity of each term of the document is computed independently. SA is one of the hottest studied area currently in Natural Language Processing (NLP), and recently plunge into the Semantic Web world which anticipate evidence that including semantic features to SA algorithms upgrades their performance 1. In any case, existing methodologies at SA, even those that fuse semantic components, are for the most part supervised and depend on the accessibility of experiments and tests, henceforth we can consider them as domain-dependent. This paper has an objective of profoundly hybridizing NLP and technologies by a means of ontology and constructs an unsupervised approach for processing SA of sentences in EMR.

The notions of sentiment in medical history: - Textual information can be broadly categorized into facts and opinions [13]. Facts are objective expressions about entities or events. Opinions are usually subjective expressions describing depicting individuals' states of mind, slants or emotions about substances.

Sentiment can be seen as a reflection of the health status of a patient which can be good, bad or normal at some point in time, expressed either implicitly or explicitly. Implicit descriptions of the health status concern the mentioning of symptoms (e.g. severe pain, extreme weight loss, high blood pressure). They require additional content information for correct interpretation. In the table 1 we have provided few examples how sentiment can be recognized from a words in medical history of patient. An explicit description of the health status is reflected in phrases such as the patient recovered well or normal.

Table 1. Words and expressions to recognize attitude

S No.	Entity Possible	sentiment values
1	Health status	Improve, worsen
2	Medical condition	Present ,improve, worsen
3	Diagnosis	Certain , uncertain, preliminary
4	Effect of a medical event	Critical, non-critical
5	Medical procedure	Positive or negative outcome, successful or unsuccessful
6	Medication	useless, serious adverse events

Patients Outlook: - As sentiment analysis is already been very popular to extract the information from the reviews, blogs and social media for rating a products in the market. Now a days companies are also using this technology to mark themselves and retrieve the position and area of development in the existing market. Similarly, through this a person who is urgently looking for a doctor to get an appointment, but he didn't know which physician is suited best for him .Instead of reading the reviews from different sources he can opt this option. Ontology can result in better results. Now a day's search engine are using RDF and Ontology to provide the best search result for the user.

Physician Outlook: - When Patients are monitored over a long period of time, at each appointment is very crucial to have an overview on the progress and changes in the patient status promptly. Such information is documented in the clinical narratives. With the proceeding of the medical treatment, the volume of these narratives for each patient increases rapidly. The large amount of patient data can easily overwhelm the processing capability of the physicians. The overflow of patient records may therefore lead to the following practical problems: First, it is increasingly difficult for the physicians to get a rapid overview of the patient status. Second, the physicians can only search for the patient records through keyword or Boolean query. The inquiry of semantic aspects such as symptoms, opinions, intentions, and judgment is impossible. Third, the summarization of patient status and treatment procedures is highly labor-intensive and time consuming. In particular, the writing of the discharge summary of differential diagnoses requires still the perusing of all the pervious judgments and diagnoses [1]. This process can be more fruitfully performed if we are able to link all its previous record and present case and also able to predict and summarize the future possibilities for the patient. This process may turn out to be fully automated and helps in generalization of the medial history.

"Finding an exact and speedy approach to figure out which patients are at high danger of building up the sickness is basically critical", said examine Medical Director of EHRs at UC Davis and co-creator Hien Nguyen, Associate Professor of Internal Medicine [17].

In order to handle the "big data" problem for medical records and offers the physician a swift access to the patient status, this paper defined a novel processing pipeline based on information extraction, sentiment analysis, Onto Graph and summarization technologies.

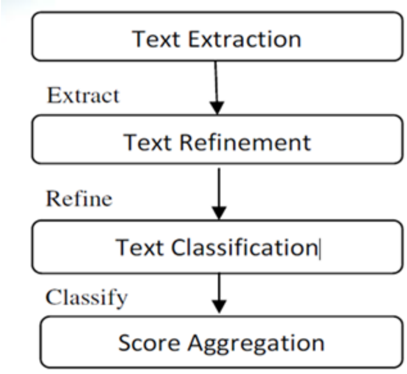


Fig 1 sentiment analysis

Basic methodology: - The basic methodologies used to extract the sentiment from the text are shown in table 2. These are the common well known approaches used. All the models are boost while training the dataset but the main drawback of the models are they persist the line order of the word. So in this paper we are using recursive deep neural tree model to train the dataset.

Table 2: Basic Methodologies used in Sentiment Analysis

S No.	SA Methodologies	Algorithm/Technique
1	Naïve Bayes Classifier	Words likelihood in a sentiment class
2	Lexicon-Based Methods	Score sentiment keywords
3	Machine Learning Algorithms	Support Vector Machine, Maximum Entropy, Neural Network
4	Concept-level techniques	Ontologies, Entity recognition, Semantic vector space

Ontology:-

The world is depicted in terms of events and circumstances, and objects are almost always involved in a special occurrence of one or the other. In light of this perspective of the world, sentences are spoken to as connected events or situations, with participating objects [1].

In this paper we are not intended to define or codify a new controlled medical vocabulary, but try to complement the existing RDF and vocabularies. Here we are more focusing on the existing relationships between the entities rather than their convection to use. In this paper we doesn't give an approach to clarify elements with codes that alludes to existing restorative vocabularies, (for example RxNorm, SNOMED, ICD, UMLS, MeSH and so on.).

RDF:-

Triples, or statements, can be considered foundational units of the Semantic Web. Triple gets its name from the number of components it contains. Each triple states a fact and consists of subject, predicate and object [18].

A set of such triples is called a Resource Description Framework [16] (RDF) graph.

RDF triplets help in finding the relation between the entities and by using OWL we can predict the unknown possibility and relationship between the objects. In this paper SPARQL queries are used for retrieving the known information. Here we used protégé as a tools to implement RDF graph and SPAQL queries.

Figure 2 visualizes piece of data from Human-Disease Ontology (DO). Blue color rectangles represent the classes and light red rounded rectangles show string values that carry type of information specified by the labels on arrows. In total there are 4 triples represented on the picture. They all have the same subject which is the URI http://purl.obolibrary.org/obo/DOID_10652. Four different predicates relate this URI to a label for the class represented this URI, subclass information, synonym for the label of class and cross-reference to another ontology. Two objects, "Alzheimer’s disease" and "Dementia of the Alzheimer’s type" carry string value and thus are literals

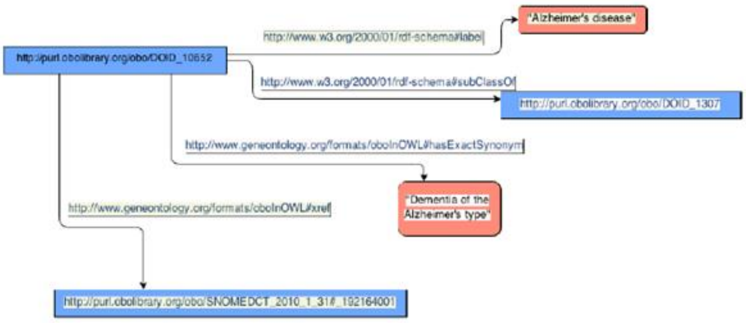

http://purl.obolibrary.org/obo/DOID_10652
http://www.w3.org/2000/01/rdf-schema#label
"Alzheimer's disease"
http://www.w3.org/2000/01/rdf-schema#subClassOf
http://purl.obolibrary.org/obo/DOID_1307
http://www.geneontology.org/formats/oboInOWL#hasExactSynonym
"Dementia of the Alzheimer's type"
http://www.geneontology.org/formats/oboInOWL#xref
http://purl.obolibrary.org/obo/SNOMEDCT_2010_1_31#_192164001

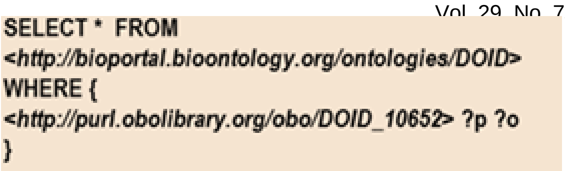
Fig 2. Example of RDF from DO

The other two objects are resources.

SPARQL:

SPARQL for the RDF plays the similar role of SQL for the relational databases. It is a query language designed for querying RDF databases [16]. SPARQL queries can include one or more triples where the subject, predicate and/or object can be variables. These queries are being sent to the SPARQL endpoints [22]. On the endpoint the triples in the query are being compared to the stored ones in specified RDF graphs. Here we had shown an example of a SPARQL query.

This query results in all the triples in graph http://bioportal.bioontology.org/ontologies/DOID that contain as a subject the URI http://purl.obolibrary.org/obo/DOID_10652. Four of these triples are shown in Figure 3.



```
SELECT *  FROM
<http://bioportal.bioontology.org/ontologies/DOID>
WHERE {
<http://purl.obolibrary.org/obo/DOID_10652> ?p ?o
}
```

Fig 3. SPARQL Query Example

A vocabulary for events and situations, and Bio Portal as reference for thematic roles of events. We compute its sentiment score by combining the scores of its associated opinion features, which are extracted from the RDF graph representing the opinionated sentence.

2. Problem Definition

Given a review r, using a sentiment analysis methodology to classify its sentiment polarity either positive or negative. We use the result 1 stands for positive and -1 for negative as shown in figure 4.

$$F_{polarity}(r) = \begin{cases} 1 & \text{when r's label is score 5 or score 4 (positive)} \\ -1 & \text{when r's label is score 1 and score 2 (negative)} \end{cases}$$

Fig 4. Dual Sentiment Analysis

Given a review r, using a sentiment analysis methodology to classify its sentiment polarity extend in 5 levels, a scale of 1 to 5.

$$F_{polarity}(r) = \begin{cases} 1 & \text{when r's label is score 1 (strong negative)} \\ 2 & \text{when r's label is score 2 (weak negative)} \\ 3 & \text{when r's label is score 3 (objective)} \\ 4 & \text{when r's label is score 4 (weak positive)} \\ 5 & \text{when r's label is score 5 (strong positive)} \end{cases}$$

Fig 5. Sentiment Analysis up to 5 levels

3. Literature survey

Neural systems have been broadly received for different PC vision assignments, including digit acknowledgment question discovery and facial acknowledgment. As of late they have likewise turned out to be prominent for an assortment of NLP issues, beginning with learning of word portrayals and being reached out to data recovery producing models machine interpretation discourse acknowledgment and sentiment analysis.

Sentiment analysis can be done at three levels namely Document level, Sentence level, Entity and aspect level. Level of document expresses the positive or negative opinion of a single entity in the document as global.

In sentence level each sentence in the document is analyzed to determine the positive, negative or neutral opinion [3].

Andrea Esuli and Fabrizio Sebastiani [15] proposed semi-supervised technique that helps in the classification that determines the orientation of subjective term. Their basic idea of this methodology is to do quantitative analysis of the glosses of these terms. But the proposed review does not have enough contextual information to determine the actual sentiment, Chunxu Wu [16] proposed a method in which contextual information present in other reviews about the same topic is gathered and analyzed, then by using semantic similarity among them, one can judge the orientation of that sentiment. An exception is the 6-layer CNN proposed by Zhang et al. 2015 for character-level text classification. More sophisticated alternatives are recursive neural networks, exhibited among others by Dong et al. 2014 although the best-performing approaches of the SemEval Twitter sentiment analysis contest of recent years mostly featured simpler architectures Limitations of these kind of setups include their requirement of large amounts of training data and that many algorithms base their prediction on short (contiguous) phrases, but struggle when information between far-apart word groups has to be combined. State of the art algorithms mostly rely on neural networks (predictive model) or factorization of the co-occurrence matrix (count-based model) in order to generate these dense vectors. Songho [7] also used K-Nearest Neighbor (KNN) approach which finds 'K' nearest neighbors of a text document among the documents in the training data set. Then classification is performed on the basis of similarity score of a class with respect to a neighbor.

Collobert and Weston, 2008; Huang et al., 2012 believe that the vectors can be trained in an unsupervised fashion to capture distributional similarities but then also be fine-tuned and trained to specific tasks such as sentiment detection. Grefenstette and Sadrzadeh (2011) analyze subject-verb object triplets and find a matrix-based categorical model to correlate well with human judgments. Y. Deng, 2014 et al. [20] proposed the summarization techniques used in medical domain. The paper to provide the basic and effective methodology for summarization.

4. Proposed Methodology

The technique of estimation sentiment analysis has following steps:

- First step: First concentrate the information to be investigated from the web. Tokenization of the medical terms are performed in this progression.
- Second step: For preprocessing and polarity calculation of the data we have created a training dataset for positive, negative, average sentiment words and stop words.
- Third step: Preprocessing- In the pre-processing of the data the words which are not carrying any sentiments or opinion are removed from the data. Another task performed in pre-processing is stemming. It is the process of reducing derived words into their root forms e.g. word happiness is reduced into root form happy. So, after preprocessing we get only the meaningful data on which we can easily apply the techniques.
- Fourth Step: "Calculate Polarity" provides us the count of positive, negative and average sentiment words in the entered data which is used by the techniques as an RNN for tree model for further processing.

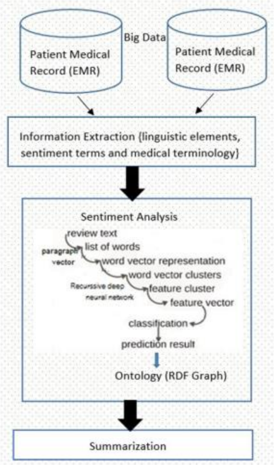

Patient Medical Record (EMR)
Big Data
Patient Medical Record (EMR)
Information Extraction {linguistic elements, sentiment terms and medical terminology}
Sentiment Analysis
review text
list of words
paragraph vector
word vector representation
word vector clusters
Recursive deep neural network
feature cluster
feature vector
classification
prediction result
Ontology (RDF Graph)
Summarization

Fig 6. Proposed Architecture Model for Sentiment and Prediction Analysis

- *Fifth Step*: Multiple tree is formed and each tree polarity is calculated. The average polarity of each tree is taken for predicting the sentiment.
- *Sixth step*: Ontology OWL is formed on the token medical keywords performed in step 1.By using the ICD-10 code we are able to create RDF and find the relationship between the disease and the symptoms.
- After Prediction Summarization is performed which is in unsupervised learning. Hence the Summary will predict the future possibility too.

4.1 Proposed Architecture:

4.1.1 Information Extraction

The first step of the processing pipeline is the extraction process which forms the basis for the further analysis. First, linguistic elements, sentiment terms and medical terminology are extracted. Linguistic elements include numbers, stop words, punctuation and part of speeches (noun, verb, adjective, adverb, pronoun, etc.), which are the surface symbols from the text. As next, the sentiment terms in clinical narratives are obtained through dictionary taggers based on Subjectivity Lexicon [14]. The sentiment terms indicate a patient's situation, e.g. ”least” shows negative outcome, while “overcome” represents a positive result of a clinical investigation. Medical terminology referring to symptoms, diseases and anatomical concepts are identified and matched with concepts of standard medical terminology. The extracted information is exploited by further analysis methods described in the following diagram.

4.1.2 Sentiment Analysis

In this paper we are comparing the old method/ techniques of find the sentiment Analysis and summarization to the approach of sentiment prediction and summarization via ontology. Clinical sentiments are distributed in the recommendation, suggestion, and judgments in the differential diagnosis, suspicion, operation result, treatment outcomes. The impression from nurses on a patient's health status documented in nurse letters or similar text are also considered. In this paper we try to achieve the objective through a voting algorithm to calculate the number of positive and negative sentiment, so that the polarity categories (negative, neutral, positive) of a document can be assigned. Referring from the table 2 the common methodology can be discussed as-

A skip- gram model: An Objective is to find word representations which are helpful in predicting the surrounding or related words

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-e\leq j<e,j\neq 0}\log p(w_{t+j})/w_t$$

To maximize the average log probability where c is the size of the training context

$$p\,(w_0/w_1) \quad = \quad \frac{\exp\,(v'_{wo}v^{T}_{w1})}{\sum_{w=1}^{W}\exp\,(v'_{wo}v^{T}_{w1})}$$

.

Defines $p(w_{t+j}/w_t)$ using the softmax function where vw and v'w re the input and output vector representations

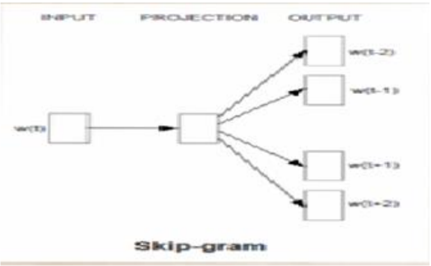


Fig 7. Skip gram model

w, and W is the number of words in the vocabulary.

Naïve Based Approach: Bayesian equation for calculating the posterior probability can be given as:

$$P(\frac{c}{x}) = \frac{P(\frac{x}{c}) \times P(c)}{P(x)}$$

$$P(c/x) = P({}^{x_1}/_{c}) \times P({}^{x_2}/_{c}) \times P({}^{x_3}/_{c}) \ldots\ldots\ldots P({}^{x_n}/_{c}) \times P(c)$$

Where $P(x_{1/c})$ *= posterior probability,*
$P(c/x)$ *= Likelihood,*
$P(c)$ *= Class Prior Probability,*
$P(x)$ *= predictor prior probability*

K-mean Clustering Algorithm: Classification is done using Naïve Bayes Classification and Support Vector Machine. Naïve Bayes Classification is based on supervised learning. It is a statistical method for classification. It computes the probabilities of the outcomes to determine whether a sample belongs to a particular class or not. It is used for both diagnostic and predictive problems. Clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as with-cluster sum-of-square [12].

$$\sum_{i=0}^{n} \min_{\mu_{j\epsilon C}}(||x_j - \mu_i||^2)$$

Recursive Neural deep learning tree Model*:* The Stanford Sentiment Treebank is the principal corpus completely marked parse trees that that takes into account an entire analysis of the compositional impacts of sentiment in language [19]. So as to catch the compositional impacts with higher precision, we are using the model called the Recursive Neural Deep Learning Network (RNDN). Recursive Neural Deep Learning Networks take as info expressions of any length. They represent a sentence by forming word vectors and a parse tree and then compute the compositional vectors for upper-level nodes in the tree by using the same recursive-based organization work.

Initially, we assume the phrase as a slider and set the slider to the neutral position. In this section, we represent the model which computes the representations of the compositional vector for English phrases having variable length and of different syntactic type. The working of the recursive model can be termed as:

1. These vector representations of the phrases elements will then be utilized as components to characterize each phrase. In Fig. 9 we try to displays this approach.
2. At the point when an n-gram is given to the compositional models, it parsed them into a binary tree and each leaf node, as analogous to a word, is represented as a vector.
3. The Recursive neural model works in a bottom to up fashion to calculate the similar utility for the parents by using diverse types of compositionality functions g.
4. Now the parent vectors are again considered as a vector and will be resulted as features to a classifier. For simplicity of article, we will utilize the tri-gram in this figure to clarify all models. Here each word of a phrase represents as an m-dimensional vector.

In this model, we initialize the value of the word vector by giving the value obtained from a uniform distribution in a random sampling as U (-k; k), where k = 0:0001.

Here after all the formed word vectors are stacked in the matrix called as word embedding matrix L.

$L \in R^{m \times |Vs|,}$

Where |Vs| is the size of the vocabulary.In order to optimize the word vector we consider them as a feature inputs to a softmax classifier. For classification into five level of class, we computed the posterior probability of each word vector given as:

$$Y_a = \text{softmax}\ (W_s\, a)$$

Where $W_s \in R^{5 \times m}$ is the sentiment classification matrix.

For the given tri-gram, this is rehashed for vectors b and c. The primary task of and distinction between the models is to compute the hidden vectors $p_i \in R^m$, from bottom to up fashion

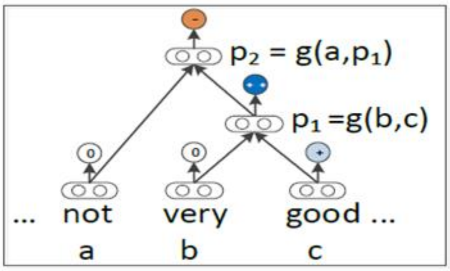


Fig 9. Recursive neural tree model where literal token semantic is processed by using the function g

4.1.3 Ontology:

A vocabulary for events and situations, and Bio Portal as reference for thematic roles of events. We compute its sentiment score by combining the scores of its associated opinion features, which are extracted from the RDF graph representing the opinionated sentence. If the topic participates in an event or a situation occurrence, we say that such occurrence provides a context to it, and affects its sentiment score. The flow diagram of the proposed model is shown in figure 6.This paper trying to enhance the emotion of the EMR using predication analysis with the sentiment predication as an unsupervised learning.

In this paper we are using Bio-Portal as a source of datasets. The Bio-portal dataset is structured as Ontology, Metadata and Mapping. The predicate are mapped by using common properties such as subPropertyof In order to map the predicate terms to its label the skos:prefLabel property is used. Similarly for synonym the used property for mapping is skos:altLabel. We have human-disease ontology, Symptom ontology and UMLs need to show the relation between these vast vocabularies. We need to find the relation between

a. Person characteristics and possible human diseases.
b. Human diseases and their symptoms
c. Symptoms and medicine and so on.

Human-Disease Ontology

Human-Disease Ontology represents an extensive learning base of inherited, developmental and acquired diseases. It coordinates sickness and therapeutic vocabularies through the utilization of cross-mappings and combination of NCI's thesaurus, ICD, MeSH, SNOMED CT and OMIM disease-specific terms and identifiers. The DO is used for describing the disease annotation by major biomedical databases (e.g. NIF, IEDB, Array Express), At the initial point of writing it consist of 8681 disease classes, 2260 of which have printed definitions annotated with medical terminology and disease attributes, such as side effects, symptom, phenotype, anatomical body location and etc. *has symptom* property utilized as a part of triples to comments on the pre-defined definitions and its metadata such as symptom information and only 388 distinct classes are explained with this predicate in the definition.

Symptom Ontology

The Symptom Ontology was produced as a component of the Gemina project [9]. It is made around the idea of a manifestation being: "An apparent change in function, sensation or appearance detailed by a patient symptomatic of a disease". SYMP is primarily structured by body locales with a branch for general symptoms. In July 2008 the Symptom Ontology was submitted for inclusion and survey to the OBO Foundry and was embraced [18].

Unified Medical Language System (UMLS)

UMLS started in 1986 by US National Library of Medicine is a system for integrating major vocabularies and standards from biomedical domain, such as SNOMED CT, MeSH, ICD, LOINC, RxNorm and several others Image[18].We select only those ontologies in Bio Portal, that we are certain about the presence of the required information in them. In the biomedical domain, classes and expressions play essential parts in ontologies as opposed to the occurrence information inside different areas. In this manner, we recover the applicable classes to us in the chose ontologies and characterize them as center classes. At that point, we request for the mappings that contain the required class as a mapping hotspot for each of the center classes.

Then it stores the objectives of these mappings as potential classes, together with the basic source of mappings and ontologies they come from. Later, on the basis of facts and characteristics, we decided them to keep which will appear in the course of analysis.

5. Result and Discussion

Table 3. Medical history of the patient Example reported by physician and nurses in EMR.

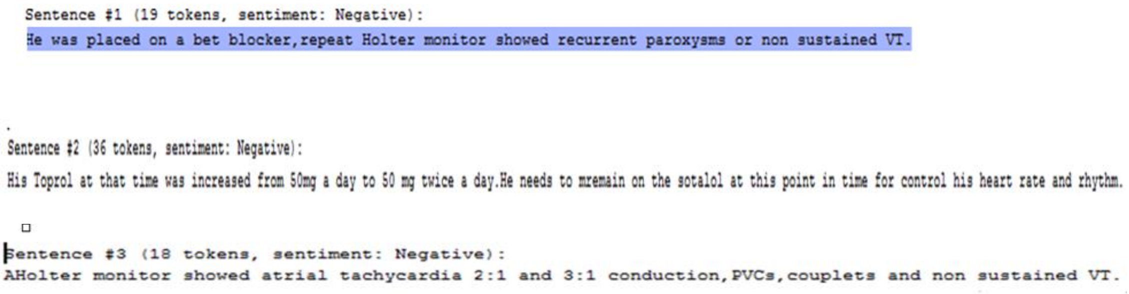
S No	***Medical history***
1	He was placed on a bet blocker, repeat Holter monitor showed recurrent paroxysms or non-sustained VT. His Toprol at that time was increased from 50mg a day to 50 mg twice a day. He needs to mremain on the sotalol at this point in time for control his heart rate and rhythm. AHolter monitor showed atrial tachycardia 2:1 and 3:1 conduction, PVCs, couplets and non-sustained VT.

Initially we extract the phrases and facts from the given history as.

1. He was placed on a bet blocker, repeat Holter monitor showed recurrent paroxysms or non-sustained VT
2. His Toprol at that time was increased from 50mg a day to 50 mg twice a day. He needs to remain on the sotalol at this point in time for control his heart rate and rhythm.
3. A Holter monitor showed atrial tachycardia 2:1 and 3:1 conduction, PVCs, couplets and non-sustained VT.

Information extraction:-

As we defined the stop words in section 4, it help us to phrase the raw data. With the help of these words we can claculat5e the collection frequency of the word in a phrase and it also helpful in estimating the next preceding word in a sentence. Here is the list of few stop words shown in figure 10



```
Sentence #1 (19 tokens, sentiment: Negative):
He was placed on a bet blocker,repeat Holter monitor showed recurrent paroxysms or non sustained VT.

Sentence #2 (36 tokens, sentiment: Negative):
His Toprol at that time was increased from 50mg a day to 50 mg twice a day.He needs to mremain on the sotalol at this point in time for control his heart rate and rhythm.

Sentence #3 (18 tokens, sentiment: Negative):
AHolter monitor showed atrial tachycardia 2:1 and 3:1 conduction,PVCs,couplets and non sustained VT.
```

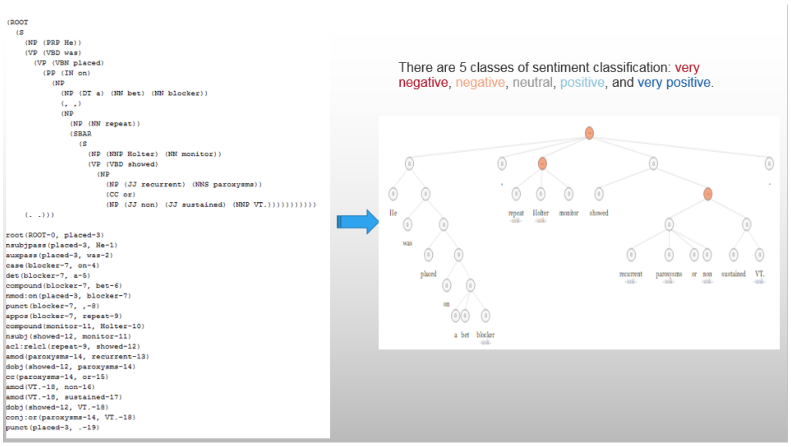
Fig 10. Extracting the phrases and tokens from the given history.

Here we try to create Medical terminology database based on ICD CODE10. Sentiment terms are extracted with the help of semantic simplification having key relation .Connectors are useful in explaining the semantic type information. In the table 4 we have given the few connectors to extract the information and simplify the raw data to get information from it. This information is further helpful in summarization of the data.

Table 4. Example of Connectors used to relation between disease and test or symptoms

Original Term Sem. Group	Connector	Explanation Term Sem. Group
Disease name	*a condition affecting*	Anatomical Structure
Anatomical Structure	*a part of*	Anatomical Structure
Device	*a device/instrument used in*	Procedure
Procedure	*a procedure performed on*	Anatomical Structure
Medication	*can have a tradename of*	Medication

Sentiment Analysis: - In this analysis, we divide the sentence into words and we create binary tree of that Phrase. The leaf nodes are small word vector. The tree consists of right and left subtrees. We apply recursive neural network for training the dataset to achieve unsupervised learning for further prediction. We retrieve polarity of each node and combine its result with neighboring nodes, recursively to form a single node at root level. Different trees are formed from same sentence by exchanging the nodes considered at each level. For each tree, we form a vector of leaf nodes and we predict polarity of the root. On the basis of polarity of root we find the ontology of root. The ontology of medical terms are identified which are retrieved during information extraction. The ICD CODE 10 is the standard ontology used in medical field which gives unique code for a particular disease. This standard also gives information about symptoms related to a specific disease. We can predict the future diseases based on the side effects of treatment and medical history of patient.



```
(ROOT
  (S
    (NP (PRP He))
    (VP (VBD was)
      (VP (VBN placed)
        (PP (IN on)
          (NP
            (NP (DT a) (NN bet) (NN blocker))
            (, ,)
            (NP
              (NP (NN repeat))
              (SBAR
                (S
                  (NP (NNP Holter) (NN monitor))
                  (VP (VBD showed)
                    (NP
                      (NP (JJ recurrent) (NNS paroxysms))
                      (CC or)
                      (NP (JJ non) (JJ sustained) (NNP VT.))))))))))
    (. .)))

root(ROOT-0, placed-3)
nsubjpass(placed-3, He-1)
auxpass(placed-3, was-2)
case(blocker-7, on-4)
det(blocker-7, a-5)
compound(blocker-7, bet-6)
nmod:on(placed-3, blocker-7)
punct(blocker-7, ,-8)
appos(blocker-7, repeat-9)
compound(monitor-11, Holter-10)
nsubj(showed-12, monitor-11)
acl:relcl(repeat-9, showed-12)
amod(paroxysms-14, recurrent-13)
dobj(showed-12, paroxysms-14)
cc(paroxysms-14, or-15)
amod(VT.-18, non-16)
amod(VT.-18, sustained-17)
dobj(showed-12, VT.-18)
conj:or(paroxysms-14, VT.-18)
punct(placed-3, .-19)
```

There are 5 classes of sentiment classification: very negative, negative, neutral, positive, and very positive.

Fig 11. Extraction of noun verb adjective from the sentences and construct the tree.

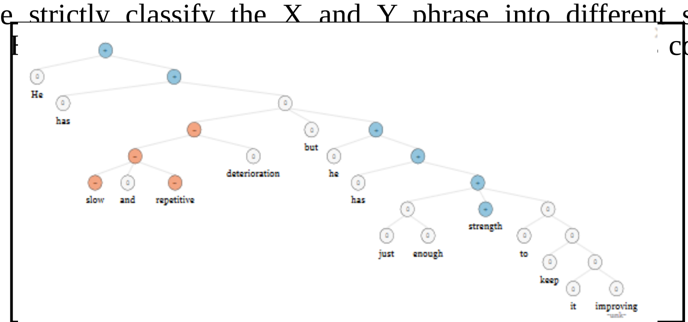
In this paper we considered to extract the Full Sentence Binary Sentiment and analyze the model on the basis of Contrastive Conjunction. It can be further has the cases like-

1. High Level Negation ⇐ Negating Positive Sentence
   Negating Negative Sentence
2. Most Positive and Negative Phrase.

Contrastive Conjunction:

In this section, we utilize a subset of the test set which incorporates just sentences with an *'X but Y'* structure: A phrase *X* being trailed by *but* which is further followed up by a phrase *Y.* The conjunction is deciphered as a contention for the second conjunct, with the principal working concussively (Lakoff, 1971; Blakemore, 1989; Merin, 1999) [19] Fig. 12 contains an example. Here we uses phrase as "He has slow and repetitive deterioration but he has just enough strength to keep it improving".

In this example we strictly classify the X and Y phrase into different sentiment including neutral also. F counted as



correct, only when the Phrases X and Y annotated sentiment are correct.

Fig 12. Showing X but Y tree formation.

High Level Negation:

We evaluated that there are two sorts of negation. In order to understand each kind we are taking different datasets for evaluation.

Set 1: ***Negating Positive Sentences***. The first set contains positive sentences and their negation. In this set, the negation changes the general assumption of a sentence from positive belief to negative sentiment. Hence, the model successfully computes the accuracy by using negative n-gram word vector .Here we try to compute sentiment having reversal from positive to negative. Fig. 13 shows examples of positive negation where the model correctly classified, even if n-gram weight for 'least' is very less in percentile.

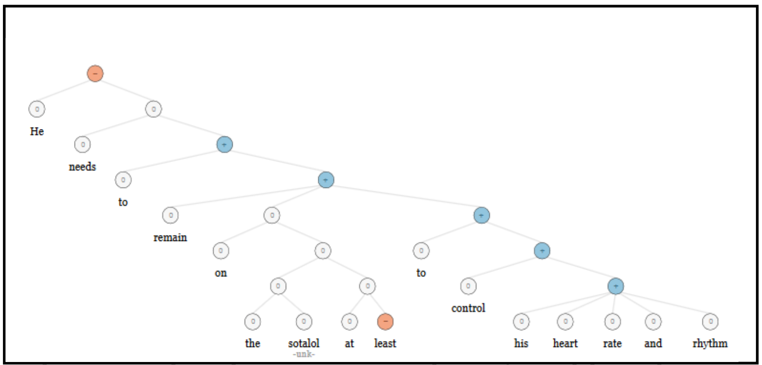


Fig 13. Sentiment estimation over negation of positive phrase

Set 2: ***Negating Negative Sentences***. In the second set we consider a negative sentiment sentence which has been followed up by its negation. As we all assume according to our mathematical assumption two negation will form a positive output but in phrases this theory is not always true.

When negative phrases are negated, the sentiment tree shows that overall sentiment which may become less negative, but not always necessarily to be positive. For instance, 'His health condition becomes bad is negative but the 'His health condition becomes least bad' says just that it was less awful than a horrible one, not that it was great (Horn, 1989; Israel, 2001) [19]. Henceforth, we assess precision as far as how frequently each model could increment non-negative actuation in the sentiment of the phrase. We show the example in fig 14.

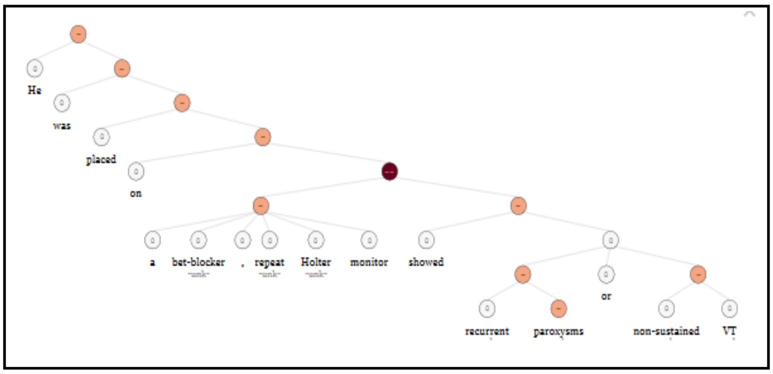


Fig 14 Sentiment estimation over negation of positive phrase

Most Positive and Negative Phrases:

In this case, we consider two sub phrases in which one of the sub phrase is highly positive and other sub phrase is negative .The overall contextual polarity be remain slightly negative. Here in the Fig 15 it shows like word vector recovers and rapidly have positive impact and remain has negative polarity.

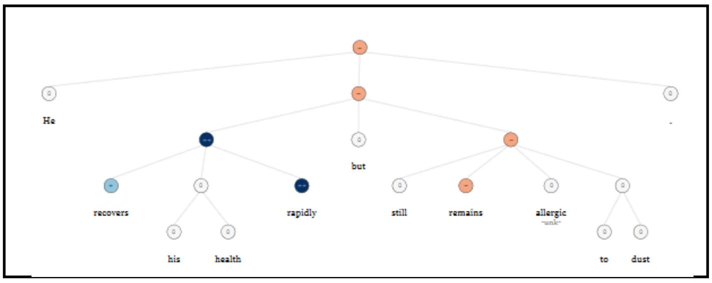


Fig 15. Most positive and Negative phrase tree construction

Ontology:

In this step we check the polarity of the phrase and use the tokenized medical terminologies which are extracted in during information retrieval. In this project we use ICD codes to identify the diseases and bio-portal to link these diseases with the symptoms and identify the relationship between the symptoms and the diseases.

RDF is available for human- diseases datasets. (DO).the diseases are categorized based on the body parts, tissues, genetic etc. We consider our previous example where in the phrase 3 we extracted *atrial tachycardia*. The RDF for the token is given below in fig .The

Now we form the SPRQL query to find the relation between the symptoms and medical history. The SPARQL query can be given as:

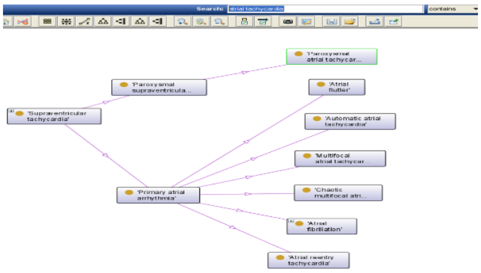
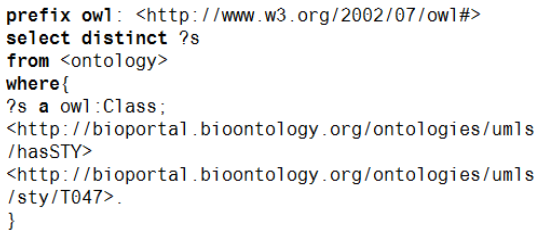

Search: atrial tachycardia
contains
'Paroxysmal atrial tachycar...
'Paroxysmal supraventricula...
'Supraventricular tachycardia'
'Atrial flutter'
'Automatic atrial tachycardia'
'Multifocal atrial tachycar...
'Primary atrial arrhythmia'
'Chaotic multifocal atri...
'Atrial fibrillation'
'Atrial reentry tachycardia'

Fig 16. RDF graph for token atrial tachycardia in protégé.

In Order to find the metadata for the term we can use SPARQL query as shown in figure. Since DO contains information just about the diseases, we just consider each of the OWL classes in DO a disease [18].

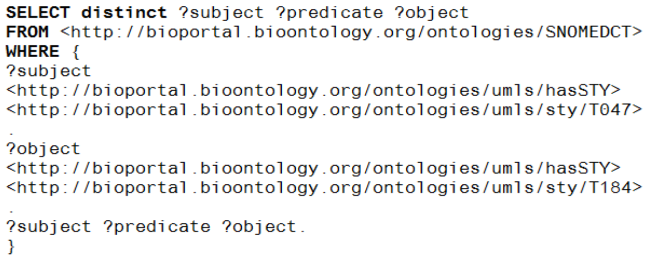
For retrieval of the medical term we can use semantic type T047 which elaborate about the diseases or the syndrome and in order to retrieve the metadata the *has STY* property can be use. The resulted data is from filtered UMLs .Here we can have core classes, potential classes and so on. Potential classes we consider those who resulted via mapping. These sort of classes may have overlapping but core classes are unique ones.



```
prefix owl: <http://www.w3.org/2002/07/owl#>
select distinct ?s
from <ontology>
where{
?s a owl:Class;
<http://bioportal.bioontology.org/ontologies/umls
/hasSTY>
<http://bioportal.bioontology.org/ontologies/umls
/sty/T047>.
}
```

Fig 17. SPARQL query for extraction of diseases from DO.

Now for retrieving more metadata we need to go for symptom ontology. Here we can use T184 that refereed to the SYMPTOMS OR SIGNS for the classes. Here we can use the property *hasSYMP* for retrieving data. According to the estimation around 34000 classes related to symptoms in 161 core classes amongst them around 11,000 are overlapped to the potential classes.



```
SELECT distinct ?subject ?predicate ?object
FROM <http://bioportal.bioontology.org/ontologies/SNOMEDCT>
WHERE {
?subject
<http://bioportal.bioontology.org/ontologies/umls/hasSTY>
<http://bioportal.bioontology.org/ontologies/umls/sty/T047>
.

?object
<http://bioportal.bioontology.org/ontologies/umls/hasSTY>
<http://bioportal.bioontology.org/ontologies/umls/sty/T184>
.

?subject ?predicate ?object.
}
```

Fig 18. SPARQL query for extraction of diseases and their symptoms from DO.

ISSN: 2005-4238 IJAST

The main connection in UMLS can defined on the overlapping of the classes of the diseases as well as the symptoms. In the figure we has explained the overlapping by considering classes as Core Disease class(C), potential Disease class(P),potential Symptom class(S),core Symptom class(D).In the figure few classes are labeled as diseases, few are symptoms but some of the classes are not able to label , so we keep them in both disease and symptoms .

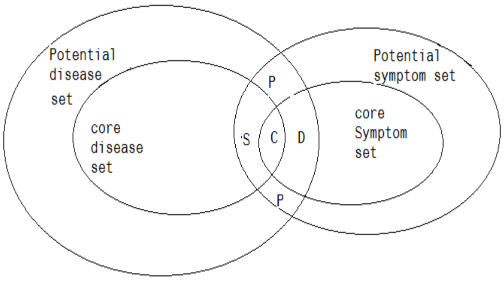


Fig 19. Overlapping/Relationship between diseases and symptoms from DO.

Summarization:

Many techniques are used to summarize the paragraphs. In the model we use simple technique given by A.Laxmisan et al. in 2014 has given the following steps using a previously described conceptual model: AORTIS (Aggregation, Organization, Reduction, Interpretation and Synthesis) [20].The algorithm approach can be referred to from the paper. In summarization we try let us take a small example like in a medical history the symptoms noted as A person is having headache and cold the prediction will be maybe he has the viral fever. Now let us suppose the patient got rashes or being allergic to some medicine due to high sugar level as described in medical history then the summarization part can be reviewed as “The person is a sugar patient and may be suffered by a viral fever and it is allergic to specific medicine.”

6. Conclusion

EMR is the emerging field which provides an infrastructure that allows hospitals, medical practices, insurance companies, and research facilities to tap improved computing resources at ***lower initial capital*** outlays. The focus of the analysis of medical health records was to establish a solid baseline for systems predicting health conditions – specifically ICD-9 codes from text data. The ability to reliably predict outcomes to improve quality of care. The main aim of the paper is to summarize the medical history and make the process to be automated along with having the predication of Sentiment Analysis is the underlying way which helps to improve the notion of EMR. The recursive deep neural learning improves the predication and learning of test cases in a better way. Ontology relatively support the prediction and enhance the learning to extract the valid and useful information.

7. Future Scope

Automating the process is major aim of this paper. These improvements should concentrate on the extraction phase: first, the proximity-based extraction is required, since the most desirable information exists in the special sections such as impression, conclusion. We can also look over the factor where the model simply makes phrases very negative when negation is in the sentence? The next experiments show that the model captures more than such a simplistic negation rule and so on.

References

[1] Gonzenbach, Maurice. Sentiment Classification and Medical Health Record Analysis using Convolutional Neural Networks. Diss. 2016.

[2] Pathan, Farheen, and Anjali Phaltane. "DUAL SENTIMENT ANALYSIS."

[3] Siegrist Jr, R. B., and S. Madden. "Sentiment analysis turns patients' feelings into actionable data to improve the quality of care." The Science of Emotion (2011): 27-35.

[4] Ranjitham, G. Grace, S. Mohana, and B. Vinothini. "Sentiment Analysis For Two Sides Of Reviews Using Dual Prediction Algorithm."

[5] Haque, Md. "Sentiment analysis by using fuzzy logic." arXiv preprint arXiv:1403.3185 (2014).

[6] Dragoni, Mauro, Andrea GB Tettamanzi, and Célia da Costa Pereira. "A fuzzy system for concept-level sentiment analysis." Semantic Web Evaluation Challenge. Springer International Publishing, 2014.

[7] Recupero, Diego Reforgiato, Aldo Gangemi, Andrea Giovanni Nuzzolese, Sergio Consoli, Daria Spampinato, and Valentina Presutti. "Semantic Web-based Sentiment Analysis." In SSA-SMILE@ ESWC, pp. 25-28. 2014.

[8] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexiconbased methods for sentiment analysis. Computational linguistics, volume 37, number2, 267–307, MIT Press

[9] Songbo Tan, Jin Zhang, "An empirical study of sentiment analysis for chinese documents", Expert Systems with Applications 34 (2008) 2622–2629.

[10] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, Andr´es Montoyo. 2010. A survey on the role of negation in sentiment analysis. Proceedings of the workshop on negation and speculation in natural language processing 60–68, Association for Computational Linguistics.

[11] Kaushik, Chetan, and Atul Mishra. "A scalable, lexicon based technique for sentiment analysis." arXiv preprint arXiv:1410.2265 (2014).

[12] https://prezi.com/pzl-yt3epq4a/using-word-vector-cluster-for-sentiment-analysis

[13] S.Madupalli et al., Structural performance of non-linear analysis of turbo generator building using seismic protection techniques, International Journal of Recent Technology and Engineering, 8(1), (2019), 1091-1095.

[14] Thye, J., M. C. Straede, J. D. Liebe, and U. Hübner. "GMDS 2014: 59. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V.(GMDS)." (2014).

[15] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.

[16] Chunxu Wu, Lingfeng Shen, "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions", 2009 International Conference on Artificial Intelligence and Computational Intelligence

[17] http://healthitanalytics.com/news/four-use-cases-for-healthcare-predictive-analytics-big-data

[18] Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, p. 1642. 2013.

[19] Laxmisan, A., A. B. McCoy, A. Wright, and D. F. Sittig. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. Appl Clin Inf. 2012; 3 (1): 80–93. doi: 10. 4338. ACI-2011-11-RA-0066. Retrieved from PM: 22468161.