LEVERAGING OF WEIGHTED ENSEMBLE TECHNIQUE FOR IDENTIFYING MEDICAL CONCEPTS FROM CLINICAL TEXTS AT WORD AND PHRASE LEVEL

Dipankar Das and Krishna Sharma

Department of Computer Science & Engineering, Jadavpur University, India

ABSTRACT

Concept identification from medical texts becomes important due to digitization. However, it is not always feasible to identify all such medical concepts manually. Thus, in the present attempt, we have applied five machine learning classifiers (Support Vector Machine, K-Nearest Neighbours, Logistic Regression, Random Forest and Naïve Bayes) and one deep learning classifier (Long Short Term Memory) to identify medical concepts by training a total of 27.383K sentences. In addition, we have also developed a rule based phrase identification module to help the existing classifiers for identifying multi- word medical concepts. We have employed word2vec technique for feature extraction and PCA and T- SNE for conducting ablation study over various features to select important ones. Finally, we have adopted two different ensemble approaches, stacking and weighted sum to improve the performance of the individual classifier and significant improvements were observed with respect to each of the classifiers. It has been observed that phrase identification module plays an important role when dealing with individual classifier in identifying higher order n-gram medical concepts. Finally, the ensemble approach enhances the results over SVM that was showing initial improvement even after the application of phrase based module.

KEYWORDS

Medical Concepts, Phrase Identification, Ensemble, Machine Learning.

1. INTRODUCTION

In recent trends of digital platforms, people in general are relying on electronic data because the number of active internet users is increasing in medical domain[11]. It is very necessary to develop a state of the art tool to extract medical phrases from raw unstructured text.

We have developed a structured dataset of bio-medical concepts by manually annotating each and every term as either medical or not by collecting huge amount of raw data from the web archives. The training data consists of 27383 sentences while 7283 sentences are available as test data.

In the present work, we have developed a model that identifies medical concepts from texts as well as helps medical practitioners as well as novice users to deal with unstructured data. One instance of input and output of our model is shown as follows.

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

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.161-173, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111213

Input: *Amlodipine is used with or without other medications to treat high blood pressure*.

Output: ***Amlodipine_MC*** *is_O used__O with__O or_O without_O* ***[other medications] MC*** *to__O* ***treat__MC*** ***[high blood pressure]__MC .*** *_O .*

In the above example, the words (phrases) tagged with "__MC" are medical terms and the words that are tagged with "__O" are non-medical terms. The phrases are separated with brackets "[ ]". It has been observed that the presence of non-medical words also invokes the sense of a medical concept. For example, the words in italic are non-medical words whereas their appearance along with medical words forms a phrase level medical concept (e.g., "*Rat* Fever", "*Indian* Medical *Association*" etc.). For this reason, phrase identification module plays an important role and some set of rules are defined by considering medical as well as linguistic features. Moreover, support and confidence are also measured in order to identify the best possible phrase identification rules to tag multi-word medical concepts. Performances of the individual classifier before and after applying phrase identification are less while comparing the performance of the ensemble approach.

Finally, we have applied an ensemble approach to combine multiple classifiers to predict better than that of the individual classifier. The evaluation result shows that the ensemble approaches outperform other classifiers. We have applied two ensemble approaches i.e. stacking and weighted sum. Stacking helps to identify unigram medical concepts whereas weighted sum out performs multiword n-grams where n lies between 2 to 5**.**

The rest of the paper is organized as follows. The literature survey on extracting medical entities by machine learning classifiers is discussed in Section 2. The dataset preparation is discussed in Section 3 whereas machine learning and deep learning frameworks are described in Section 4. The phrase identification module is described in Section 5 followed by its evaluation results over ML approaches as discussed in Section 6. In contrast to ML and DL, Section 7 illustrates the implications of ensemble approaches and Section 8 highlights the feature selection strategies for improving results along with critical observations. Finally, Section 9 concludes the paper by mentioning future tasks.

2. RELATED WORK

Biomedical information extraction from the unstructured data is considered as one of the emerging challenges in the research field of NLP. Hence, a domain specific lexicon has become an essential component for converting a structured corpus from the unstructured corpus. Also, it helps in extracting the subjective and conceptual information related to medical concepts from the corpus.

Various researchers have tried to build various ontologies and lexicons such as UMLS, SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), MWN (Medical WordNet), SentiHealth, and WordNet of Medical Events (WME 1.0 and WME 2.0) etc. in the domain of healthcare.

UMLS helps to enhance the access to biomedical literature by facilitating the development of computer systems that understand biomedical language (Bodenreider, 2004). SNOMED-CT is a standardized, multilingual vocabulary that contains clinical terminologies and assists in exchanging the electronic healthcare information among physicians (Donnelly, 2006).

Furthermore, Fellbaum and Smith (2004) proposed Medical WordNet (MWN) with two sub-

networks e.g., Medical FactNet (MFN) and Medical BeliefNet (MBN) for justifying the consumer health. The MWN follows the formal architecture of the Princeton WordNet (Fellbaum, 1998). On the other hand, MFN aids in extracting and understanding the generic medical information for non-expert groups whereas MBN identifies the fraction of the beliefs about the medical phenomena (Smith and Fellbaum, 2004). Their primary motivation was to develop a network for medical information retrieval system with visualization effect.

Being in the similar trends, SentiHealth lexicon was developed to identify the sentiment of the medical concepts (Asghar et al., 2016; Asghar et al., 2014). In recent times, WME 1.0 and WME 2.0 lexicons were designed to extract the medical concepts and their related linguistic and sentiment features from the corpus (Mondal et al., 2016; Mondal et al., 2018).

These mentioned ontologies and lexicons assist in identifying the medical concepts and their sentiments from the corpus but unable to provide the complete knowledge of such concepts. Hence, in the current work, we are motivated to design a full-fledged lexicon in healthcare which provides the linguistic and knowledge-based features together for the medical concepts.

3. DATA PREPARATION

A total of 170 medical e-books of various sub-domains such as anatomy, internal, medicine, physiology, biochemistry etc. were collected from various web archives. Such books are mainly recommended for medical degree courses. Some of the books are text books[2], some books are medical encyclopedia[3], and a few are medical dictionaries[4]. We have extracted texts from the pdf files of all such books using open source tika[5] python library.

Finally, we have trained a word2vec word embedding model (Embedding size ~ 100) using these texts. We have used gensim[6] python library for training purpose. This large collection of text is used only for training our own word embedding whereas we have selected only a part of these texts for training and test purposes of the machine learning and deep learning classifiers, separately.

On the other hand, we collected a total of 34666 sentences from a medical dictionary[7]. These sentences are split into 27383 sentences for training and 7283 sentences for test purposes. The training set contains 498734 words whereas test set contains 130662 words, respectively. We have mentioned the brief details of our training and test data. In this Table 1, the statistics denote for medical words / phrases only.

[2] https://medicostimes.com/all-mbbs-books-pdf/
[3] Gale encyclopedia vol 1 to 5
[4] Dictionary of Medical Terms 4th Ed.- (Malestrom) and Black's medical dictionary etc.
[5] https://pypi.org/project/tika
[6] https://radimrehurek.com/gensim/models/word2vec.html
[7] BLACK'S MEDICAL DICTIONARY 41ST EDITION

Table 1. Statistics of the dataset

Dataset	# of N in N-gram					# phrases
	N=1	*N=2*	*N=3*	*N=4*	*N=5*	
Training	43734	41273	14904	4134	1247	134826
Test	11242	12245	3889	1048	334	29534
Total	54976	51518	18793	5182	1581	164360

4. SYSTEM FRAMEWORK

We have used five machine learning models followed by one deep learning model. We have used SVM (degree of SVM polynomial kernel is 3, and C=1.0.), K-NN (K=4), Logistic Regression, Gaussian Naïve Bayes and Random Forest algorithms for developing our machine learning framework.

We have applied these 5 machine learning classifiers to explore a comparative study among their performances with respect to the classification of medical concepts. Apart from that, we have selected multiple classifiers because we wanted to enhance the performance of classification framework by applying ensemble technique. As the classifiers require features for learning, we have used word embedding to convert word to feature vector and employed as features.

For machine learning classifiers, we have used scikit-learn[8] python library and for our deep learning framework using LSTM (Long and Short Term Memory) model, we have used keras[9] python library. In the LSTM, we have used time distributed character embedding with output dimension 20. In this layer, we have used LSTM unitof 64 with recurrent dropout=0.1. In the next layer, we have used LSTM unit of 256 with recurrent dropout=0.1 and in the last layer, we have employed a dense layer with *softmax* activation function. In this model, we have used *adam* optimizer with *binary_crossentropy* loss function.

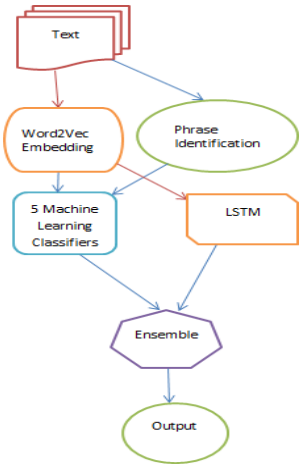

Figure 1: System Framework

[8] https://www.scikit-learn.org
[9] https://www.keras.io

We have discussed various steps for concept identification from medical texts. If we summarize all the process and understand the steps one by one, we have to look into the following diagram. In the above diagram, texts are sent to word2vec model for feature extraction that extracts the words and their features. The same text is sent to phrase identification module and it returns a list of words and phrases according to the sequence of the input text. Each word and its corresponding vectors were extracted from the text are sent to machine learning classifiers. If our classifiers observe a single word, it predicts whether it is medical term or not and if it does not see a multi word, it predicts such that if a multiword expression contains at least one medical term or notand declares this as a medical phrase. This same text with word2vec word embedding is sent to LSTM module. We send the outputs of the classifiers to ensemble module to increase the performance. After ensemble, we receive the final output.

However, these classifiers are not good enough with sequential data and thus unable to classify multi-word or phrase level medical concepts. In order to identify such phrase level medical concepts, we have developed a rule based phrase identification module for our task. The phrase level module helps in machine learning classifiers whereas in case of developing our deep learning framework using LSTM model with character embedding, we did not employ phrase identification module. We have also compared the performance between LSTM and the 5 machine learning classifiers with phrase identification.

5. PHRASE IDENTIFICATION

As the presence of non-medical words also invokes the sense of a medical concept, we have developed the rule based phrase identification module. Further, in order to handle sequential data using 5 machine learning classifiers, we have built this.

If we want to predict a medical phrase having a non-medical term as a whole, our classifier will predict that non-medical term as a medical whenever it occurs. E.g., **"***Indian Medical Association***"** is a medical phrase where the words, "*Indian*" and "*Association*" are not medical terms. However, if we want to predict this phrase as medical, our phrase identification module plays a vital role. We have used nltk[10] python library for phrase identification. The algorithm is as follows:

Step 1: *In the first pass, our algorithm will extract single and multiword medical concepts using phrase identification module.*

Step 2: *In the second pass, our classifier will predict all the single word expression and multiword such that if a multiword expression contains at least one medical term it predicts this as medical phrase*.

We have defined some phrase identification rules. Support and Confidence are measured in order to identify the best possible phrase identification rules to identify multi word medical concepts. The descriptions are given below.
In the above Table 2, we can observe that rule 1, 2, 3, 4, 10, 12, 14 are crucial to find multi word medical concepts and finally, we selected these rules only while

[10] https://www.nltk.org

applying into the framework of machine learning. In the next section, we have compared the performance of LSTM model with respect to each of the five machine learning classifiers and the importance of this phrase identification module is visible. Some example of the medical phrases with respect to each of the rule is given below.

Table 2: List of all phrase identification rules

RULES	**Support**	**Confidence**
RULE1 : {<JJ|NN|NNP><NNS>}	4523	0.258
RULE2{<VBP|VBN|NN|DT><NN>+}	7237	0.413
RULE3: {<RB><VBD|VB><NN>}	485	0.027
RULE4: {<DT><JJ>+<NN>}	3753	0.214
RULE5: {<DT><NN><JJ>}	26	0.0014
RULE6: {<VB><IN><NN>}	38	0.0021
RULE7: {<VB><TO><NN>}	8	0.0004
RULE8: {<NN><IN><VBG>}	84	0.004
RULE9: {<NN><CC><NN><VBZ>}	62	0.003
RULE10:{<NN.><NN.><NN>*}	430	0.024
RULE11: {<DT><RB><JJ><NN>*}	87	0.004
RULE12: {<CD>*<NN><IN><NN>}	384	0.021
RULE13: {<NNP><NN>+}	88	0.005
RULE14: {<RB>*<CD><NNS>}	311	0.017

Table 3: Phrase identification rules with examples

RULES	**Examples**
RULE1 : {<JJ|NN|NNP><NNS>}	Severe symptoms
RULE2{<VBP|VBN|NN|DT><NN>+}	The liver
RULE3: {<RB><VBD|VB><NN>}	Significantly lower cholesterol
RULE4: {<DT><JJ>+<NN>}	The tympanic membrane
RULE5: {<DT><NN><JJ>}	The autonomic nervous System
RULE6: {<VB><IN><NN>}	Lack of oxygen
RULE7: {<VB><TO><NN>}	Leads to death
RULE8: {<NN><IN><VBG>}	difficulty in breathing
RULE9: {<NN><CC><NN>}	Anoxia and hypoxia
RULE10:{<NN.><NN.><NN>*}	Catecholamine substances
RULE11: {<DT><RB><JJ><NN>*}	A potentially life-threatening condition
RULE12: {<CD>*<NN><IN><NN>}	Encyclopedia of medicine
RULE13: {<NNP><NN>+}	X Chromosome
RULE14: {<RB>*<CD><NNS>}	22 autosomes

6. EVALUATION

The test dataset consists of 7283 sentences (130662 words) manually annotated. We have analyzed 5 traditional ML algorithms (SVM, K-NN, LR, Naïve Bayes, Random Forest), and we have shown that these classifiers can also perform well in sequential data while using phrase identification module. We have used one deep learning (LSTM with character embedding) for classification to avoid phrase identification. As LSTM performs better in case of the sequential data by default, therefore, we did not apply phrase identification module into it. We have evaluated the performance of every classifier in phrase level. After evaluation we have increased our model's performance using ensemble method.

Table 4: Performance metrics of the classifiers

Classifiers	N in N-gram	**Precision**	**Recall**	**F1-Score After** phrase identification	**F1-Score Before** phrase identification
SVM	1	0.88	**0.95**	**0.92**	**0.92**
	2	**0.91**	0.93	**0.92**	**0.81**
	3	0.87	0.92	0.90	0.78
	4	0.72	0.94	0.81	0.60
	5	0.55	0.80	0.65	0.00
RandomForest	1	**0.77**	0.73	0.75	0.75
	2	**0.77**	0.79	**0.78**	0.72
	3	0.65	0.80	0.72	0.74
	4	0.39	**0.84**	0.53	0.41
	5	0.22	0.60	0.32	0.00
Naïve Bayes	1	0.63	**0.87**	0.73	0.73
	2	**0.78**	0.85	**0.81**	0.74
	3	0.67	0.85	0.75	0.60
	4	0.51	0.85	0.64	0.55
	5	0.35	0.73	0.75	0.00
Logistic Regression	1	0.71	0.72	0.71	0.71
	2	**0.73**	0.77	**0.75**	**0.65**
	3	0.60	0.75	0.66	0.59
	4	0.35	**0.78**	0.49	0.38
	5	0.20	0.53	0.29	0.0
KNN	1	0.88	**0.95**	**0.92**	**0.92**
	2	**0.91**	0.93	**0.92**	**0.81**
	3	0.87	0.92	0.90	0.76
	4	0.72	0.94	0.81	0.70
	5	0.55	0.80	0.65	0.00
LSTM	1	0.87	0.91	0.89	NA
	2	0.90	0.92	0.91	NA
	3	0.83	0.78	0.80	NA
	4	0.69	0.65	0.67	NA
	5	0.58	0.60	0.59	NA

From the above Table 4, we can observe that SVM, KNN and LSTM have performed well among all the other classifiers. From the last two columns (F1-Score after and before phrase identification), we can conclude that phrase identification plays a vital role in multi-word medical concept. It is also observed that the performance is decreased while predicting higher order N-grams. As phrase identification is not used in LSTM, performance analysis of before and after phrase identification of LSTM is not applicable. It is also noticed that SVM and K-NN with phrase identification rules perform better than LSTM. It means we can conclude that phrase identification is a key task of medical concept identification and classification.

7. ENSEMBLE APPROACH

In conventional machine learning, ensemble is a technique that uses multiple learning algorithms to obtain better performance which could not be obtained from any of the single learning algorithm alone. In this paper, we have used one type of ensemble approach i.e. “Weighted_Sum”. We will discuss about the performance gain as follows.

We have used six different classifiers and observed that three classifiers (e.g., SVM, LSTM and KNN) performed well and rest of the three (Random Forest, Naïve Bayes and Logistic Regression) performed moderate. As we tried to increase the overall performance, we finally selected SVM, KNN, LSTM, RF as the top performers and therefore ensemble them to improve the performance of our system.

We used weighted sum approach for ensemble and have given higher weight to the classifiers that obtain higher accuracy. Similarity in results between a pair of classifiers with respect to specific n-grams is also observed. Our motivation is to combine such output and predict better. The weighted sum is calculated as follows. Suppose, we have *n* number of classifiers and their outputs are α_1, α_2……..α_{n-1}, α_n. If we have assigned certain weights for each of the classifiers such as ω_1, ω_2……ω_{n-1}, ω_n, then the weighted sum for the output of the classifiers is If $\sum (\alpha_i * \omega_i) > £$ (£ is some threshold value), we classify it as **"MC"** class and otherwise, we classify it as "**O**" class.

In our approach, we have started by employing all the six classifiers and let the output of the classifiers are: *svm, knn, lr, lstm, nb* and *rf,* respectively. We have compared multiple possible weightages of all the classifiers and compared the F1-Score of all the combinations. In the following Table, we have given some instances of the combination of weights with respective £.

Table 5.1: Instances of the combination of weights When £=0.4

When £=0.4	**SVM**	**KNN**	**LSTM**	**LR**	**F1**
	0.25	0.25	0.25	0.25	0.91
	0.25	0.30	0.30	0.15	0.92
	0.10	0.25	0.30	0.35	0.82
	0.30	0.30	0.30	0.10	0.93
	0.35	0.30	0.25	0.10	0.93

Table 5.2: instances of the combination of weights When £=0.5

SVM	**KNN**	**LSTM**	**LR**	**F1**
0.25	0.25	0.25	0.25	0.91
0.25	0.30	0.30	0.15	0.88
0.10	0.25	0.30	0.35	0.84
0.30	0.30	0.30	0.10	0.91
0.35	0.30	0.25	0.10	0.92

Table 5.3: instances of the combination of weights When £=0.6

SVM	**KNN**	**LSTM**	**LR**	**F1**
0.25	0.25	0.25	0.25	0.92
0.25	0.30	0.30	0.15	0.91
0.10	0.25	0.30	0.35	0.82
0.30	0.30	0.30	0.10	0.92
0.35	0.30	0.25	0.10	0.94

Table 5. 4: instances of the combination of weights When £=0.7

SVM	KNN	LSTM	LR	F1
0.25	0.25	0.25	0.25	0.93
0.25	0.30	0.30	0.15	0.93
0.10	0.25	0.30	0.35	0.84
0.30	0.30	0.30	0.10	0.94
0.35	0.30	0.25	0.10	0.96

\

Table 5.5: instances of the combination of weights When £=0.65

SVM	KNN	LSTM	LR	F1
0.25	0.25	0.25	0.25	0.88
0.25	0.30	0.30	0.15	0.83
0.10	0.25	0.30	0.35	0.80
0.30	0.30	0.30	0.10	0.89
0.35	0.30	0.25	0.10	0.89

From the above tables from 5.1 to 5.5, we can find that our optimal weights are $W = \{0.35, 0.3, 0.25, 0.1\}$ and optimal threshold value, $£ = 0.65$. From our observation, we have derived the following weighted sum ensemble equation.

$$\sum (\alpha_i * \omega_i) = .35*svm+0.3*knn+0.25*lstm+0.1*lr$$

If $\sum (\alpha_I * \omega_i) > 0.65$, we will classify it as "**MC**" class, otherwise, we classify it as "**O**" class. Using this approach, we have improved the performance of our classifiers. The performances of our classifiers after ensemble are shown in Table 6. We can observe that after ensemble, the performances of our classifiers have increased, especially for the multi-gram concept identification.

Table 6. Performance metrics after ensemble

Number of N in N-gram	Precision	Recall	F1-Score
0	0.98	0.99	0.99
1	0.90	0.96	0.93
2	0.92	0.96	0.94
3	0.91	0.94	0.93
4	0.89	0.94	0.91
5	0.86	0.86	0.86

8. FEATURE SELECTION

As mentioned earlier, we have used 100 length word2vec feature vector for learning. As the length is very large, we had to reduce the feature length. For this reason we conducted an ablation study. We have used PCA and *t-sne* for ablation study. We wanted to reduce the dimension from 100 to 20. The classification report is in the

following based on precision, recall and F1-Score. We have trained and tested on same data with new features. The performance matrices are as follows.

Table 7: Performance metrics after *t-sne*

Classifiers	Number of N in N-gram	Precision	Recall	F1-Score With phrase identification
SVM	1	0.83	**0.90**	**0.87**
	2	**0.82**	0.87	**0.85**
	3	0.80	0.88	0.84
	4	0.64	0.80	0.72
	5	0.50	0.72	0.59
Random Forest	1	**0.70**	0.68	0.67
	2	**0.69**	0.73	**0.72**
	3	0.59	0.76	0.67
	4	0.33	**0.81**	0.48
	5	0.22	0.58	0.31
Naïve Bayes	1	0.57	**0.81**	0.68
	2	**0.72**	0.80	**0.77**
	3	0.59	0.80	0.68
	4	0.51	0.81	0.63
	5	0.32	0.68	0.45
Logistic Regression	1	0.64	0.68	0.69
	2	**0.69**	0.72	**0.71**
	3	0.55	0.71	0.62
	4	0.30	**0.71**	0.42
	5	0.20	0.50	0.28
KNN	1	0.81	**0.90**	**0.85**
	2	**0.85**	0.89	**0.87**
	3	0.80	0.84	0.82
	4	0.67	0.77	0.78
	5	0.50	0.70	0.59
LSTM	1	0.81	0.87	0.84
	2	0.85	0.86	0.87
	3	0.79	0.74	0.77
	4	0.65	0.61	0.63
	5	0.50	0.51	0.51

Table 8: Performance metrics after PCA

Classifiers	Number of N in N-gram	Precision	Recall	F1-Score With phrase identification
SVM	1	0.85	**0.91**	**0.88**
	2	**0.84**	0.89	**0.86**
	3	0.81	0.88	0.84
	4	0.65	0.85	0.74
	5	0.50	0.76	0.61
Random Forest	1	**0.71**	0.69	0.68
	2	**0.71**	0.74	**0.73**
	3	0.61	0.78	0.69
	4	0.34	**0.80**	0.48
	5	0.22	0.58	0.31
Naïve Bayes	1	0.59	**0.82**	0.69
	2	**0.74**	0.81	**0.78**
	3	0.61	0.82	0.70
	4	0.51	0.81	0.63
	5	0.35	0.70	0.47
Logistic Regression	1	0.65	0.69	0.70
	2	**0.69**	0.72	**0.71**
	3	0.56	0.72	0.63
	4	0.31	**0.73**	0.44
	5	0.20	0.50	0.28
KNN	1	0.83	**0.91**	**0.86**
	2	**0.87**	0.90	**0.89**
	3	0.82	0.89	0.85
	4	0.70	0.90	0.79
	5	0.51	0.71	0.60
LSTM	1	0.82	0.88	0.85
	2	0.88	0.87	0.88
	3	0.80	0.77	0.78
	4	0.65	0.61	0.63
	5	0.51	0.55	0.53

In Table7, we have shown the performance of the classifiers after applying *t-sne* feature selection technique. After reducing the dimensions, we have seen that the performances were also reduced a bit. It means that we have lost some information after dimensionality reduction. Now, we have explored the performances of the individual classifiers after using PCA selection technique. We also completed a comparative study about PCA and *t-sne*. From Table 7 and Table 8, we can observe that performances have been reduced in both PCA *and t-sne*. However, *t-sne* performed better than PCA.

8.1. Observations

We have previously discussed that the presence of non-medical words also invokes the sense of a medical concept. We have used six machine learning classifiers. These classifiers are not good for phrase identification.

For example: **"World Health Organization"** is a medical phrase. In this phrase, "World" and "Organization" are not medical terms. If we want to label all the terms as medical, our classifiers predict "Indian" and "Association" as a medical term all the time where ever it will occur. For this reason, phrase identification plays a vital role. For understanding more, let's consider two sentences in the following

Sentence 1: *World Health Organization did not recommend Hydroxychloroquine as a medicine of Covid-19.*
Sentence 2: *World is in danger for a disease called Covid-19,*

In the two sentences, the word "world" has been used for two aspects. In first sentence "world" should be classified as medical and in the second sentence it should be classified as non-medical. But our ML classifiers will predict the word "world" as same (medical or non-medical) whenever it occurs. If we follow our method, it will correctly classify the two sentences. In the first sentence there is one phrase whereas in the second sentence, there is no phrase.

In the first pass:

Sentence 1: ***World Health Organization_PHRASE*** *did not recommend Hydroxychloroquine as a medicine of Covid-19.*

Sentence 2: *World is in danger for a disease called Covid-19.*

In the 2nd pass:

Sentence 1: *[World Health Organization]_MC did_O not_O recommend_O Hydroxychloroquine_MC as_O a_O medicine_MC of Covid-19_MC.*

Sentence 2: *World_O is_O in_O danger_O for_O a_O disease_MC called Covid-19_MC.*

In the first sentence, the phrase, "*World Health Organization*" contains a medical term "*Health*", for this reason "*World Health Organization*" becomes a medical term. In the second sentence there is no phrase. In this way we have dealt with two situations using traditional machine learning classifiers.

9. CONCLUSIONS

We have developed a module for concept identification in medical text. We have identified a phrase from a given text using some rules. We have created 6 types of binary classifiers to predict a word (phrase) is medical word (phrase) or not. We have analyzed the performances of these multiple classifiers. In our observation phrase identification module with SVM or K-NN performs better than LSTM. In this way we have shown the importance of phrase identification module. We have applied ensemble (Weighted sum) module for increasing accuracy. After all of these we have built a system which can identify medical concepts from unstructured medical plain text. In future, we are planning to integrate the model with chatbot for medical assistance.

ACKNOWLEDGEMENTS

The work is supported by the project "*Sevak- an Intelligent Indian Language Chatbot*" fundedby DST-SERB, MeiTY, Govt. of India.

REFERENCES

[1] Asghar Muhammad Z., S. Ahmad, M. Qasim, S. Rabail Zahra, and F. Masud Kundi. 2016. SentiHealth: creating health-related sentiment lexicon using hybrid approach. Springer-Plus, 5(1):1139.

[2] Asghar Muhammad Z ., A. Khan, F. M Kundi, M. Qasim, F. Khan, R. Ullah and I. U Nawaz. 2014. Medical opinion lexicon: an incremental model for mining health reviews. International Journal of Academic Research, 6(1):295–302.

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

[4] Donnelly, K. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics, 121:279.

[5] 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, page 371. Association for Computational Linguistics.

[6] Miller, G. and Fellbaum, C. 1998. Word-Net: An electronic lexical database.

[7] Mondal, A., D. Das, E. Cambria and S. Bandyopadhyay. 2016. WME: Sense, Polarity and Affinity based Concept Resource for Medical Events. Proceedings of the Eighth Global WordNet Conference, pages 242–246

[8] Mondal, A., D. Das, E. Cambria and S. Bandyopadhyay. 2018 WME 3.0: An Enhanced and Validated Lexicon of Medical Concepts Proceedings of the 9th Global WordNet Conference (GWC 2018), 10-16