DENGUE DETECTION AND PREDICTION SYSTEM USING DATA MINING WITH FREQUENCY ANALYSIS

Nandini. V[1] and Sriranjitha. R[2] and Yazhini. T. P[3]

Department of Computer Science and Engineering,
SSN College of Engineering, Kalavakkam
[1]nandini.vishwa94@gmail.com
[2]sriranjitha.raghuraman@gmail.com
[3]tp.yazhini@gmail.com

ABSTRACT

Clinical documents are a repository of information about patients' conditions. However, this wealth of data is not properly tapped by the existing analysis tools. Dengue is one of the most widespread water borne diseases known today. Every year, dengue has been threatening lives the world over. Systems already developed have concentrated on extracting disorder mentions using dictionary look-up, or supervised learning methods. This project aims at performing Named Entity Recognition to extract disorder mentions, time expressions and other relevant features from clinical data. These can be used to build a model, which can in turn be used to predict the presence or absence of the disease, dengue. Further, we perform a frequency analysis which correlates the occurrence of dengue and the manifestation of its symptoms over the months. The system produces appreciable accuracy and serves as a valuable tool for medical experts.

KEYWORDS

Named Entity Recognition, Part of Speech tagging, Classification, Prediction, SMO

1. INTRODUCTION

Mining unstructured data is a very pressing issue in the field of text mining. This is especially a major subject in the area of medicine. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge-rich data hidden in the database. Dengue is attracting global concern from researchers and health care professionals over the world. Statistics reveal that almost 25,000 people die from dengue every year. Timely detection of symptoms associated with this deadly disease, and apt prevention measures will go a long way in bringing down its effects on the world populace. Hence, we need a system that will first learn the characteristics of people with dengue, and use this knowledge to predict dengue in new patients. Over the years several NLP systems like cTakes, MetaMap, etc. [3] [2] were used to extract medical concepts from clinical text. They focused on rule based, medical knowledge driven dictionary lookup approaches. While some researchers have contributed to disease prediction,

they have concentrated primarily around heart attacks [6] [7] [10] [11] [12]. Inspiration drawn from such work, combined with the increasing rate of dengue cases around the world motivates us to develop a system to model, predict and analyze dengue instances. The inability to extract useful information from clinical documents may hamper the health care experts' efforts from understanding the relationship between the prevalence of diseases and the associated factors. The frequency of diseases can also be allied with its time frame. This is especially true in the case of water-borne and air-borne diseases. Addressing this task will be a major help to doctors, experts and patients. This relation will enable health care connoisseurs to take preventive measures and reduce the prevalence of these diseases.

2. PROBLEM STATEMENT

The knowledge available in medical repositories is effectively mined and analyzed using the proposed system. The input is a set of annotated discharge summaries containing data pertaining to the disease dengue. Disorder names are extracted from these summaries and looked up in a summarized UMLS (Unified Medical Language System). The output produced in this step is supplied to classifiers which then perform detection and prediction. Further, frequency correlation is performed with the time frame.

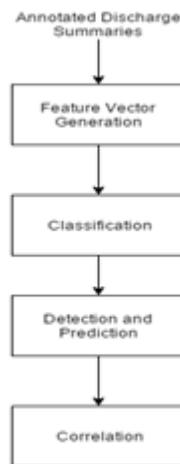## 3. OVERVIEW OF PROPOSED SYSTEM

Figure 1. Overview of the system

The annotated discharge summaries are supplied to feature extraction algorithms and the extracted features are in turn used to generate a feature vector. This is supplied as input to a classification algorithm and a prediction model is developed. The model generated can then be used to detect and predict the presence of dengue. Finally, a correlation analysis is performed to determine how the disease is spread over the months.

4. RELATED WORK

N. Aditya Sundar et al [5] use regular factors contributing to heart diseases, including age, sex, blood sugar and blood pressure, to predict the likelihood of a patient getting a heart disease. Data mining techniques of Naïve Bayesian classification and WAC (Weighted Associative Classifier)

are used to train a model on existing data. Subsequently, patients and nurses can use this model to supply features and get a prediction on a possible heart attack. Oona Frunza et al [6] present a machine learning approach that identifies semantic relations between treatments and diseases and focuses on three semantic relations (prevent, cure and side effect). Later, features were extracted from unstructured clinical text, and were used to classify the relationship between diseases and associated treatments. Jyoti Soni et al [7] have developed a predictive data mining algorithm to predict the presence of heart disease. Fifteen attributes were selected to perform the prediction and Decision Tree was found to produce the best results. Classification based on clustering algorithms was found to not perform well. [12] Proposes a Medical Diagnosis System for predicting the risk of cardiovascular disease. It uses genetic algorithm to determine the weights for a neural network. This feed forward neural network is subsequently used for classification and prediction of heart diseases. A data set of 303 instances of heart disease with 14 attributes each is used for training the system. Devendra Ratnaparkhi, Tushar Mahajan and Vishal Jadhav in their paper [10] describe a system for prediction of heart disease using Naïve Bayes algorithms. They further propose a web interface to help healthcare practitioners assess the possibility of a heart problem in patients. A similar attempt proposes a heart disease prediction system [11] using Decision Tree and Naïve Bayes and its implementation in .NET platform by I.S.Jenzi et al in their paper. Some data mining techniques used for modeling and prediction of dengue include SVM [13], decision tree [14] and neural network [15].

5. SYSTEM DESIGN

The system design is divided into 2 parts.

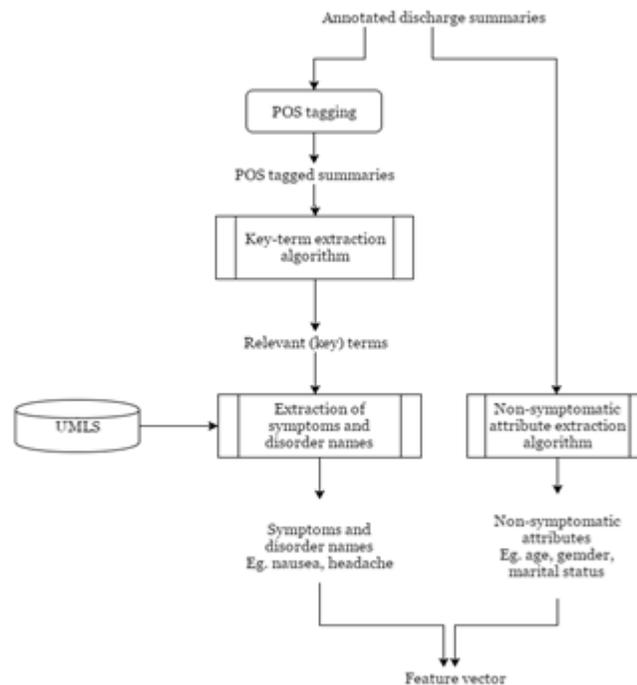### 5.1. Feature Vector Generation

Figure 2. Feature Vector Generation

5.1.1. POS Tagging

The Stanford POS Tagger is used to tag the discharge summaries. An instance of the tagger class is created. The input data is stored in a folder. The program iterates through the folder's files and tags all the input files using the tagger instance created. The tagged data is stored in a file.

```
POSTagger tag = new POSTagger();
File f = new File("DischargeSummary.txt");
while str=f.read()!=NULL do
    tag.tagSample(str);
end
```

Figure 2. POS tagging pseudo code

5.1.2. Key Term Extraction

The key terms such as nouns and adjectives (specified by the tags NN NNP NNS NNPS JJ JJS etc) are extracted from the tagged data and stored in a file.

```
for all word sequences of the form (adjective,noun),(noun),(gerund preposition
    noun) do
    Add Sequence to keyterms file;
end
```

Figure 3. Key term extraction pseudo code

5.1.3. Duplicates Removal

The file generated might contain redundant attributes. To avoid this, the duplicates are removed. The pseudo code for the same is given below

```
Load keyterms file;
Create or open distinctKeyterms file;
Create an empty ArrayList of Strings;
for  each word sequence in keyterms file do
    if sequence is not already in ArrayList then
    |   Add sequence to ArrayList
    end
end
```

Figure 4. Duplicates removal pseudo code

5.1.4. Dictionary Look Up

UMLS (Unified Medical Language System) serves as a repository of mentions. The UMLS is used to extract the relevant symptoms from the tagged file. FileSearcher Class is imported and the FindWordInFile method is used to search for a word in a given file.

```
Load UMLS dictionary;

Load ArrayList of distinct keyterms file;

Create or open distinctFilteredKeyterms file;

Import FileSearcher class ;

Create new FileSeacher object;

for  each keyterm in distinctKeyterms file do
    Call findWordInList(word,file) function of FileSearcher object as

    findWordInList(keyterm,umls) if returned value is true, keyterm is present in

    UMLS then
    |   Add keyterm to distinctFilteredKeyterms file
    else

    end
end
```

Figure 5. Dictionary lookup pseudo code

5.1.5. Temporal Data Extraction

The discharge summaries are fed as input to the temporal data extraction algorithm. The admission months are extracted using regular expressions.

```
String pattern= Expression;

Pattern p = new Pattern(pattern);

Matcher m = p.matcher(Line to be processed);

if m.find() then
    |   return month;
else
    |   return null;
end
```

Figure 6. Temporal data extraction pseudo code

5.1.6. Non-Symptomatic Feature Extraction

Non-Symptomatic features such as age, gender, marital status, family history and past medical history are extracted using regular expressions from the annotated discharge summaries.

- The age is computed using the date of birth of the patient.

- The gender can be either M or F (Male or Female)

- The marital status can be Y or N (Yes or No)

- The family history can be Y or N (Yes or No)

- The past medical history can be Y or N (Yes or No)

- The disease can be Y or N (Yes or No)

5.1.7. Feature Vector Generation

The feature vector is the input supplied to the classifier. The features extracted are combined in a comma separated format and a feature vector is generated. The vector can be represented using frequency value representation or using binary representation. The frequency value format implies that, the frequency of occurrence of the feature in the document is considered. The binary representation on the other hand only considers the presence or absence of the feature in concern. Dengue has very few prominent symptoms and therefore it is not advisable to use the frequency value representation to retrieve them from clinical text. Binary representation of symptoms is therefore preferred in this case. Non-symptomatic features are represented as nominal attributes.

```
Load distinctFilteredKeyterms file;

Create or open featureVector file;

Import FileSearcher class ;

Create new FileSeacher object;

for each file in input data folder do
 |   Add input file name to featureVector file
end

for each keyterm in distinctFilteredKeyterms file do
 |   Call findWordInFile(keyterm,input file) on FileSearcher object
end

if returned value is true then
 |   Print 1 ;
else

end
Print 0 ;

Print newline;
```

Figure 7. Feature Vector Generation pseudo code

The feature vector generated is supplied to a set of classifiers. To identify the best classifier an analysis is performed.

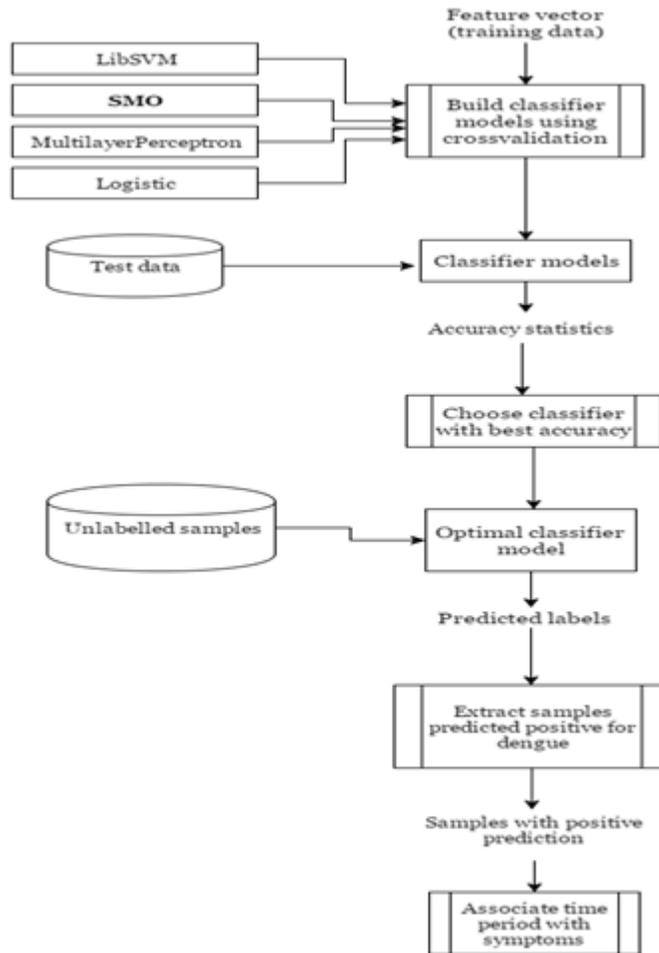## 5.2. Classification and Analysis

Figure 8. Classification and analysis

5.2.1. Classification

The following is the gist of steps followed during classification process:

1. Prepare Training set

2. Supply training set to the classifiers

3. Build the classification models

4. Save the models that have been built

5. Prepare the test set

6. Evaluate the test set on the saved models

7. Analyze the performance of the classifiers

8. Choose the best classifier

9. Supply unlabeled dataset to the best classifier

10. Obtain the prediction

5.2.2. Frequency Analysis

Frequency analysis aims at correlating the frequency of occurrence of the disease over the months. Eight most common and highly contributing symptoms for dengue have been chosen. The occurrences of these symptoms over the months is represented using graphs to give a better understanding of which symptom contributes the most to the presence of dengue.

6. IMPLEMENTATION

6.1. Data set used

We have used 100 samples of annotated discharge summaries as input to this system. The personal details of the patients are already preprocessed to ensure patient confidentiality. They contain details like age, date of birth, date of admission, patient's medical history, medication administered to the patient during the period of stay in the hospital. And the final diagnosis of the patient is also mentioned.

6.2. Tagged file

The above dataset is sent to a POS tagger to perform the part of speech tagging. An instance of the tagger is created and its TagFile method is used to tag the data. This tagged file is sent to a key term extraction algorithm and the relevant features are extracted. The duplicate terms are removed from using the duplicates removal algorithm. These terms are stored in a file.

6.3. UMLS Look up

A subset of the UMLS containing terms relevant to the disease are used as basis to perform the dictionary look up. The file containing the key terms is then compared with the thesaurus and symptoms that contribute to dengue are stored in another file.

6.4. Feature Extraction and Vector Generation

6.4.1. Symptomatic features

To extract the symptomatic features, the following steps are performed:

1. A file reader object is created

2. The discharge summaries are read line by line

- Each line is split into words
- The words are compared with the file containing filtered output

- If there is a match , 1 is written to the feature to the feature vector

3. If there is no match, 0 is written to the vector

6.4.2. Non- Symptomatic features

The non-symptomatic features are extracted using regular expressions. The features are extracted and written to the feature vector file. The feature vector is saved as an arff file.

Snapshot of the generated vector is as shown:

```
@relation DENGUE
@attribute age numeric
@attribute gender(M/F) {M,F}
@attribute ad_mon(01-12) numeric
@attribute marital_stat(M/N) {M,N}
@attribute med_his(Y/N) {Y,N}
@attribute fam_his(Y/N) {Y,N}
@attribute fever(0/1) numeric
@attribute headache(0/1) numeric
@attribute nausea(0/1) numeric
@attribute vomit(0/1) numeric
@attribute Pain_behind_eye(0/1) numeric
@attribute Fatigue(0/1) numeric
@attribute muscle_pain(0/1) numeric
@attribute skin_rash(0/1) numeric
@attribute disease(Y/N) {Y,N}
@data
35, M, 7, M, Y, N, 1, 0, 1, 1, 0, 1, 0, 0, Y
45, M, 8, M, N, N, 1, 1, 0, 0, 0, 0, 1, 1, Y
39, M, 7, N, N, Y, 0, 0, 1, 1, 0, 0, 0, 0, N
57, M, 9, M, N, Y, 1, 1, 0, 0, 0, 0, 0, 1, Y
29, M, 9, M, N, N, 1, 1, 0, 1, 1, 0, 0, 0, Y
22, M, 12, N, Y, N, 1, 1, 0, 1, 0, 0, 1, 1, Y
31, F, 1, N, N, Y, 1, 0, 0, 1, 0, 0, 0, 0, N
37, F, 5, M, Y, Y, 1, 1, 1, 0, 0, 1, 1, 1, Y
78, M, 2, M, Y, N, 1, 1, 0, 1, 1, 0, 0, 1, Y
62, M, 8, N, Y, Y, 1, 0, 0, 0, 0, 1, 1, 1, Y
```

Figure 9. Feature vector

6.5. Classification

The training set is supplied as input to 6 classifiers. Classification analysis was performed on the classifiers. The steps involved in this analysis are:

- Import the weka and java packages
- Call function useClassifier with the data to be classified as parameter
- Create the classifier object
- Build the classifier model
- Save the model
- Create an Evaluation object
- Cross validate using 10 fold cross validation
- Print the confusion matrix

The results of the analysis are discussed in the Results and Discussions section of the paper.

6.5.1. Prediction on Test Set

The test set contains the samples that aren't known to the classification model yet. The saved model is then evaluated on the test set and the accuracy is obtained.

6.5.2. Prediction on Unlabeled Dataset

Unlabeled dataset is fed to the saved model. The disease label is a "?" in this case. The model then predicts the labels for these samples.

6.5.3. Graphical User Interface

A GUI was developed to simplify access to the dengue detection system. Separate panels, one for researchers and another for common users were developed. Researchers can upload a folder consisting of discharge summaries which will be used as the training set. Common users can indicate which symptoms they are experiencing and get a prediction from the system.

Figure 10. Patient GUI

Figure 11. Researcher GUI

6.6. Frequency Analysis

To perform frequency analysis, we have used bar charts. The bar charts are generated using JFreeCharts. The correlations of the spread of the symptoms and in turn the disease over the months are reported briefly to give a clear picture to the researchers. This feature is only available to the researchers.
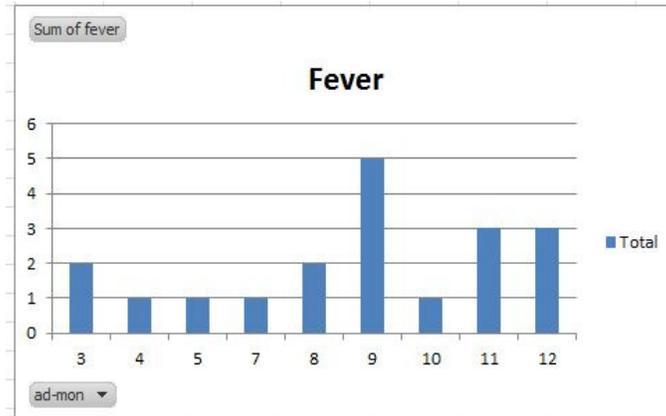

Figure 12. Fever vs month

7. RESULTS AND DISCUSSIONS

The feature vector is supplied to various supervised learning algorithms and classifier models are generated. LibSVM is integrated software for support vector classification, regression and distribution estimation. It supports multi-class classification. Logistic regression classifier uses a sigmoid function to perform the classification. Multilayer perceptron is a classifier based on Artificial Neural Networks. Each layer is completely connected to the next layer in the network. Naïve Bayes methods are a set of supervised learning methods based on applying Bayes theorem with the naïve assumption of independence between every pair of features. The Sequential Minimal Optimizer uses John Plat's sequential minimal optimization algorithm for training a support vector classifier. It also normalizes all attributes by default. The Simple Logistic Classifier is used for building linear logistic regression models. These classifiers are subject to two types of classifications – 10-fold cross-validation and percentage split ($2/3^{rd}$ training and $1/3^{rd}$ test). Accuracies obtained from the 2 methods are compared. In addition, accuracy of the various classifiers are analyzed based on five performance metrics (Accuracy, Kappa statistics, Mean absolute error, Root mean squared error, Relative absolute error) [16] and the best model is chosen.

- Accuracy: The number of samples that are correctly classified from the given 100 input samples.

- Kappa Statistic: The Kappa Statistic can be defined as measuring degree of agreement between two sets of categorized data. Kappa result varies between 0 to 1intervals. Higher the value of Kappa means stronger the agreement/ bonding. If Kappa = 1, then there is perfect agreement. If Kappa = 0, then there is no agreement. If values of Kappa statics are varying in the range of 0.40 to 0.59 considered as moderate, 0.60 to 0.79 considered as substantial, and above 0.80 considered as outstanding.

- <u>Mean Absolute Error:</u> Mean absolute error can be defined as sum of absolute errors divided by number of predictions. It is measure set of predicted value to actual value i.e. how close a predicted model to actual model. The lower the value of MAE the better the classification.

- <u>Root Mean Squared Error :</u> Root mean square error is defined as square root of sum of squares error divided by number of predictions. It is measure the differences between values predicted by a model and the values actually observed. Small value of RMSE means better accuracy of model. Lower the value of RMSE, better the prediction and accuracy.

- <u>Relative Absolute Error:</u> Relative error is the ratio of the absolute error of the measurement to the accepted measurement. A lower percentage indicated better prediction and accuracy.

Classifier Analysis					
Classifier	Accuracy	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
LibSVM	73%	0.0927	0.27	0.5196	68.0175%
Logistic	85%	0.6239	0.1439	0.3735	36.2592%
MultiLayer Perceptron	80%	0.3812	0.2412	0.3806	64.8464%
Naive Bayes'	83%	0.542	0.2261	0.342	56.9706%
SMO	91%	0.7634	0.09	0.3	22.6725%
Simple Logistic	86%	0.6449	0.2004	0.3197	50.4718%

Figure 13. Classifier analysis using 10-fold cross validation

Based on the above analysis, SMO is identified to be the most optimal classifier.

7.1 Analysis and correlation

The predicted results are visualized in graphical form subsequent to prediction. Counts of occurrences of various symptoms over the months are depicted using bar charts, and these values are compared with the graphs generated for the original training dataset. The month with maximum manifestation of all symptoms was found to be September. This was also the month with maximum cases of dengue, according to the prediction. This inference was also corroborated by the graph generated from the initial training dataset, and we gather from these graphs that August, September and October are the months most vulnerable to dengue.
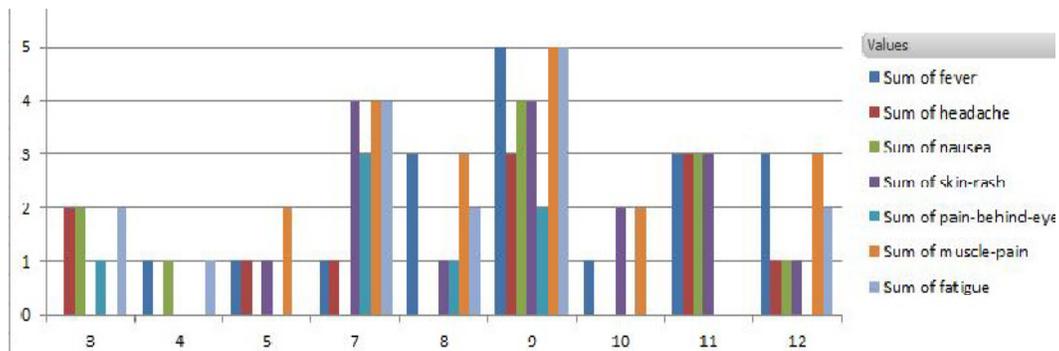
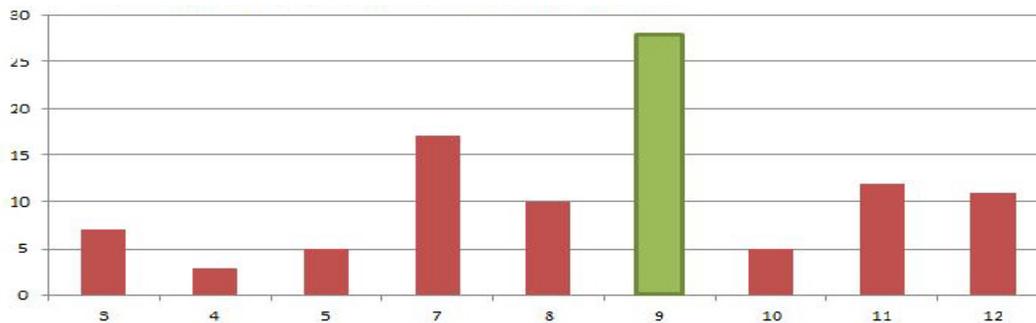Figure 14. Overview of all symptoms spread over the months

Figure 15. Occurrences of all symptoms over the months

8. CONCLUSION

To conclude, we have discussed, in this report, the detailed design and related algorithms for a system to identify disorder mentions from clinical text and correlate its frequency with the time frame. The annotated discharge summaries are tagged and feature extraction algorithms are used to obtain the features relevant to the disease, Dengue. This is followed by the generation of a feature vector (Binary representation). This vector is then used to train and build various classification models and SMO is found to produce the best results. The model generated further aids in the prediction of the disease. Bar graphs are then used to succinctly represent this correlation. Additionally the correlation of training samples with time frame was compared with the correlation obtained from predicted results and the disease occurrence was found to concentrate in the months of August, September and October in both the cases.

9. LIMITATIONS

Our system uses only 15 features. Extracting more features might increase the accuracy of the model. The feature vector is depicted using the binary representation. Using the frequency value representation might improve overall classification.

10. FUTURE WORK

As a part of our future work, we intend to write an implementation to produce bag of words and extract more features to produce an extensive analysis. Further, we also intend to implement

tagging of the discharge summaries using BIOS tagging [5]. Whenever hospitals receive new samples showing a tendency for dengue, those samples must be integrated with the existing training set. This was, the training and predictive capacity of the model will grow, possible giving better results in the future. To provide and up-to date analysis, we could extend the project to be used as a desktop app or browser plugin which will automatically synchronize with new data received from the hospitals' end.

REFERENCES

[1]   Sameer Pradhan, Noemie Elhadad, Wendy Chapman, Suresh Manandhar & Guergana Savova, (July 2014) "Analysis of Clinical Text", SemEval-2014 Task 7.

[2]   Melinda Katona & Richard Farkas, (June 2014) "SZTE-NLP: Clinical Text Analysis with Named Entity Recognition", SemEval-2014.

[3]   Koldo Gojenola, Maite Oronoz, Alicia Perez & Arantza Casillas, (December 2014) "IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts", SemEval-2014.

[4]   Parth Pathak, Pinal Patel, Vishal Panchal, Narayan Choudhary, Amrish Patel & Gautam Joshi, (July 2014) "ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes", SemEval-2014.

[5]   Oana Frunza, Diana Inkpen & Thomas Tran, (June 2011) "A Machine LearningApproach for Identifying Disease-Treatment Relations in Short Texts", IEEE transactions on knowledge and data engineering, vol. 23, Issue no. 6.

[6]   Deepali Chandna, (2014) "Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (2), pp1678-1680.

[7]   Jyoti Soni, Ujma Ansari, Dipesh Sharma & Sunita Soni , (March 2011) "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, vol 17.

[8]   Smitha T & Dr.V Sundaram, (2012) "Knowledge Discovery from Real Time Database using Data Mining Technique", International Journal of Scientific and Research Publications, Volume 2, Issue 4.

[9]   M.A.Nishara Banu & B Gomathy, (Nov-Dec 2013) "Disease Predicting System Using Data Mining Techniques", International Journal of Technical Research and Applications, Volume 1, Issue 5, PP. 41-45

[10]   Devendra Ratnaparkhi, Tushar Mahajan & Vishal Jadhav, (November 2015) "Heart Disease Prediction System Using Data Mining Technique", International Research Journal of Engineering and Technology, Volume: 02 Issue: 08.

[11]   I.S.Jenzi, P.Priyanka & Dr.P.Alli, (March 2013) "A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3.

[12]   N. G. Bhuvaneswari Amma, (February 2012) "Cardiovascular Disease Prediction System using Genetic algorithm and Neural Network", International Conference on Computing, Communication and Applications, IEEE, pp1-5.

[13] A.Shameem Fathima & D.Manimeglai, (March 2012) "Predictive Analysis for the Arbovirus-Dengue using SVM Classification", International Journal of Engineering and Technology, Volume 2 No. 3

[14] Daranee Thitipryoonwongse, Prapat Suriyaphol & Nuanwan Soonthornphisaj, (2012) "A Data Mining Framework for Building Dengue Infection Disease Model", 26thAnnual Conference of the Japanese Society for Artificial Intelligence

[15] N.Subitha & Dr.A.Padmapriya, (August 2013) "Diagnosis for Dengue Fever Using Spatial Data Mining", International Journal of Computer Trends and Technology, Volume 4 Issue 8

[16] Yugal kumar & G. Sahoo, (July 2012) "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA", I.J. Information Technology and Computer Science, Volume 7, pp43-49.

AUTHORS

Nandini V is currently pursuing her final year, Computer Science and Engineering in SSN College of Engineering. She has published a paper on Machine Vision in the ARPN Journal of Engineering and Applied Sciences. Her research interests include Artificial Intelligence, Robotics, Machine Learning, Machine Vision and Data Mining.

Sriranjitha R is currently pursuing her final year, Computer Science and Engineering in SSN College of Engineering. She is a member of CSI (Computer Society of India). Her research interests include Machine Learning, Artificial Intelligence, Data Mining and Data Structures.

Yazhini T P is currently pursuing her final year, Computer Science and Engineering in SSN College of Engineering. Her research interests include Computer Networks, Data Mining, Artificial Intelligence and Web Technology.