A FUZZY LOGIC BASED ON SENTIMENT CLASSIFICATION

J.I.Sheeba[1] and Dr.K.Vivekanandan[2]

[1]Assistant Professor,
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry, India
[2]Professor,
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry, India

ABSTRACT

Sentiment classification aims to detect information such as opinions, explicit , implicit feelings expressed in text. The most existing approaches are able to detect either explicit expressions or implicit expressions of sentiments in the text separately. In this proposed framework it will detect both Implicit and Explicit expressions available in the meeting transcripts. It will classify the Positive, Negative, Neutral words and also identify the topic of the particular meeting transcripts by using fuzzy logic. This paper aims to add some additional features for improving the classification method. The quality of the sentiment classification is improved using proposed fuzzy logic framework .In this fuzzy logic it includes the features like Fuzzy rules and Fuzzy C-means algorithm.The quality of the output is evaluated using the parameters such as precision, recall, f-measure. Here Fuzzy C-means Clustering technique measured in terms of Purity and Entropy. The data set was validated using 10-fold cross validation method and observed 95% confidence interval between the accuracy values .Finally, the proposed fuzzy logic method produced more than 85 % accurate results and error rate is very less compared to existing sentiment classification techniques.

KEYWORDS

Sentiment classification, Fuzzy logic, Fuzzy rules, Fuzzy C-means algorithm, Text mining

1. INTRODUCTION

Now a days, the immense amount of text data is available in the online, text mining has been applied to discover hidden knowledge from the text in many applications and domains. Users express their opinions about TV shows ,Programs ,Products or Services and they consume in blog posts, shopping sites or review sites. For example in the business field, sentiment classification is mainly used to find out consumers, sentiments and opinions about company's products and services and identifying new business opportunities. Manually discovering sentiments and opinions from large volume of data is very difficult and also, another issue of this problem is to extract the expressions like implicit or explicit from large volume of unstructured text data (Ex : Meeting transcripts)[1].So it is needed to detect the expressions automatically from the unstructured text data. Automatic classification of sentiment is important for several applications like credibility analysis of news sites on the Web, survey of marketing report, recommendation system ,mining online discussion, opinion mining and opinion summarization.

Sentiment analysis is one of the latest, highly active fields in Natural Language Processing. It is a task of classifying sentiments according to their transcripts. Sentiment analysis is also known as

DOI : 10.5121/ijdkp.2014.4403

sentiment classification or opinion mining. It is a computational technique which seeks to understand and explain expressions and sentiments by analyzing large amounts unstructured data in such an efficient way as to assist in human decision making. It is essentially useful for all the fields in day to day life. For example In politics, it can predict shift in public opinion regarding election candidates and also it will give the suggestion to select the candidate.

On a daily basis life, it suggests the people to select electronic products like mobile, washing machine, air conditioner and it will also give a better decision to see the movies or books to read[1].Sentiment classification has been modelled as the problem of training a classifier using reviews annotated for Positive, Negative and Neutral sentiment and it will express differently in various domains[2].It is very useful for both the consumers as well as for the producers to know what general public think about a particular product or service .The proposed framework has been introduced for extracting sentiment classification from conversations. So that the users can get the information about the conversation and also they can get the important keywords, topic and expressions. Automatic sentiment classification is the task of classifying given transcripts with respect to the sentiment expressed by the viewers of the conversations.

The most existing approaches are proposed for identifying only explicit expressions of sentiments in the document. Generally, emotions are expressed directly for example "I like to see Sony Television" in the transcripts. Detecting implicit expression from the transcripts is little bit hard, the meanings appeared in the transcripts are indirectly like " This Sony Television is costly but sound system performance is less compared to other Televisions" and the system has to identify these type of words, it would require the training dataset[3].In this proposed approach incorporating features for both implicit and explicit expression detection with in a single framework itself. This paper aims to identify the both expressions and topic of the particular conversations and also to add some additional features for improving the sentiment extraction method . Here, this method is performed by both Human transcripts and ASR transcripts and the sentiments are extracted through MaxEnt and SVM classifier. Finally, the quality of the extracted sentiments are improved using Fuzzy Logic technique.

2. RELATED WORKS

In Sentiment Classification method (Chenghua Lin, Yulan He 2011) proposed unsupervised framework called Joint Sentiment-Topic (JST) model, it works based on Latent Dirichlet Allocation (LDA). In JST model detects sentiment and topic simultaneously from text [4]. In this JST model detects only explicit emotions are presented in the transcripts. It can not detect for implicit emotions and neutral words from the transcripts.In another method(Giacomo Inches,2011) proposed the ideas of author characterization, author identification and also additionally identify the Topic of the particular document, using these methods it is easy to detect the people who are involved in the conversation, find out author character and topic of particular transcripts[5]. In this method detect author's characterization only and it can not support to detect emotions present in the input. In (Yulan He Chenghua Lin,2011)proposed a modified JST model by incorporating word polarity priors through modifying the topic-word Dirichlet priors. Here polarity bearing topics are extracted by modified JST model[6].Another approach (Ahmed abbasi,2011)proposed Feature Relation Network for Sentiment Classification, it will identify N-gram based sentiments only. In this method, two new concepts like subsumption, parallel relation introduced for reducing redundant words[7]. Using this approach, it can not supported for identifying the topic of the particular input In (Alexandra Balahur ,2012) proposed to detect implicit expression from text. In this method it will identify only seven implicit words which have taken from EmotiNet dataset [3]. In this method, it will identify only seven implicit words which have taken from EmotiNet dataset other words, it cannot detect. In another method (J.I.Sheeba and Dr.K.Vivekanandan ,2012) proposed for improved sentiment classification from meeting

transcripts .They incorporate the features for both JST and FRN method and also included for the author classification method [8]. In (Sheeba and Vivekanandan,2012) proposed a another method to combine both keyword extraction and sentiment classification into a single model which will perform both the works at a single time[15].but in this model it cannot detect the implicit words. In (Long Thanh Ngo and Dinh Sinh Mai ,2012) introduced a method to improve the computational efficiency of the interval type-2 Fuzzy C-Means Clustering(IT2-FCM) based on GPU platform and applied it to land-cover classification from multi-spectral satellite image. In this process the performance of GPU is more faster than CPU[16]. In (Nayana Mariya Varghese,2012) proposed a cluster optimization methodology based on fuzzy logic which is used for eliminating the occurrence of redundancies in data after the clustering was completed by the web page mining methods. For Clustering Fuzzy C-Means algorithm is used in this method [17].In (G. Tadayon Tabrizi,2012) proposed a new algorithm that avoids being trapped in local peaks using fuzzy logic and PSO algorithm. It also finds a global optimal response or optimal place of cluster centers[18]. In (Tushar,2009) introduced Fuzzy logic based approach for determining the input-output relationship of some manufacturing processes. Here, fuzzy logic controller is used to determine regression equations and also developed two types of clustering algorithm namely entropy-based approach and fuzzy C-means algorithm[19].In (S.Selva Kumar,2013) introduced a new technique mixed C-means clustering .This method is mainly used to test against a brain tumour gene expression[20]. In (Maciej Piasecki,2013) proposed a new tool WordNetLoom an application for WordNet development it mainly consists of two methods namely a form-based and graph-based and it also discussed the role of the application in WordNet development[21].Based on the above papers inferred fuzzy logic and wordnet techniques and it was adopted for this proposed work.

Generally most existing approaches are detecting either explicit expressions or implicit expressions of sentiments in the text. In the existing method it can identify only seven implicit expressions which was available in the emotinet dataset . But in this proposed framework along with emotinet dataset we can add word net dataset and general words also so that it can able to detect more implicit words compare to existing method. In this proposed method is going to detect both implicit and explicit expressions ,identify topic , author characterization and author identification using fuzzy logic technique likes fuzzy rules and fuzzy C-means algorithm.

3. PROPOSED FRAMEWORK

In this proposed framework the process is done step by step for the best result in extraction of sentiment classification . In this proposed framework along with emotinet data set , the word net dataset has been used. So that it can detect any type of implicit emotions that are available in the transcripts. The data whichever matches with emotions available in the data set will be listed. Based on this list of words, it is concluded whether the Conversation is based on the Positive conversation, Negative conversation or Neutral conversation. Here Convote debate dataset has been taken as input. The data preprocessing has been implemented for this text file and the output will be sent to input to the sentiment classification step. The sentiments are classified both implicit and explicit expression of emotions using training dataset. The output of the explicit and implicit emotions will be sent to the input of the fuzzy logic based sentiment classification for reducing the emotions. Here ,the input will be taken from the output of the sentiment classification step. The weights are assigned for each word in which the weights are calculated from SentiWordNet. Based on the weight of the word, the “Threshold” value is calculated. . The threshold value is calculated based on the average of each listed word. The list of positive, negative and neutral words are listed which is greater than or equal to the Threshold value. This method will returns the reduced emotions list. The output of the fuzzy emotions list will be given to the input of clustering step.

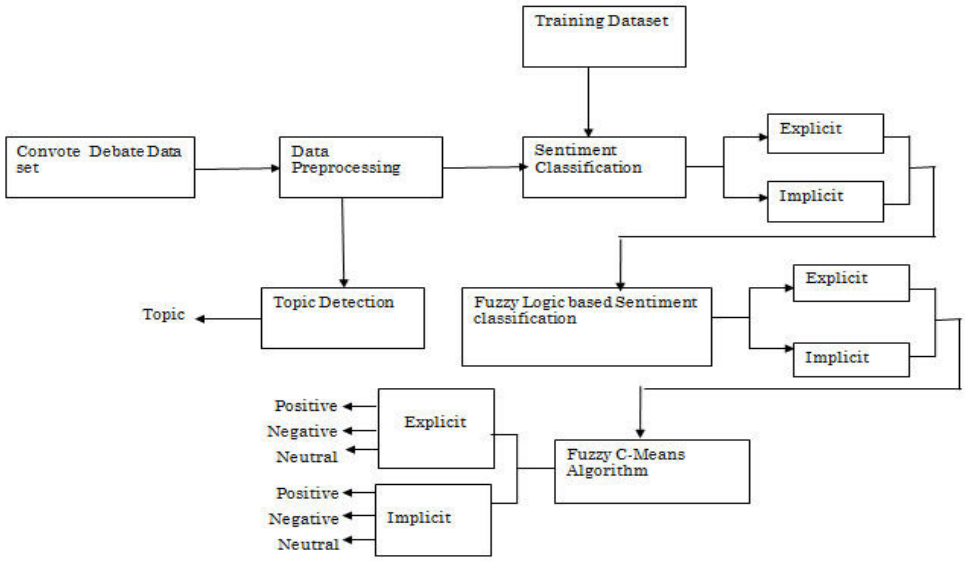

Figure 1: Fuzzy logic based Sentiment classification using proposed framework

Here Fuzzy C-means algorithm used to cluster the similar words for reducing the emotions list and it is able to group the emotions based on the cluster centroid. This framework also identify the topic of the particular inputs. Finally this proposed method will return the reduced and accurate emotions list.

The Figure 1 shows a proposed frame work for Sentiment Classification for given dataset. In this task, 5 Steps have been included

1. Sentiment classification for both implicit and explicit emotions

2. Sentiment classification using Fuzzy logic method

3.To implement for fuzzy C-means algorithm

4. To detect Author classification(Identification, Characterization)

5. To identify the Topic .

3.1 Sentiment classification for both implicit and explicit emotions

Data Preprocessing:

Data preprocessing is mainly used for removing irrelevant and redundant information present or noisy and unreliable data in the inputs. It includes cleaning, selection and feature extraction, normalization, transformation etc. In this method, stop word removal process has been done. Stop words are natural language words which have been filtered out after processing. Some search engines do not record extremely common words in order to save space or to speed up searches. As the list of stop words cannot be removed automatically, it must be trained by human input. Examples of Stop words: A, About, Being, Can, Ever, of, the, You etc.

Sentiment Classification:

In Sentiment classification, the preprocessed data is given as an input to find the list of Positive words, Negative words and Neutral words for both implicit and explicit emotions. The Neutral words are common words. The emotions are detected by comparing each preprocessed data with the list of positive, negative and neutral emotions that are available in the data set for both implicit and explicit emotions .

In this existing work implicit emotions are compared with emotinet dataset, it consists only seven words like "joy, sadness ,fear, anger, shame, disgust and guilt" and it is able to detect and consider these words, except "joy" all the other words are negative related words in which it is only concentrate on the negative part[3]. But in this proposed framework along with emotinet data set , the word net dataset has been used. So that it can detect any type of implicit emotions that are available in the datasets. The data whichever matches with emotions available in the data set will be listed. Based on this list of words, it is concluded whether the debate is based on the Positive conversation, Negative conversation or Neutral conversation.

3.2. Sentiment classification using Fuzzy logic method

Fuzzy Logic:

In this paper it is proposed for a new framework that is capable of answering to the needs of the users in their process of retrieving the desired information when it is enormous ,heterogeneous, vague, imprecise or not in order. Here fuzzy logic is used for retrieving and extracting emotions from transcripts. Fuzzy logic is an ideal tool for the management of this kind of vague and heterogeneous information[14].Fuzzy logic is a form of many-valued logic or probabilistic logic and it deals with reasoning that is approximate rather than fixed and exact. Generally traditional logic they can have varying values between true or false, but in the fuzzy logic variables may have a truth value that ranges in degree between 0 and 1 and it has been extended to handle the concept of partial truth and the truth value may range between completely true and completely false. Here ,the input will be taken from the output of the sentiment classification step. The weights are assigned for each word in which the weights are calculated from SentiWordNet. Based on the weight of the word, the "Threshold" value is calculated. The threshold value is calculated based on the average of each listed word. Finally, the list of positive, negative and neutral words are listed which is greater than or equal to the Threshold value.

3.3. To implement for Fuzzy C-means algorithm

Cluster:

The Fuzzy C-Means algorithm is mainly used for clustering. A cluster brings together instances in the data that share a common set of attributes .For each of these clusters a central value representative of the cluster's principal value is calculated .This is called centriod of the cluster. An array of these centroids for a collection of data elements from a database provides the cluster centers,thus telling us which set of rows in the database are closely related. This algorithm depends on the selection of the initial cluster center and the initial membership value. It involves two processes : the calculation of cluster centers and the assignment of points to these centers using a form of Euclidean distance .This process is repeated until the cluster centers have stabilized .The main goal of this algorithm is the assignment of data points into clusters with varying degrees of membership. This membership reflects the degree to which the point is more representative of one cluster than another[22]. In cluster, the output of the fuzzy logic step acts as the input to this step .Initially, the number of clusters and cluster centroid values are randomly selected. Next, each data element is computed with each cluster centroid to find which data

element belongs to which cluster. After computation, each cluster will have some data elements. Next, the cluster centroid value is updated based on the data elements . This updated cluster centroid value is the final output of the cluster. Here the fuzzy clustering and the data elements can also belong to one or more cluster.Fuzzy C means algorithm works by assigning membership to each data point corresponding to each cluster center ,on the basis of distance between the cluster center and the data point. Most of the data is nearer to the cluster center and its membership is towards the particular cluster center. Summation of membership of each data point should be equal to one. This algorithm is used to cluster the similar words and it is able to group the emotions based on the cluster centroid. Finally , this algorithm gives a reduced and accurate emotions list.

3.4. To Detect Author classification(Identification, Characterization)

Author Classification

Author identification and Author Characterization are called author classification. To identify the author, POS tagger is used here.

POS tagger(Qtag tool)

The tool reads text from datasets and for each token in the text returns the part-of-speech which is used for both stop word removal and author identification.Author identification is used to identify the people who are involved in the conversation. It helps to identify the authors and their participants in the inputs. There are many authors involved in the conversations, here, the author has been identified with the help of POS(Part of Speech) concept.POS tag has been done for the datasets using QTAG tool. It tags the each word's POS. There are some conditions to identify the authors

- The POS of the word must be NP (Proper Noun)
- Author name should be the first word of a sentence.
- Next of the word must be a semicolon (:)

If a particular word satisfies the above mentioned rules, the word will be selected as the name of the author of the conversation. Author characterization is used to identify the characteristics of each author and categorize the people involved in conversation. The sentiment of each author has been identified for detecting the attitude of the author. Depending on the author identification, the conversation is spitted as each author's participation. Then the sentiment will be detected for every author using SVM classifier. Classification concepts are used to group the authors depending on their sentiment[5].In the Figure 2 shows the output of Author identification and characterization from the conversation.

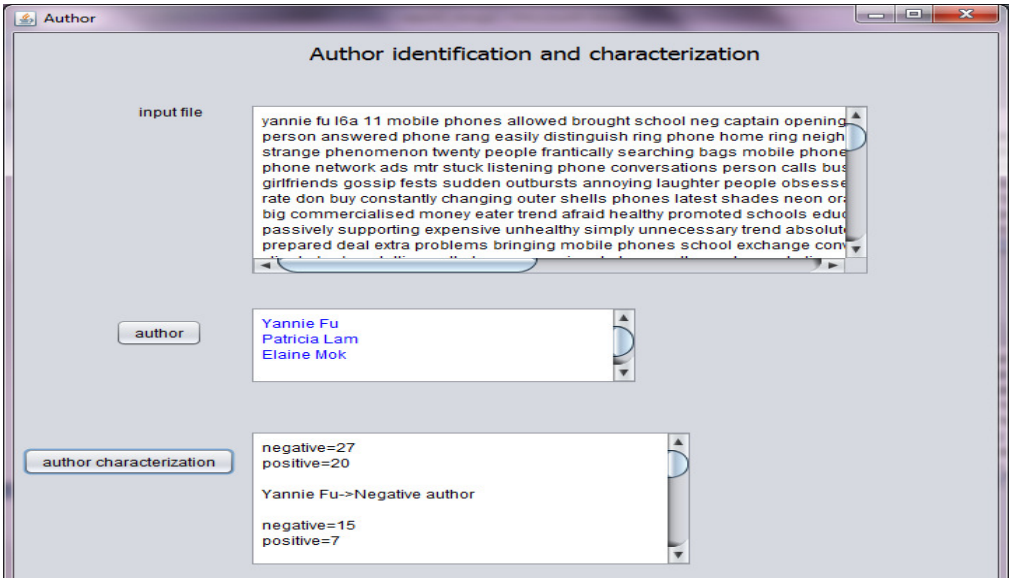

Figure 2:Author identification and characterization from the conversation.

3.5. To identify the topic

Topic Detection:

In Topic Detection, the preprocessed data is given as an input which is used to find the topic, which the debated is based on. In topic detection, each word will assign a weight which is a number of occurrences in the whole conversation. The word which has a maximum weight is selected as a topic for that debate. If two or more words have the same maximum weight, then all the words are listed as a topic separated by numbers which will be useful to select any one.

4. EXPERIMENTAL RESULTS

Data Set

In this paper, the domain meeting transcripts has been focused. Meeting speech is significantly different from written text and other speech data. For example, in meeting transcripts, many people can participate, even the discussions are not organized well and the speech is unplanned one .The speech contains disfluencies and also the sentences are not constructed well .The people who are involved in the meeting ,speak different pronunciations and they use different types of words. Each people can act as different roles and topics in the transcripts. So extracting keywords from meeting transcripts is difficult one compared to the documents[9][10].

Many people can involve in this conversation. The proposed system validated using data set from Convote debate dataset .The text file is similar to meeting transcripts format. In Convote dataset(http://www.Cs.Cornell.Edu/Home/Llee/Papers/Tpl-Convote.Home.Html) It contains 17,940 conversations . Here, randomly conversations have been chosen from the Corpus. By applying this proposed framework sentiment classifications has been done from these Corpus. These inputs are tested by using both Existing systems and Proposed systems. The data sets are tested with the existing systems such as JST , FRN ,ISCF and DIEE Methods and the performance of the proposed system is compared with the existing systems. The results are shown that the fuzzy logic based proposed system provide the results are very accurate one.

Existing Systems

Joint Sentiment-Topic (JST) Model

Joint Sentiment-Topic (JST) model is mainly used to detect sentiment and topic simultaneously from text. It is mainly based on Latent Dirichlet allocation (LDA) method. LDA is a generative probabilistic model of a corpus which is widely used in document analysis. It models the semantic relationships between words based on their co-occurrences in documents[23]. Some sentiment classifications methods often fail to produce satisfactory performance when shifting to other domains but JST makes it highly portable to other domains. The LDA model has three hierarchical layers, where topics are associated with documents, and words are associated with topics. But in the Joint Sentiment-Topic (JST) model, an additional sentiment layer has been added between the document and the topic layer. JST is basically a four-layer model, here sentiment labels are associated with documents and topics are associated with sentiment labels and words are associated with both sentiment labels and topics[4].

Topic Extraction

JST extracts the topics and evaluates the effectiveness of topic sentiment captured by the model. Topics extracted from dataset under positive and negative sentiment labels. Using SVM classifier classifies the dataset and the data set compared with training data set for identifying the topic. Topic identification using SVM classifier is done for the best accuracy from the conversation[4].Figure 3 shows the result of Sentiment classification using JST model.

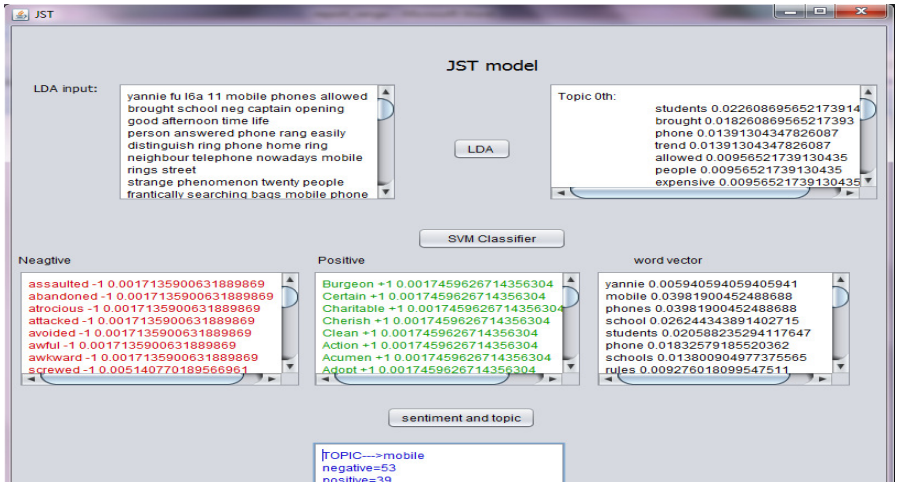

Figure 3:Sentiment Classification using JST Model

Feature Relation Network (FRN)

A Rule-based multivariate text feature selection method is called Feature Relation Network that considers semantic information and also leverages the syntactic relationships between N-gram features. Feature Relation Network is intended for the inclusion of extended sets of heterogeneous N-gram features for enhanced sentiment classification .It includes the following steps for classifying the sentiments like N-Gram Feature generation, Feature extraction, Feature selection and Relation between features[7].Figure 4 shows the result of sentiment classification using FRN Model.

Difference between JST and FRN Model

Using FRN the user can detect sentiment classification only. It can't detect topic of the particular transcripts. It detect the sentiment classification using N-gram based methods .But in JST model the sentiments are detected by using Svm classifier and training dataset.

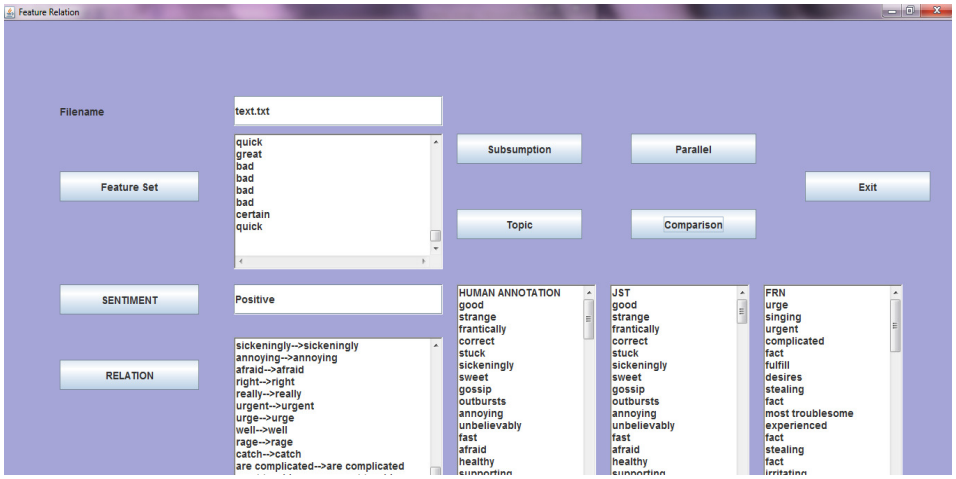

Figure 4 :Sentiment Classification using FRN Model

Improved Sentiment Classification Framework (ISCF)

This framework proposes a method for improved selection of text attributes for enhanced sentiment classification. The result shows whether the given transcript is a positive, negative and neutral oriented topic. Word sense disambiguation is also done which governs the process of identifying sensible words used in a sentence, when the word has multiple meanings. Our proposed framework aims to identify sentiment and topic and also author classification from meeting transcripts. By improving its performance, this framework will incorporate the features of both JST and FRN method . Author classification method mainly used to identify both Author identification and Author characterization from the transcripts and sentiments, topic and author classification are extracted using SVM classifier. This method classifies the author ,the sentiment of the topic and then groups the author based on the result. It benefits to identify the members of the meeting transcripts and their opinion .The results are proved in JST, most of the features match with human annotated than FRN. When comparison is done between JST outperforms and FRN. This operations are performed in a single framework. This method reduces time and space complexity[8].

Detecting Implicit Expression using Emotinet (DIEE)

In this method implicit emotions are detected from text using Emotinet. Emotinet is like a knowledge base which is based on commonsense knowledge stored. It is mainly used to captures and stores emotional reaction to real-world situations. Here commonsense knowledge plays a significant role in the affective interpretation .This method can identify only seven implicit expressions like joy, fear, anger, sadness, disgust, shame, and guilt are available in the emotinet . Among these words, except "joy" all the other words are negative related words[3].

Fuzzy Logic based Sentiment Classification Framework (FL-SCF)

Fuzzy logic can handle the problems with imprecise and incomplete data and it can model nonlinear functions of arbitrary complexity [24].Fuzzy logic technique is introduced to improve the accuracy of the keyword list and also it includes the benefits of simplicity and flexibility. The proposed fuzzy logic method is mainly used to classify the sentiments both implicit and explicit expression of emotion from the Dataset. In this proposed framework along with emotinet , the word net dataset has been used. This method can able to detect any type of implicit emotions that are available in the various datasets compare to existing method. The output of sentiment classification like list of implicit and explicit emotions words weights are identified using Sentiwordnet. The average of the weights are calculated .threshold weight are assigned .The weight which is having greater then or equal to threshold value are considered as output of this method. Again this output will be sent to the clustering step. Here Fuzzy C-means algorithm used to reduce the emotions by applying cluster method. In this cluster method similar words are grouped and it is able to group the emotions using cluster centroid. Based on this list of words, it is concluded whether the input is based on the Positive conversation, Negative conversation or Neutral conversation .It can able to find the accurate implicit and explicit emotions from the sentiment classification .This method is able to find the Topic and also identify both Author identification and Author characterization of the particular inputs. Finally the output will be reduced two times namely fuzzy logic method and fuzzy C-means algorithm method.

Proposed Algorithm for Fuzzy based Sentiment Classification

Input: Convote debate dataset

//Data Preprocessing

1.Read Input Data

2.Collect Datasets D={D1,D2,…..Dn}

3. Remove Stop words from the inputs.

//Sentiment Classification

4. Read Preprocessed input.

5. Foreach (i = 1 to n) (n= number of inputs)

6. Seperate the words based on explicit, implicit and neutral database.

if input match with explicit database words

move into explicit group

else

if input match with implicit database words

move into implicit group

else

if input match with neutral database words

move into neutral group.

End if

End if

End if

7. end Foreach.

8. Repeat steps 5 to 7 for all inputs.

//Fuzzy based Sentiment classification

9. Read the sentiment classification output as input.

10. Each explicit, implicit database words assign weight from 1-10.

11. Foreach (i= 1 to n)

12. Find the average weight.

13. end Foreach.

14. Repeat steps 11 to 13 for each list of words in implicit and explicit.

15.Assign the threshold value and Calculate the words that is equal or greater than threshold for each list.

16.Display the results

//Topic Detection

17. Read Preprocessed input.

18. Foreach (i= 1 to n)

19. Count no of occurrences of each word.

20. end Foreach.

21. Find the word has maximum count.

22. Display the word (topic)

23. If more then one word has the same maximum count then

display word one by one.

end if

// Fuzzy C-means Clustering

24. Read fuzzy output as input.

25. Initialize three clusters namely 4 ,7 ,10

26. Foreach (i=1 to n)

27.Cluster the words based on its associated weight

if the word weight is 5

move it to cluster group 4

end if

28. end Foreach

29.Repeat steps 25 to 28 for all words.

30.Recalculate the clusters name by the list of words it containing.

31.Find the average weight of the list (called centroid)

32.Recalculate the cluster based on centroid value.

33.Repeat steps 30 to 32 for all clusters.

34. Display the result

Output : Reduced the Sentiment classification list

In this algorithm the process is done step by step for the best result in extraction of sentiment classification . Here Convote datasets has been taken as input. Steps 1 to 3 data preprocessing has been implemented for this given text file. In step 4 the data preprocessing output will be sent to input to the sentiment classification .In step6 the sentiments are classified like implicit and explicit expression of positive ,negative and neutral emotions compared with training dataset. In step 9 the output of sentiment classification will be sent to fuzzy based sentiment classification step. In step 10 to 15 the weights are assigned for each word and the weights are calculated from SentiWordNet. Based on the weight of the word, the "Threshold" value is calculated. The threshold value is calculated based on the average of each listed word. Finally, the list of positive, negative and neutral words are listed which is greater than or equal to the Threshold value. In step 17 the pre processed data will be sent to the topic detection step. In step 19 count the number of occurrences of each word in the given input. The word which has a maximum weight is selected as a topic for the given conversation. In step 23 , If two or more words have the same maximum weight, then all the words are listed as a topic separated by numbers which will be useful to select any one. In step 24 the fuzzy based sentiment output will be sent to the fuzzy C-means clustering. In step 25 initialize the three clusters namely 4,7, 10. In step 27 Cluster the words based on its associated weight .for example if the word weight is 5 then the word is move it to cluster group 4.In step 30 assigning membership to each data point corresponding to each cluster center ,on the basis of distance between the cluster center and the data point. Most of the data is nearer to the cluster center and its membership is towards the particular cluster center and finally summation of membership of each data point should be equal

to one. In step 32 cluster the similar words and it is able to group the emotions based on the cluster centroid. Finally this method will given the reduced and accurate emotions list.

Metrics Considered for Evaluation

The performance of the proposed framework is measured in terms of the quality measures namely Precision, Recall and F-Measure.

Precision

Precision is the fraction of retrieved emotions that are relevant

$$\text{Precision} = \frac{\{\text{Number of Relevant emotions}\} \cap \{\text{Number of Retrieved emotions}\}}{\{\text{Number of Retrieved emotions}\}}$$

Recall

Recall is the fraction of the emotions that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{\{\text{Number of Relevant emotions}\} \cap \{\text{Number of Retrieved emotions}\}}{\{\text{Number of emotions Keywords}\}}$$

F-Measure

F-Measure computes both precision and recall as the test to compute the score. Here precision is the number of correct emotions divided by number of all returned emotions. Recall is the number of correct emotions divided by the number of emotions.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

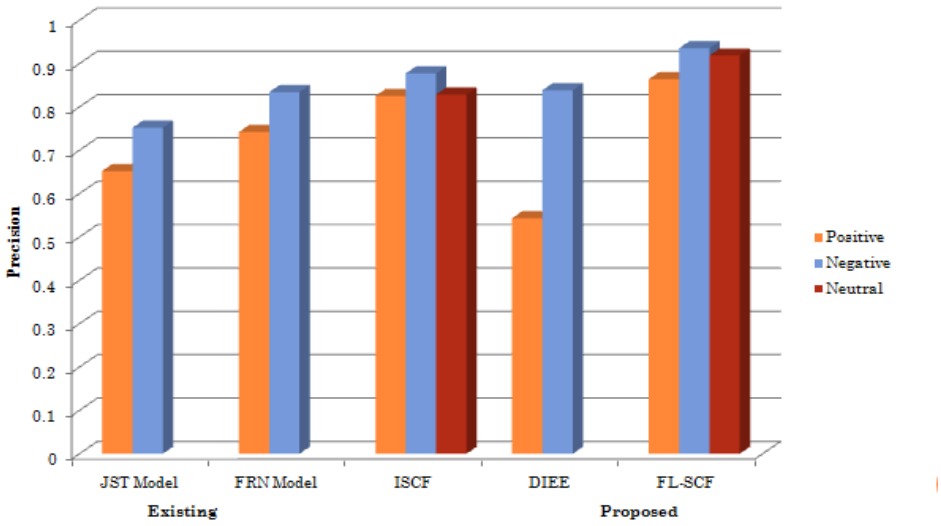

Figure 5. Precision achieved

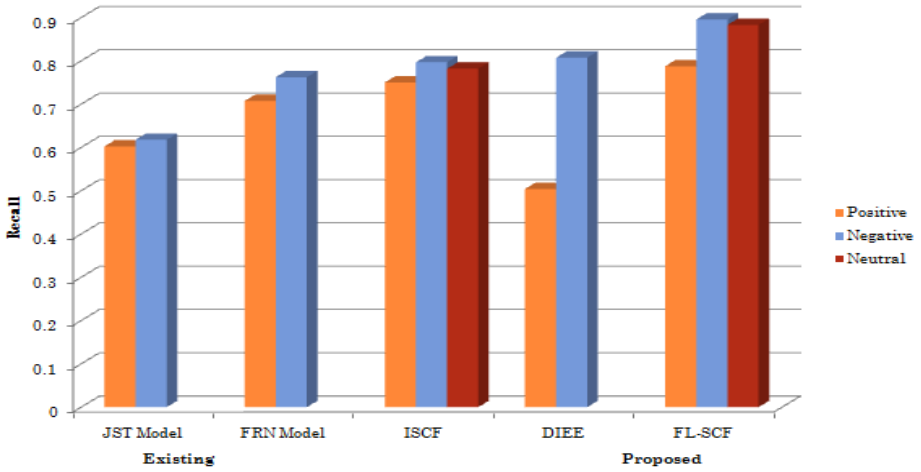

Figure 6. Recall achieved

It is observed from Figure 5 that the FL-SCF achieves the best Precision and from Figure 6 the best Recall .It is also seen that the use of Fuzzy logic improves the performance of all the techniques with SCF performing better than ISCF.

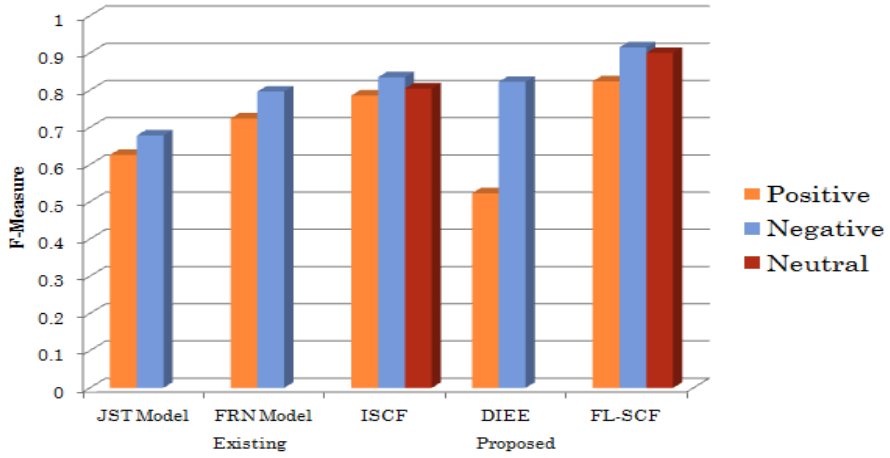

Figure 7. F-Measure

It is observed from Figure 7 that the FL-SCF achieves the best F-Measure. It is also seen that the use of Fuzzy Logic improves performance of all the techniques with FL-SCF performing better than other techniques.Figure 8 shows the Classification Accuracy obtained from the various Techniques.

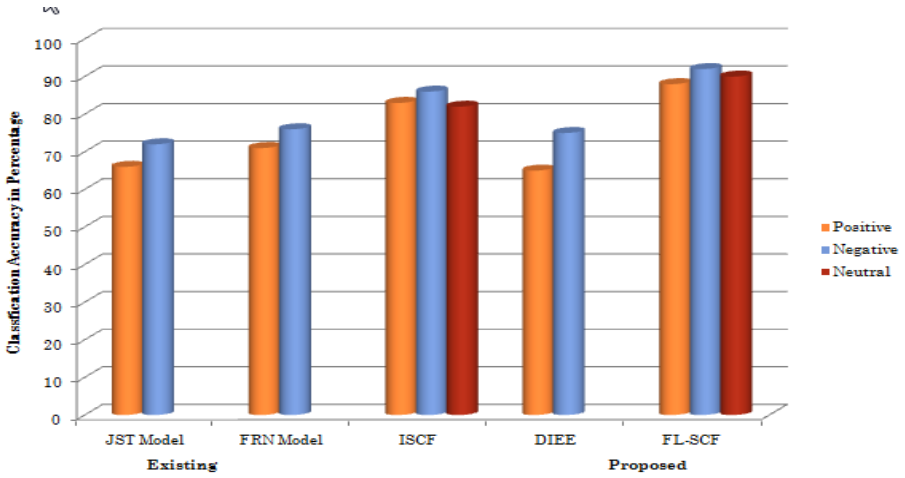

Figure 8. Classification Accuracy

It is observed from Figure 8 that the FL–SCF achieves the best Classification Accuracy .It is also seen that the Fuzzy Logic based SCF method improves the performance compare with ISCF Clustering is a technique of automatically grouping related data into clusters. A wide variety of distance functions and similarity measures have been used for clustering, such as Euclidean distance, cosine similarity, Jaccard distance, Pearson Correlation distance and Kullback –Leibler Divergence. Fuzzy C means algorithm is operated in this process by assigning membership to each data point which is corresponding to each cluster center from the basis of distance between the cluster center and the data point. This proposed framework is used to compare , analyze and evaluate the effectiveness of five similarity measure functions with Fuzzy C means clustering approach. Here purity and entropy validity measures are used to estimate the optimal number of clusters . The results are compared subsequently and shown that Euclidean similarity measure function provides better and faster results as compared to the other distance functions with Fuzzy C means clustering approach .Figure 9 shows the purity performance with different techniques. Figure 10 shows Entropy performance with different techniques.

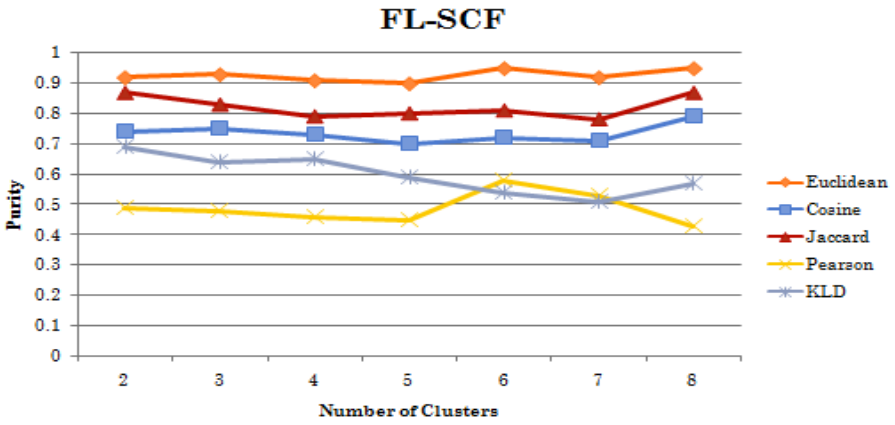

Figure 9: Purity values with different techniques using Convote debate dataset

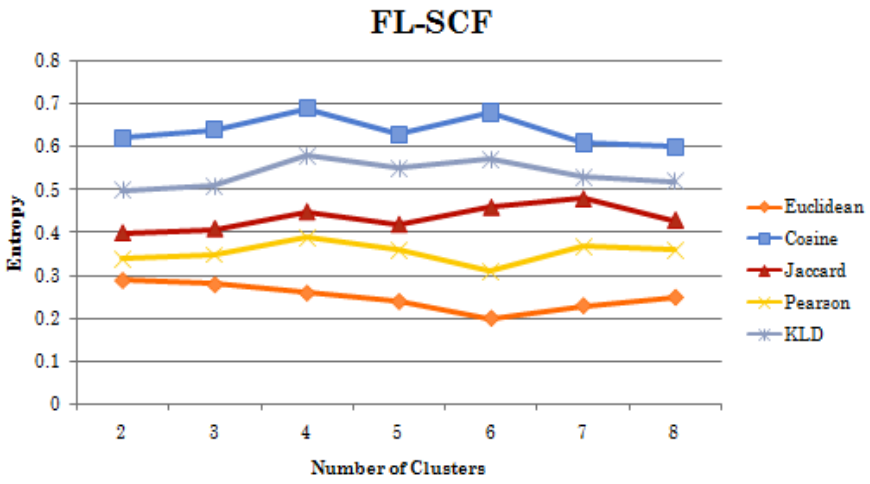

Figure 10: Entropy values with different techniques using Convote debate dataset

Data set validation

The data set was validated using 10 fold cross validation method. Split the data into 10 samples . Fit a model to the training for 9 samples and use the test 1 sample .Repeat the process for the next sample, until all samples have been used to either train or test the model. This table 1 shows the overall mean accuracy values for convote debate dataset. A confidence interval is a range around a measurement that conveys how precise the measurement is .Table 2 shows the 10 fold cross validation performed on 95% confident Interval on convote data set.

Table 1:Mean Accuracy values for different K values

Data set	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Convote debate	86.65	89.43	93.01	94.67	87.63	88.08	89.03	90.43	92.36	93.80

Table 2:10 fold cross validation performed on convote dataset with 95%confidence interval.

Data Sets	T	$Z_{.95}$	Df	Sig.(2 tailed) α	P	Mean	Lower	Upper
Convote debate	28.58	2.262	9	0.025	0.063	92.59	88.59	96.58

In the above table 2.T value represents difference between mean values . $Z_{.95}$ can be found using t-distribution table based on df and α values. where $Z_{.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area. Df gives degree is of freedom. P denotes null hypothesis. lower and upper values denotes 95% confidence

interval between mean accuracy values. The results obtained from this work obviously shows that the proposed algorithms achieving better accuracy when trained with fuzzy logic .

5. CONCLUSION

This paper proposes a method for improved selection of text attributes for enhanced sentiment classification. It shows whether the given conversation is a Positive or Negative or Neutral oriented topic. The proposed framework is able to identify both implicit and explicit expression of emotions and topic detection from the datasets using Fuzzy Logic technique and also applied for the Fuzzy C-means clustering algorithm for clustering the words. Author identification and characterization are also performed. Hence, this method classifies the author and the topic sentiment and then it groups the author. To adopt more accuracy this framework has been implemented. It is performed in a single framework, so it reduces time and space complexity. It benefits to identify the members of the conversation and their opinions. It also enhances the process of classification sentiments from the corpus which are compared to the existing systems. Here Fuzzy logic balances the importance of all the features and also returns the accuracy of the all important emotions from the inputs compared to existing methods. The data set was validated using 10-fold cross validation method and observed 95% confidence interval .The proposed fuzzy logic method produced more than 85 % accurate results and error rate is very less compared to existing sentiment classification techniques.

REFERENCES

[1] Zied Kechaou , Mohamed Ben Ammar, (2011)"Improving elearning with sentiment analysis of users opinions", 2011 IEEE Global Engineering Education Conference (EDUCON) "Learning Environments and Ecosystems in Engineering Education", Amman, Jordan, April 4 - 6, 2010,pp.1032-1038.

[2] Danushka Bollegala, David Weir,(2012) "Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus", IEEE transactions on knowledge and data engineering.

[3] Alexandra Balahur , Jesús M. Hermida,(2012) "Detecting implicit expressions of emotion in text: A comparative analysis", Decision support system ,Elsevier journal, pp.1-12.

[4] Chenghua Lin, Yulan He, Richard Everson,(2012) "Weakly-Supervised Joint Sentiment-Topic Detection from Text," IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 6, June 2012, pp .1134-1145.

[5] Giacomo Inches, Fabio Crestani, (2011)"Online Conversation Mining for Author Characterization and Topic Identification" ,PIKM'11, October 28,ACM, pp. 19-26.

[6] Yulan He Chenghua Lin, Harith Alani,(2011) "Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification". Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, June 19-24, pp. 123–131.

[7] Ahmed Abbasi, Stephen France, Zhu Zhang, Hsinchun Chen,(2011) "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," IEEE Transactions on Knowledge and Data Engineering, Mar. 2011,vol. 23, no.3, pp. 44-42.

[8] J.I.Sheeba , Dr.K.Vivekanandan,(2012) "Improved Sentiment Classification From Meeting Transcripts", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[9] Feifan Liu, Deana Pennell, Fei Liu,(2009) "Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts", ACM .

[10] Fei Liu, Feifan Liu,(2011) "A Supervised Framework for Keyword Extraction From Meeting Transcripts", IEEE Transactions On Audio, Speech, And Language Processing, vol. 19, No. 3, March 2011,pp.538-548.

[11] http://www.support-vector.net.

[12] Veselin Stoyanov , Claire Cardie,(2008) "Topic Identification for Fine-Grained Opinion Analysis", International Conference on Computational Linguistics (Coling 2008), Manchester, August 2008, pp. 817–824.

[13] Eibe Frank ,Gordon W.Paynter, "Domain specific Keyphrase extraction".

[14] Jorge Ropero, Ariel Gomez (2012)" A Fuzzy logic intelligent agent for information extraction introducing a new Fuzzy logic-based term weighting scheme", Expert systems with applications39(2012),Elsevier journal, pp. 4567-4581

[15] J. I. Sheeba , K. Vivekanandan,(2012) "Low Frequency Keyword and Keyphrase Extraction from Meeting Transcripts with Sentiment Classification using Unsupervised Framework", CCSEIT-2012, October 26~28, Coimbatore, Tamilnadu, India. ACM .pp.212-216, ISBN 978-1-4503-1310-0.

[16] Long Thanh Ngo and Dinh Sinh Mai,(2012) "GPU-based Acceleration of Interval Type-2 Fuzzy C-Means Clustering for Satellite Imagery Land-Cover Classification", 12th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE, pp. 992-997.

[17] Nayana Mariya Varghese and Jomina John,(2012) " Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", World Congress on Information and Communication Technologies,IEEE, pp.948-952

[18] M. Mir and G. Tadayon Tabrizi ,(2012) "Improving Data Clustering Using Fuzzy Logic and PSO Algorithm", 20th Iranian Conference on Electrical Engineering, (ICEE2012), May 15-17,Tehran, Iran. IEEE, pp .784-788

[19] Tushar,Dilip Kumar Pratihar,(2009)"Design of cluster-wise optimal fuzzy logic controller to model input-output relationships of some manufacturing processes", Int J.of Data Mining, Modelling and Management,vol.1,no2 pp.178-205,DOI :10.1504/IJDMMM 2009.026075.

[20] S. Selva Kumar; H. Hannah Inbarani ,(2013)"Analysis of mixed C-means clustering approach for brain tumour gene expression data", Int. J. of Data Analysis Techniques and Strategies,Vol.5, No.2, pp.214 – 228, DOI: 10.1504/IJDATS.2013.053682.

[21] Maciej Piasecki; Michał Marcińczuk; Radosław Ramocki; Marek Maziarz,(2013) "WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives", Int. J. of Data Mining, Modelling and Management, Vol.5, No.3, pp.210 – 232, DOI: 10.1504 / IJDMMM. 2013. 055861.

[22] Earl cox(2005)"Fuzzy modeling and Genetic algorithms for data mining and exploration", Published by Elsevier, Morgan Kaufmann Publishers. ISBN No :0-12-194275-9.

[23] Blei ,D.M.,NG, A.Y., And Jordan ,M.I,(2003) Latent Drichlet Allocation Math.Learn.Res.3,pp.993-1022.

[24] Innocent P.R and John, R.I.(2004), "Computer Aided Fuzzy Medical Diagnosis", Information sciences ,Vol .162,No.2,pp.81-104.

BIBLIOGRAPHY

J.I.Sheeba received her B.E in Computer Science and Engineering from Bharathidasan University and M.E in Computer Science and Engineering from Anna University. She is currently pursuing her Ph.D in Computer Science and Engineering, from Pondicherry Engineering College affiliated to Pondicherry University. Presently she is working as Assistant Professor in Department of Computer Science and Engineering, Pondicherry Engineering College. Her research interest includes Data mining and Fuzzy Logic.

Dr.K.Vivekanandan received his B.E from Bharathiyar University, M.Tech from Indian Institute of Technology, Bombay and Ph.D from Pondicherry University. He has been the faculty of Department of Computer Science and Engineering, Pondicherry Engineering College from 1992. Presently he is working as Professor in the Department of Computer Science and Engineering. His research interest includes Software Engineering, Object Oriented Systems, Information Security and Web Services. He has coordinated two AICTE sponsored RPS projects on "Developing Product Line Architecture and Components for e-Governance Applications of Indian Context" and "Development of a framework for designing WDM Optical Network".