Suitability of Naïve Bayesian Methods for Paragraph Level Text Classification in the Kannada Language using Dimensionality Reduction Technique

Jayashree R[1], Srikantamurthy K[1] and Basavaraj S Anami[2]

[1]Department of Computer Science and Engineering, PES Institute of Technology, Bangalore,India
[2]Department of Computer Science and Engineering, KLE Institute of Technology, Hubli,India

ABSTRACT

The amount of data present online is growing very rapidly, hence a need for organizing and categorizing data has become an obvious need. The Information Retrieval (IR) techniques act as an aid in assisting users in obtaining relevant information. IR in the Indian context is very relevant as there are several blogs, news publications in Indian languages present online. This work looks at the suitability of Naïve Bayesian methods for paragraph level text classification in the Kannada language. The Naïve Bayesian methods are the most primitive algorithms for Text Categorization tasks. We apply dimensionality reduction technique using Minimum term frequency, stop word identification and elimination methods for achieving the task. It is evident that Naïve Bayesian Multinomial model outperforms simple Naïve Bayesian approach in paragraph classification tasks.

KEYWORDS

performance,classifier,paragraph level classification, text classification,Naïve Bayesian,Multinomial .

1.INTRODUCTION

Text Classification (also known as Document Classification, Text Categorization) is a common problem studied in Natural Language Processing. The idea is to identify a *class* or a group to which a particular "*document*" belongs to. The *supervised* form of this task requires the availability of "*training data*" – the data that can "train" a classifier to correctly identify the class of an unknown text – which is called the "*test data*". There are two variations in Text Classification: Single Label, binary and Multi label. If there are two classes involved, then it is called binary classification, if a single class is involved, then it is single label classification; if there are more classes involved then it is multi label classification.

Mathematically, Text categorization is the task of assigning a Boolean value to each pair $(d_j, c_i) \in D \times C$, where D is a domain of documents and $C = \{c_1 , \ldots , c_{|C|}\}$ is a set of predefined categories. A value of T assigned to(d_j, c_i) indicates a decision to file d_j under c_i , while a value of F indicates a decision not to file d_j under c_i .

Paragraph level classification has gained due importance in many real world applications such as Text Document summarization, on line Essay scoring, sentiment analysis, etc, as information present online is mostly in the form of paragraphs. There is also a need to update these data regularly. The problem of finding the correct location to insert new updated information in a hierarchical text is quite a challenging task. Normally, the structure of a given document is

DOI : 10.5121/ijaia.2013.4509

hierarchically divided into sections, paragraphs and sentences. Hence finding the correct location to insert updated information takes due importance which can be done using paragraph classification. As mentioned in the Literature, Topic shift detection, discourse parsing, text segmentation etc may be treated as paragraph classification tasks. Essay assessing application requires processing text within paragraph, passage or whole document.

With this background, we have made a new attempt to analyze how paragraph level text classification works for the Kannada language using Naïve Bayesian and Naïve Bayesian Multinomial methods.

The rest of the paper is organized as follows. Section- 2 highlights the literature carried out in paragraph classification in particular and Text categorization in general. Section-2.2 describes how the corpus was prepared for use in this work. Section-3 discusses the methodology of our work. Section –4 is about Results and Discussion.

2. LITERATURE SURVEY

Text Classification is an important Natural Language Processing(NLP) application. Jayashree.R Et.al(2011) have investigated two classical approaches such as Naïve Bayesian and Bag of Words to Sentence Level Text Classification in the Kannada Language and looked at the possibility of extending sentence level classification task to Paragraph Level Text Classification in their future work.

Erdong Chen Et.al(2007) in their work on 'Incremental Text structuring with on line hierarchical ranking' work on an online ranking model which exploits the hierarchical structure of a given document. The testing is performed on a corpus of Wikipedia articles. They also present a sentence insertion model which determines the best paragraph within which to place a sentence.

Alex Smola Et al(2007) worked on Semi Markov model using Max Margin method.
In speech to text applications, the output is text, is usually a raw text without any punctuation or paragraph markings, such texts requires paragraph segmentation.

'Modelling organization in student Essays' carried out by Isaac Persing Et.al(2010) discusses the structure of the essay. The authors develop computational model for the organization of the student essays. They adopt heuristic approach to paragraph function labelling.

'TextTiling:segmenting text into multi-paragraph sub topic passages is a novel contribution of Marti.A.Hearst Et al(1997).TextTiling is an approach for subdividing text into multi-paragraph units that represent passages, paragraphs or subtopics. This basically plays an important role in many Information Retrieval and Text Summarization tasks.

'Genre Based Paragraph for Sentiment Analysis' classification system for differentiating different paragraphs within movie reviews is the work carried out by Maite Taboada Et.al (2009).

We also had to do survey of Naïve Bayesian Multinomial models in literature since the suitability of this model to paragraph level classification was considered.

Work by Andrew Mccallum Et.al (2007) makes an attempt to clarify the confusion between Naïve Bayesian models; Multi variant Bernoulli model and multinomial model. They claim that multinomial model is better than the multi variant Bernoulli model.

Multinomial model for Text Categorization is a novel work of Ashraf.M.Kibriya Et.al(2004).They discuss about transformations from Multinomial Naïve Bayes to Transformed Weight Normalized Complement Naïve Bayes.

Reversing and smoothing the Multinomial Naïve Bayes Text classifier is the work of Alfons Juan Et.al(2001). The paper highlighted the effects of parameter smoothing and document length normalization.

Jason.D.M.Rennie Et.al(2003) proposes heuristic solutions to some of the problems with Naïve Bayes classifiers. They first review Multinomial Naïve Bayes model for classification and discuss several systemic problems with it.

Applications like sentiment analysis, student online essay scoring, ext summarization etc are predominant for the present information era. This prompted us to look at paragraph level text classification from the perspective of the classifiers.

2.1 CORPUS

A custom built corpus called TDIL(Technology for Development of Indian Languages) is a comprehensive Kannada text resource, which is developed by Central Institute of Indian Languages (CIIL). The text resource is manually categorized which makes it a ready source of data for this problem. Four classes of data collected, namely: Commerce, Social Sciences, Natural Sciences and Aesthetics. Cleaning up the data involved sentence separation and removing headings in the document.

The training data comprised of ~80% of the sentences in each class, and the rest of the sentences were set aside for test.

A total of 1791 paragraphs belonging to different category documents were used for the classification process. The table below shows the class wise distribution of paragraphs.

Table 1.1 Class wise Distribution of paragraphs in TDIL corpus

Category	No. of Paragraphs
Commerce	476
Social	413
Natural	475
Aesthetics	427

2.2. METHODOLOGY

Naïve Bayes Multinomial

Presence or absence of words is an important point to be considered while characterizing documents. Treating this parameter as Boolean attribute helps in applying machine learning to paragraph classification. Naïve Baysian does not consider the number of occurrences of each word, which is potential information in determining the category of the document. Naïve Bayesian views document as bag of words with multiple occurrences of a word appearing multiple times. Hence word frequencies can be accommodated by applying a modified form of Naïve Bayes which is Naïve Bayes Multinomial.

If a document E belongs to category H , and $n_1,n_2,n_3…..n_k$ is the number of times word 'i' occurs in the document and P_1,P_2,P_3 …P_k is the probability of obtaining word 'i'. the probability of a document E given its class H is,

PR[E|H]≈N!X

$$Pr[E|H] \approx N!X \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i!}$$

……..…..(1)

Where $N = n_1+n_2+n_3……n_k$

Is the number of words in the document.

Dimensionality Reduction :

We have tried to look at the performance of the classifier using dimensionality reduction technique. This is because the size of the feature set has a significant impact on the time required for classification. The morphological richness of the Kannada language has led to the feature dimensions to be in the order of tens of thousands. For practical classification considerations, large amounts of training samples are required to train the classifier.

1)Using stopwords

The first approach for dimensionality reduction is identifying and eliminating stop words. Stop words do not hold any information about the class of the text. The function words of a language are usually identified as stop words. These words are considered as noise and are removed before the classification process. There is no standard stop word list available in the Kannada language. Hence, we sought the help of subject expert in Kannada for identifying stop words in our corpus manually.

The words were manually examined and following set of stop words was created for use in the classification process. Below is the list of stop words and their meaning in English.

Table 1.2 List of Stop words

ಈ (This)	ಮತ್ತು (And)	ಹಾಗೂ (And)	ಎಲ್ಲಾ (All)	ಬಂದ (Came)
ಎಂಬ (Called)	ಅವರ (Their)	ಎಂದು (Known)	ಹಾಗು (And)	ಹೇಗೆ (How)
ತಮ್ಮ (Yours)	ಇವರು (These)	ಯಾವ (Which)	ಇವರ (These)	ಅದೇ (That)
ಇದು (This)	ಅವರು (Those)	ಅಥವಾ (Or)	ಆದರೆ (But)	ಹೀಗೆ (How)
ಈಗ (Now)	ಎಂಬ (Called)	ಇದನ್ನು (This)	ಇದರ (This)	ಆಗದೆ (Not possible)

ಎಲ್ಲ (All)	ಅದು (That)	ಇನ್ನೂ (Even)	ಅವರಿಗೆ (Them)	ಏನೂ (Any Thing)
ಬಗೆ (Type)	ಎಲ್ಲರೂ (All)	ಅಥವ (And)	ಇಲ್ಲವೇ (No)	ಯವರ (Their)
ಆದ (Cause)	ಅದನ್ನು (That)	ಇಂದು (Today)	ಹೋಗಿ (Go)	ಆವರ (Their)
ಅಲ್ಲ (No)	ಇದೇ (This)	ಅವನು (Him)	ಅದರ (Its)	ಅವನಿಗೆ (His)
ನಾವು (Our)	ನಮ್ಮ (Ours)	ನನ್ನ (Mine)	ಇಂದ (From)	ಎನ್ನುವ (Called)
ಎಶ್ಟು (How much)	ಇದಕ್ಕೆ (For This)	ಇವು (These)	ಈಗಿನ (Present)	ಈಗಲೂ (Even now)
ಇಲ್ಲ (No)	ತಾನು (That)	ಆಗಾಗ (Often)	ಆತನ (His)	ಹಾಗೆಯೇ (This way)
ಎಲ್ಲಿ (Where)	ತನಗೆ (Self)	ಇದ್ದ (Present)	ಎರಡು (Two)	ಯಾವುದೇ (Any)
ಇತ್ತು (Present)	ಬಂದು (Come)	ಆದರ (Its)	ಅಂದರೆ (Called)	ಯಾಗುವ (Cause)
ಅಲ್ಲಿ (There)	ಇದರಿಂದ (Hence)	ನಿಮ್ಮ (Yours)	ಹಾಗಗಿ (Hence)	ಎಂಬುದು (Called)
ಹೀಗೆ (This Way)	ಇವರಿಗೆ (These)	ನಾನು (Me)	ಅಲ್ಲಿಂದ (From)	ಇವೆಂದರೆ (These)
ಇದೆ (There)	ಅಲ್ಲಿನ (There)	ನನಗೆ (Me)	ಆಗಿನ (Then)	ಇವೆಲ್ಲವೂ (All These)
ತಾವು (You)	ಅವರೇ (Those)	ಅವನ (His)	ಅದಕ್ಕೆ (For that)	ಎಲ್ಲವನ್ನೂ (Everything)
ತಾನೆ (Self)	ಎಂದು (When)	ಅವನ್ನು (Those)	ನನ್ನನ್ನು (Me)	ಅದರಿಂದ (For that)
ಇವೂ (These)	ಅಂಥ (Such)	ಅದಕ್ಕಾಗಿ(For that)	ಈತನ (His)	ಏಕೆಂದರೆ (Because)
ಅಂದು (Then)	ಇರುತ್ತದೆ (Present)	ಇದ್ದರೆ (Present)	ಇವಳಿಗೆ (For her)	ಆದುದರಿಂದ (Hence)

2) Using a restriction based on word occurrence

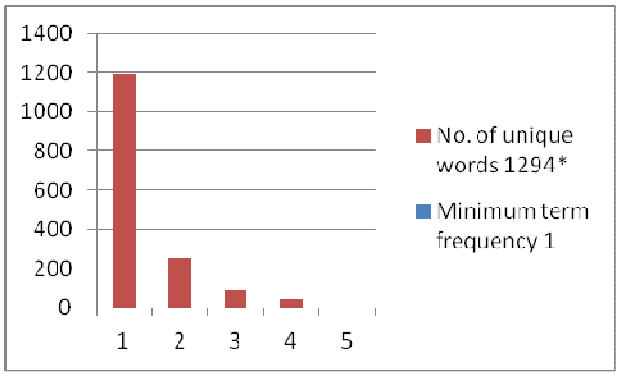
Removing words that occur in the data base only once would help in improving the performance of the classifier, as the time taken for classification is reduced. But it is evident from our results that, words which occur once may definitely matter to the document in deciding the category of

the document, the amount of words that occur once is very large, which add unrealistic requirement to the training data.

In this work, the behaviour of the classifier for varying values of the minimum word occurrence requirement (m) is analyzed; the values being 2 to 5 and the evaluation measure for m = 1 is derived by fitting it on the curve using least squares method. The points are plotted on the graph in an excel sheet and a best fit curve is derived which makes it possible to interpolate and infer co-ordinates for other points .Figure 1.1 shows that for m=6 or above the number of terms is too low to be considered.

Table 1.3 Word Occurrence and the Number of Unique words from CIIL corpus

Minimum term frequency	No. of unique words
1	1294*
2	1189
3	251
4	86
5	41



.

Figure 1.1 graph of Minimum Term frequency

Naïve Bayesian model

A word vector is created based on the training data. A Naive Bayesian classifier was used as an alternative approach to Bag of Words approach. The dimensions in the vector indicated the presence of the word and no special weight age parameter was used in classification.

The word occurrence probability is given by the relation

$$W_k = P(W_k|C_j) = \frac{1+\sum_{i=1}^{|D|} N(W_k,d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N(W_s,d_i)} \quad \ldots\ldots\ldots(2)$$

Where W_k is the probability vector of the characteristic word that belongs to every class .

To find the class that the text belongs to, the following relation must be maximized:

$$P(C_j|d_i;\Theta)^n = \frac{P(C_j|\Theta)\prod_{k=1}^{n} P(W_k|C_j;\Theta)^{N(W_k,d_i)}}{\sum_{r=1}^{|C|} P(C_k|\Theta)\prod_{k=1}^{n} P(W_k|C_r;\Theta)^{N(W_k,d_i)}} \quad \ldots\ldots(3)$$

$P(C_j|d_i;\Theta)^n$ is the probability of text d_i belonging to class C_j. $P(C_k|\Theta)$ is resemble meanings, $|C|$ is the sum of the class, $N(W_k, d_i)$ is the word frequency of W_k in d_i, n is the account of characteristic words.

IV. METHODOLOGY

K-fold Cross Validation is used in this work for evaluation of the classifier performance. This technique involves splitting the document into K disjoint partitions and carrying out K rounds of testing with one of the partitions as the test set and the remaining as training. It is ensured that each partition is used as a test set only once

The parameters used to evaluate the classifier performance are: Precision (P), Recall (R)(also called as TP rate) and F-Score (F). The definitions of the parameters are as shown:

Precision

$$= \frac{\text{Proportion of the examples which truly have class } x}{\text{Total classified as class } x} \quad \text{.....(1)}$$

TP rate/True Positive(TP)

$$= \frac{\text{Proportion classified as class } x}{\text{Actual total of class } x} \quad \text{.....(2)}$$

$$\text{False Positive (FP)} = \frac{\text{Proportion incorrectly classified as class } x}{\text{Actual total of all classes, except } x} \quad \text{.....(3)}$$

$$F - \text{measure} = \frac{2 \times \text{Precison} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \quad \text{.....(4)}$$

VI. RESULTS AND DISCUSSION

The two models developed are evaluated against the test set using 10-fold cross (k= 10) Validation and the results are as shown. We have also compared the results obtained with that of the sentence level text classification.

Naive Bayes

Table 1.4 Weighted Averages for precision, recall and F-scores with stop words

	Precision	Recall	F-Score
m=1*	0.867	0.883	0.888
m=2	0.803	0.799	0.8
m=3	0.735	0.717	0.719
m=4	0.661	0.601	0.613
m=5	0.607	0.546	0.542

* interpolated

Table 1.5 Weighted Averages for precision, recall and F-scores without stop words

	Precision	Recall	F-Score
m=1*	0.803	0.859	0.863
m=2	0.785	0.773	0.774
m=3	0.734	0.681	0.685
m=4	0.715	0.583	0.588
m=5	0.698	0.508	0.507

* interpolated

With decreasing 'm', the evaluation parameters show a significant rise, which shows the effect of the words that have low occurrence, have on classification. To estimate the impact if single-occurrence words were included, the values from the table are used and the corresponding values for m = 1 are derived using best-fit regression.

Taking m=2, the class-wise breakup for the classification results is as shown:

Table 1.6 classification results

	TP rate	FP rate	Precision
Commerce	0.811	0.027	0.915
Social Sciences	0.661	0.05	0.798
Natural Sciences	0.773	0.09	0.757
Aesthetics	0.838	0.135	0.661

Table 1.7 Confusion Matrix for Naïve Bayesian Approach

a	b	c	d	← classified as
386	28	36	26	a = Commerce
11	273	46	83	b = Social Sciences
16	17	367	75	c = Natural Sciences
9	24	36	358	d =Aesthetics

Error Analysis here indicates confusion between Natural Sciences and Aesthetics classes and Commerce and Natural Sciences classes.

The Aesthetics class seems to have low precision and high recall. Social Sciences class has high precision but low recall. Natural Sciences class has equal precision and recall values, which is an interesting pointer for further research.

Naïve Bayesian Multinomial

Table 1.8 Weighted Averages for precision, recall and F-scores with stop words using Naïve Bayesian multinomial:

	Precision	Recall	F-Score
m=1*	0.983	1.005	1.014
m=2	0.877	0.874	0.874
m=3	0.75	0.736	0.736
m=4	0.594	0.559	0.554
m=5	0.537	0.474	0.455

Table 1.9 Weighted Averages for precision, recall and F-scores without stop words using Naïve Bayesian multinomial:

	Precision	Recall	F-Score
m=1*	1.11	0.970	0.985
m=2	0.867	0.86	0.86
m=3	0.719	0.666	0.667
m=4	0.612	0.492	0.471
m=5	0.59	0.445	0.416

Taking M=2, the class-wise breakup for the classification results using Naïve Bayes multinomial is as shown:

Table 1.10 classification results

	TP rate	FPrate	Precision	Recall
Commerce	0.943	0.095	0.782	0.943
Social Sciences	0.833	0.036	0.873	0.833
Natural Sciences	0.878	0.021	0.937	0.878
Aesthetics	0.775	0.034	0.876	0.775

Table 1.11 Confusion Matrix for Naïve Bayesian multi nomial Approach

a	b	c	d	← classified as
449	15	7	5	a = Commerce
35	344	11	23	b = Social Sciences
27	12	417	19	c = Natural sciences
63	23	10	331	d = Aesthetics

.

Error analysis in this case shows the confusion between the Natural Sciences and Commerce Classes and also Aesthetics and Commerce classes.

One important observation made in this work s that a paragraph may not have sufficient information to decide about the category, a classifier may behave poorly in such cases.

Fig 1.1 A sample paragraph which contains insufficient information about its category

commerce,'ಪದಾರ್ಥ ವಿಂಗಡಣೆ (Classification of Materials) ಪದಾರ್ಥಗಳನ್ನು ಅವುಗಳ ಪ್ರಕೃತಿಗನುಗುಣವಾಗಿ ವರ್ಗೀಕರಣ ಅಥವಾವಿಂಗಡಣೆ ಮಾಡಬೇಕು. ಘನ ಮತ್ತು ದ್ರವ ಪದಾರ್ಥಗಳನ್ನು ಪ್ರತ್ಯೇಕವಾಗಿಇಡಬೇಕಾಗುತ್ತದೆ. ಈ ಎರಡು ಪ್ರಕೃತಿಗಳಲ್ಲಿ ಅನೇಕ ವಿಧವಾದ ಪದಾರ್ಥಗಳುತಯಾರಿಕೆಗೆ ಅವಶ್ಯಕವಾಗಿರಬೇಕಾಗುತ್ತದೆ. ಪ್ರತಿ ವಿಧವಾದ ಪದಾರ್ಥದಲ್ಲಿ ವಿವಿಧ ಅಳತೆ,ತೂಕ, ಗಾತ್ರ ಮತ್ತು ಗುಣಮಟ್ಟಗಳಿರುತ್ತವೆ. ಕಬ್ಬಿಣ ಮತ್ತು ಉಕ್ಕು, ಎಣ್ಣೆಇತ್ಯಾದಿ ಭಿನ್ನ ಪದಾರ್ಥಗಳು, ರಸಾಯನಿಕ ಪದಾರ್ಥಗಳು, ಬಣ್ಣ, ಬಿಡಿ ಭಾಗಗಳುಪ್ರತ್ಯೇಕ ಸ್ಥಳದಲ್ಲಿಡಬೇಕಾಗುತ್ತದೆ. ಈ ಗುಂಪುಗಳ ಸ್ಥಳಕ್ಕೆ ಪ್ರಧಾನವಾಗಿಕಾಣುವ ಹಾಗೆ ನಾಮ ಫಲಕ (Sign Boards) ಗಳನ್ನುತಗುಲಿಸಿರಬೇಕು. ಇದು ಆಯಾಯ ಗುಂಪಿನ ಪದಾರ್ಥಗಳಿಗಿರಬೇಕಾದ ಎಚ್ಚರಿಕೆಕೊಟ್ಟಂತಾಗುತ್ತದೆ. ಉದಾ : ಒಂದು ಕಾರ್ಖಾನೆಗೆ ಬೇಕಾಗುವ ರಾಸಾಯನಿಕಪದಾರ್ಥ ದಾಸ್ತಾನು
ವಿಭಾಗದಲ್ಲಿಟ್ಟಿರುವಾಗಅಪಾಯಕಾರಿಯಾಗಿರುವಸಂದರ್ಭದಲ್ಲಿಎಚ್ಚರಿಕೆಯ ನಾಮ ಫಲಕ ಅಪಾಯದಿಂದ ದೂರವಿರುವ ಸೂಚನೆಯಾಗುತ್ತದೆ'

The above paragraph belongs to category Commerce but does not contain key words belonging to commerce category, which may lead to possible misclassification since the information may pertain to different categories.

3. CONCLUSIONS

Error analysis indicated the fail points were due to paragraphs being neutral of the class, as there is a significant possibility of paragraphs belonging to multiple classes. Paragraphs may use neighbouring paragraphs to get the class information. This can be captured to increase the performance of the classifier.

The work can be made use of in customer reviews in Kannada blogs. It can also be used in extracting opinions in posted Kannada articles online.

High/low precision and high /low recall may be desirable for certain applications. Hence depending on the requirements of an application – whether it requires high precision/recall, the appropriate methods can be chosen.

ACKNOWLEDGEMENTS

The authors thank Dr Anandarama Upadhyaya for his suggestions.

REFERENCES

[1] Alfons Juan,Hermann. Ney,'Reversing and Smoothing the Multinomial Naive Bayes Text Classifier',Work supported by the Spanish "Ministerio de Ciencia y Tecnolog´ia" under grant TIC2000-1703-CO3-01.

[2] Andrew McCallumzy,Kamal Nigamy,'A Comparison of Event Models for Naive Bayes Text Classication',AAAI-98, Workshop on learning for Text Categorization.1998.

[3] Maite Taboada,Julian Brooke,Manfred Stede,'Genre-Based Paragraph Classification for Sentiment Analysis',Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pp 62–70,Queen Mary University of London, 2009.

[4] Marti A. Hearst,'TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages',) Association for Computational Linguistics, pp 33-64,1997.

[5] Isaac Persing and Alan Davis and Vincent Ng,'Modeling Organization in Student Essays',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 229–239,MIT, Massachusetts, USA, pp 9-11 ,2010.

[6] Ashrof.M.Kibriya,EibeFrank,Bernard.P,Geoffrey Holm`es,'Multinomial Naive Bayes for Text',lecture notes in Artificial Intelligence,pp 488-499,2004.

[7] Jason D. Rennie,Lawrence,Shih,Jaim,Teevan,David R. Karger,'Tackling the Poor Assumptions of Naive Bayes Text Classifiers',Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003

[8] Jayashree.R,Srikantamurthy.K,'Analysis of Sentence Level Text Cassification in the Kannada Language',Proceedings of the 2011 International Conference on Soft Computing and Pattern Recognition'(SOCPAR_-11),pp 147-151,Dalian,China, 2011.

[9] Erdong Chen, Benjamin Snyder and Regina Barzilay,'Incremental Text Structuring with Online Hierarchical Ranking',Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational linguistics, pp. 83–91, Prague, 2007.

[10] Qinfeng Shi,Yasemin Altun,Alex Smola,S. V. N. Vishwanathan,'Semi-Markov Models for Sequence Segmentation',Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational linguistics, pp. 640–648, Prague, 2007.

Authors

Jayashree.R is an Associate Professor in the Department of Computer Science and Engineering, at PES Institute of Technology, Bangalore. She has over 18 Years of teaching experience. She has published several papers in international conferences and journals. Her research area of interest is Natural Language Processing and Pattern Recognition.

Dr.K.Srikanta Murthy is a Professor and Head, Department of Computer Science and Engineering, at PES School of Engineering, Bangalore. He has put in 24 years of service in teaching and 5 years in research. He has published 8 papers in reputed International Journals; 41 papers at various International Conferences. He is currently guiding 5 PhD students .He is also a member of various Board of Studies and Board of Examiners for different universities. His research areas include Image Processing ,Document Image analysis, Pattern Recognition, Character Recognition ,Data Mining and Artificial Intelligence.

Basavaraj S Anami is presently working as the Principal, K.L.E's Institute of Technology, Hubli, since August 2008. He completed his Bachelor of Engineering in Electrical Stream during November 1981. Then he completed his M.Tech in Computer Science at IIT Madras in March 1986. Later he received his Doctrine(PhD) in Computer Science at University of Mysore in January 2003.He began his academic journey as the Lecturer in Electrical department in BEC, Bagalkot from September 1983 up to December 1985. Then he was promoted as the In-charge Head of Department of Computer Science in the same college and in February 1990 2008.