Performance Analysis of Naïve Bayesian Methods for Paragraph Level Text Classification in the Kannada Language

R. Jayashree[1], K. Srikanta Murthy[2], and Basavaraj S. Anami[3]

[1] Department of Computer Science, PES Institute of Technology, Bangalore, India
[2] Department of Computer Science, PES School of Engineering, Bangalore, India
[3] Department of Computer Science, KLE Institute of Technology, Hubli, India
{jayashree,srikantamurthy}@pes.edu, anami_basu@hotmail.com

Abstract. Text Categorization plays a predominant role in Natural Language Processing (NLP) and Information Retrieval (IR) applications. This work highlights the performance of different Naïve Bayesian methods for paragraph level Text Classification in the Kannada language. The dimensionality reduction technique is achieved using minimum term frequency, stop word identification and removal methods.

Keywords: performance, classifier, paragraph level classification, Kannada text classification, Naïve Bayesian, Multinomial, naïve Bayesian upbeat able, Bayesnet.

1 Introduction

When we browse information present on the internet, the point worth noting is that information is mostly present as documents, paragraphs and sentences. Another important point to be noted is the way in which the information gets updated. Whenever the information updation needs to be done, the problem of finding the correct location to update information in a document is challenging.

This work looks at the possibility of paragraph classification in the Kannada language which helps in information updation online.

The rest of the paper is organized as follows. Section-II highlights the literature about paragraph level text classification in particular, Text categorization and Research on Naïve Bayesian models in general. Section-III describes how the corpus was prepared for use in this work. Section-IV discusses the methodology of our work. Section –V is about Results and Discussion.

2 Literature Survey

Jayashree.R Et.al, have investigated two classical approaches such as Naïve Bayesian and Bag of Words to Sentence Level Text Classification in the Kannada Language

C. Stephanidis (Ed.): Posters, Part II, HCII 2013, CCIS 374, pp. 435–439, 2013.

and looked at the possibility of extending sentence level classification task to Paragraph Level Text Classification in their future work [1].

Erdong Chen Et.al [2] in their work on 'Incremental Text structuring with on line hierarchical ranking' present an online ranking model which exploits the hierarchical structure of a given document.

The importance of paragraph segmentation is highlighted by Alex Smola Et.al [3], the application is speech to text conversion, wherein there is necessity to identify punctuations, paragraphs etc.

Isaac Persing Et.al [4] have worked on' modeling organization in student Essays', wherein the organization could be treated as collection of paragraphs with respect to the structure of the Essay.

'Genre Based Paragraph for Sentiment Analysis' is an interesting work carried out by Maite Taboada Et.al [6]. They present a classification system for representing different paragraphs within movie reviews.

Work by Andrew Mccallum Et.al[7] 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.

ABOUT The Corpus

The TDIL(Technology for Development of Indian Languages) corpus developed by Central Institute of Indian Languages(CIIL) is considered for use in this work. TDIL corpus contains pre categorized documents.

Table 1. Class wise Distribution of paragraphs in TDIL corpus

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

3 Methodology

Dimensionality Reduction:
In this work, we have achieved dimensionality reduction technique using two methods:

1. Stop word identification and removal
2. Using a restriction based on the word occurrence.

3.1 Naïve Bayes

The Naïve Bayesian is a probabilistic classifier. The dimensions in the vector indicate the presence of the word and no special weight age parameter was used in the classification process.

3.2 Naïve Bayesian Multinomial

According to Naïve Bayesian methods, a document is to be treated as bag of words wherein multiple occurring words appear multiple times. Hence we have used a modified form of Naïve Bayes which is Naïve Bayes Multinomial.

3.3 Naïve Bayes Multinomial Upbeatable

Bayes net is a probabilistic graphical model that represents a set of random variables.

3.4 Bayes Net

Is an incremental version that processes one instance at a time.

4 Results and Discussion

K-fold Cross Validation is used in this work for evaluation of the classifier performance, which is needed to ensure that each partition is used as a test set only once.

A classifier's performance can be measured by using parameters: Precision (P), Recall (R) (also called as TP rate) and F-Score (F). The definitions of the parameters are as shown:

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

$$\text{TP rate/True Positive(TP)} = \frac{\text{Proportion classified as class x}}{\text{Actual total of class x}} \quad (2)$$

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

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

4.1 Naive Bayes

with decreasing 'm', which is minimum term frequency, the evaluation parameters showed a significant rise, hence, we need to consider such words and their impact on classification. Taking m=2, the class-wise break up for the classification results by taking into consideration stop word removal, is as shown.

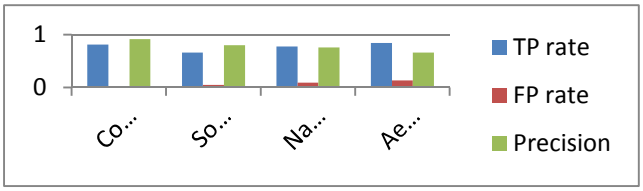

Fig. 1. Classification Results

4.2 Naïve Bayesian Multinomial

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

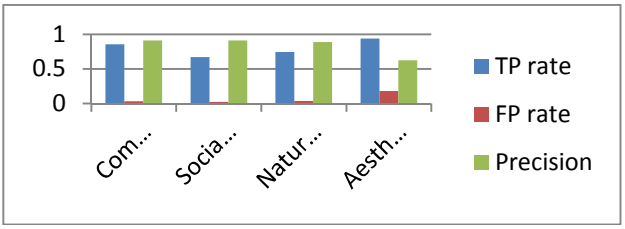

Fig. 2. Classification results for BayesNet

4.3 Bayesnet

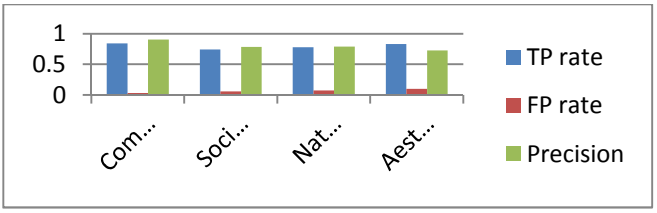

Fig. 3. Classification results

4.4 Naïve Bayes Multinomial Upbeatable

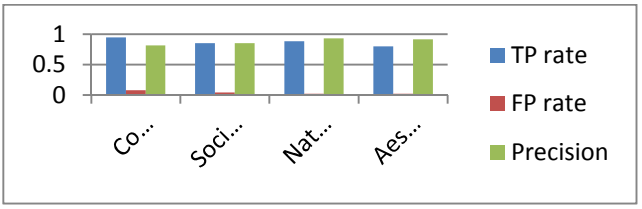

Fig. 4. Classification results

5 Conclusion

The distribution of the Minimum term frequency across categories varies in our experiments. Manual error analysis has shown that there is a significant possibility of paragraphs belonging to multiple classes. In some cases, paragraphs might not have sufficient information which indicate category and hence might use neighboring paragraphs to convey the class information.

References

1. Jayashree, R., Srikantamurthy, K.: Analysis of Sentence Level Text Cassification in the Kannada Language. In: Proceedings of the 2011 International Conference on Soft Computing and Pattern Recognition (SOCPAR 2011), Dalian, China, pp. 147–151 (October 2011)
2. Chen, E., Snyder, B., Barzilay, R.: Incremental Text Structuring with Online Hierarchical Ranking. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Linguistics, Prague, pp. 83–91 (June 2007)
3. Shi, Q., Altun, Y., Smola, A., Vishwanathan, S.V.N.: Semi-Markov Models for Sequence Segmentation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Linguistics, Prague, pp. 640–648 (June 2007)
4. Persing, I., Davis, A., Ng, V.: Modeling Organization in Student Essays. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, October 9-11, pp. 229–239. MIT, Massachusetts (2010)
5. Hearst, M.A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Association for Computational Linguistics (1997)
6. Taboada, M., Brooke, J., Stede, M.: Genre-Based Paragraph Classification for Sentiment Analysis. In: Proceedings of SIGDIAL 2009: The 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, Queen Mary University of London, pp. 62–70 (September 2009)
7. McCallumzy, A., Nigamy, K.: A Comparison of Event Models for Naive Bayes Text Classication. In: AAAI 1998, Workshop on Learning for Text Categorization (1998)