Text News Classification System using Naïve Bayes Classifier

Shruti Bajaj Mangal
Punajbi University Patiala
Sagi_bajaj@yahoo.co.in

Dr. Vishal Goyal
Punjabi University Patiala
Vishal.pup@gmail.com

Abstract
This paper describes the Naive Bayes text News classification system developed for Punjabi Language. News corpus is used for training and testing purpose of the classifiers. Language specific preprocessing techniques are applied on raw data to generate a standardized and reduced-feature lexicon. Punjabi language is morphological rich language which makes those tasks complex. Statistical characteristics of corpus and lexicon are measured which show satisfactory results of text preprocessing module. We are able to get satisfactory results using Naive Bayes Classifier.

Keywords
Naïve Bayes, Text classification, Punjabi.

Introduction

Text classification is a process of classifying unknown text automatically by suggesting most probable class to which it belongs. Text classification is involved in many applications like text filtering, document organization, classification of news stories, searching for interesting information on the web, spam e-mail filtering etc. These are language specific systems mostly designed for English and European languages but very less work has been done for Punjabi language. So, developing classification systems for Punjabi documents is a challenging task due to morphological richness and scarcity of resources of the language like automatic tools for tokenization, feature selection, stemming etc. Inductive models like Naïve Bayes classifier, Support Vector Machines (SVMs); Boosting or MaxEnt models are commonly used for text classification. They are trained from labeled training data, under the assumption that the data is drawn from the same distribution as the test set. Manually labeling such a training set is often expensive for practical applications. It is also difficult when the label set contains more than a few destination classes – inter-labeler agreement rate is often low in this case, which leads to inconsistently labeled training data. On the other hand, a large amount of labeled data, which may have been drawn from a distribution that is discrepant from that of the test set, is often available in the same application domain.

Related Work

Text Classification for Indian Languages not widely explored. Exponential enhancement in the information related to Indian languages on the web, automatic information processing and retrieval become a vital need. So far very little work has been done for text classification with respect to Indian languages as compared to European languages. The problems faced by many Indian Languages face various problem like

free-word order language ,no capitalization, non-availability of large gazetteer lists, lack of standardization and spelling, resources and tools. These corpus documents are classified manually into defined classes and they are used as training data. Indian Languages are extremely inflectional and derivational language, yield to a very large number of inflected word forms for each root word. This makes the classification task more complex. Prabowo and Thelwall(2009)[21] used three approaches Rule Based, Support Vector Machine and Hybrid algorithms for classification. Andrew Maccalum, Kamal Nigam (1999)[1]considers the task of learning text classifiers with no labeled documents at all.

Knowledge about the classes of interest is provided in the form of a few keywords per class and a class hierarchy. Keywords are typically generated more quickly and easily than even a small number of labeled documents. Chowdhury Mofizur, Ferdous Ahmed Sohel (2010)[4] considers its proposed method to classify text is an implementation of association rule with a combined use of Naive Bayes Classifier and Genetic Algorithm, researcher have used the features of association rule to make association sets. A.Kanaka Durga, A.Govardhan (2011)[17] System has been developed ontology based text classification for Telugu documents and retrieval system. Researcher's aim is to capture the semantics of a text. K Raghuveer and Kavi Narayana Murthy (2007)[15]presents work on automatic text categorization in Indian languages. Here they used purely corpus based machine learning techniques. William B. Cavnar and John M. Trenkle[24] In this paper they described N-gram-based approach to text categorization that is tolerant of textual errors. All these techniques are used for English and for other foreign languages. There is very less work has been done for Indian languages and for Punjabi there is only one system [19], which is sports domain specific, and used to classify document into different sports categories. Therefore, our aim was to develop a text classification for a Punjabi language.

Naive Bayes Classifier

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying BYES' theorem with strong (naive) independence assumptions between the features. Naive Bayes models are also known under a variety of names in the literature, including simple Bayes and independence Bayes. All these names refers the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. To classify document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(c)} \quad \text{(i)}$$

Most Likely class of the document:

$$C = \underset{c \in C}{argmax} P(c|d) \quad \text{(ii)}$$

Bayes Rule defined as:

$$C = \underset{c \in C}{argmax} \frac{P(d|c)P(c)}{P(d)} \quad \text{(iii)}$$

Dropping the denominator from Bayes Rule:

$$C = \underset{c \in C}{argmax} P(d|c)P(c) \quad (iv)$$

For learning the Multinomial Naive Bayes Model, Parameter estimates the maximum likelihood. Simply use the frequencies in the training data. We need to calculate frequency of the documents per class and by number of total documents.

$$P(C_j = \frac{documnet(C=c)}{N_{Doc}}) \quad \text{(v)}$$

$$P(w_i|C_j) = \frac{count(w_i, C_j)}{\sum_{w \in} Count(w, c_j)} \quad \text{(vi)}$$

For parameter estimation we need to calculate the probability of occurrences of any given word in particular document of class. $count(w_i, C_j)$ Calculates fraction of times word Wi appears among all the words in documents of topic cj. Where $\sum_{w \in} Count(w, c_j)$ represents the mega document or collection of all words belongs to all documents or classes.

Results and Evaluation

To evaluate the system, we have used unknown testing data which was not part of the training data. Testing data was collected from online news websites like ***ajitjalandhar.com*** and other Punjabi news websites. Detail of the testing data shown below:

S. No	Domain	No. of Files	Total Words in all files
1	Terror Attack News	30	1342
2.	Murder Related News	35	1522
3.	Accidental News	46	2521
4.	Suicide News	20	1023

Table 1: Testing Data

Evaluation of the system has been done using standard metrics Recall Precision and F1-Measure.

$$Recall := \frac{\#no.of\ correct\ outputs\ returned\ by\ the\ system}{\#no.of\ Total\ files\ Tested}$$

$$Precision := \frac{\#no.of\ Correct\ outputs\ returned\ by\ system}{\#no.Actual(true)predictions}$$

$$F1 - Measure := 2 * \frac{(R * p)}{R + P}$$

S.No.	Tests	Results
1	Recall	0.72
2	Precision	0.78
3	F1-Measure	0.74

Table 8: Overall System Accuracy

S.No.	Domain	Accuracy
1	Terror Attack News	0.72
2	Accidental News	0.80
3	Murder Related News	0.64
4	Suicide Related News	0.72

Table 9: Individual Accuracy of Each class

Conclusion

Text classification system is vital system in area of Natural Language Processing. Text classification is used by many online and offline system to categorize text into predefined classes. The problem of classification has been widely studied in the database, data mining, and information retrieval communities. We have successfully implemented and tested Naive Bayes classifier. The system has the capabilities to classify given text news into four different categories. We are able to achieve satisfactory results based on our training data, which was not available at that moment. We have collected training data from various online resources, which was a very challenging and time-consuming task for this system. Punjabi is a resource poor language as compared to European language like English where one can find enough resources for training and testing the system. Based on our collected training data, which was not in much amount, we are able to achieve satisfactory results from this classifier system.

References

[1]. Andrew McCallum, Kamal Nigam, "Text Classification by Bootstrapping with Keywords, EM and Shrinkage", In Proceedings of the Workshop Held in conjunction with The 37th Annual Meeting of the Association for Computational Linguistics, 21 June 1999, pp 52-58.

[2].Abbas Raza Ali et.al., Urdu Text Classification, Accessed From: http://www.percipience.eu/papers/Urdu%20Text%20Classification.pdf Accessed on: 01/03/2014

[3].Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay. "Language Independent Named Entity Recognition in Indian Languages.",in Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India,. Asian Federation of Natural Language Processing , January 2008, pp. 33–40

[4].Chowdhury Mofizur, Ferdous Ahmed Sohel et.al. "Text Classification using the Concept of Association Rule of Data Mining," 2010. pp. 234-241

[5].David Lewis. Naive Bayes at forty: The independence assumption in information retrieval. In ECML'98: Tenth European Conference On Machine Learning, 1998.

[6].http://en.wikipedia.org/wiki/File:CART_tree_titanic_survivors.png Accessed on: 01/03/2014

[7].http://en.wikipedia.org/wiki/KNN Accessed on: 01/03/2014

[8].http://www.solver.com/xlminer/help/neural-networks-classification-intro Accessed on: 01/03/2014

[9].http://en.wikipedia.org/wiki/Artificial_neural_network Accessed on: 01/03/2014

[10].http://en.wikipedia.org/wiki/Naive_Bayes_classifier Accessed on: 01/03/2014

[11].http://en.wikipedia.org/wiki/Rocchio_algorithm Accessed on: 01/03/2014

[12].http://en.wikipedia.org/wiki/Nearest_centroid_classifier Accessed on: 01/3/2014

[13].http://en.wikipedia.org/wiki/Decision_tree_learning Accessed on: 03/3/2014

[14].Jake D. Brutlag Challenges of the Email Domain for Text Classification, Accessed form: http://research.microsoft.com/pubs/73532/AF1-1.pdf Dated 28/02/2014

[15].K Raghuveer and Kavi Narayana Murthy, Text Categorization in Indian Languages using Machine Learning Approaches, Accessed from: http://202.41.85.68/knm- publications/text-cat-iicai-2007.pdf Accessed on: 28/02/2014

[16]. Mita K. Dalal, Mukesh A. Zaveri, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications, Volume 28– No.2, August 2011, pp 37- 40

[17]. Mrs.A.Kanaka Durga, Dr.A.Govardhan, Ontology Based Text Categorization - Telugu Documents, International Journal of Scientific & Engineering Research Volume 2, Issue 9, September 2011, pp 1-4

[18]. Nidhi, Vishal Gupta. "Algorithm for Punjabi Text Classification", International Journal of Computer Applications, Volume 37– No.11, January 2012, pp.30-35.
[19]. Nidhi, Vishal Gupta," Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach " in proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, COLING 2012, Mumbai, December 2012, pp 109–122.
[20]. Praveen Kumar P "A Hybrid Named Entity Recognition System for South Asian Languages" In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing, January 2008, pp.83–88.
[21]. Rudy Prabowo, Mike Thelwall," Sentiment analysis: A combined approach ", Journal of Informetrics Volume 3, Issue 2, April 2009, pp 143–157.
[22]. Susan Dumais,John Platt, Inductive Learning Algorithms and Representations for Text Categorization, ACM New York, NY, USA 1998,pp 148-155
[23].Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In ICML97, 1997 pp.30-35
[24]. William B. Cavnar , John M. Trenkle. N-Gram-Based Text Categorization, 1994, pp.45- 51