International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Punjabi Text Classification using Naive Bayes Algorithm

Ubeeka Jain†* and Kavita Saini†

†Department of Computer Science Engineering, Rayat College of Engg. & IT, Ropar, Punjab, India

Accepted 06 Dec 2015, Available online 12 Dec 2015, **Vol.5, No.6 (Dec 2015)**

Abstract

Now-a-days, text classification is very necessary for an every field to organise the text documents. Till now there is no classifier available for classification of Punjabi documents. There are two new algorithms, one is ontology based and second is hybrid approach are proposed for Punjabi text classification. Here we have some Punjabi news article examples which we have to classify with the help of algorithms. Punjabi is a Indo Aryan language spoken in west Punjab (Pakistan) and East Punjab (India). So, a little work has been done in Punjabi text classification. The problem tackled by many Indian languages that is no capitalization, lack of standardization, spelling and scarcity of tools. Punjabi language has more inflectional forms than English language.

Keywords: *Punjabi text classification, news articles, ontology based and hybrid approach.*

Introduction

Text classification is a task to sort a set of documents automatically into categories from a predefined set. The large quantity of electronic data is available such as digital libraries, blogs, and electronic newspapers, electronic publication, emails, electronic books is very increasing rapidly. As the electronic data volume increases the challenges to manage the data is also increased (Nidhi *et al*, 2012)

There are two type of text classification first is automatic and second is manual text classification.

Now a day an automatic text classification becomes an important research issue in text mining. Manual text classification is very time consuming and an expensive. So automatic classification is much better than manual text classification. There are two machine learning methods to improve classification that is supervised methods where predefined classes are given to text documents with help of labelled document and unsupervised method is not involve labelled document to categorised the text documents. Text classification is included in many applications like document organization, searching of interesting information, text filtering

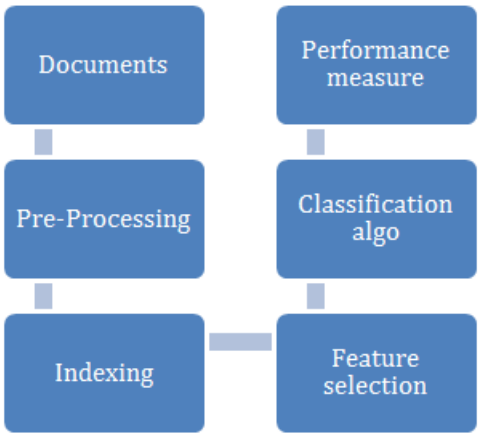
Classification news and spam e-mail etc. These are language specific machines which are mostly designed for English and foreign languages but a

Very little work has been done in Punjabi language. So Punjabi documents is challenging task to be classified. There are some automatic tools for tokenisation, stemming and feature selection. We are using statistical approach to classify Punjabi documents. Statistical approach using Naïve Bayes or support vector machine used classify particular sentences after objective sentences aimed at Urdu language. Where Urdu is morphological rich language and it is very difficult to classify the text in Urdu (Vishal Gupta *et al*, 2011)

There are three phase for processing

- Pre-processing phase.
- Feature extraction phase.
- Processing phase.

Text classification process



Text Categorization

As the dimensions of information offered on the Internet and shared intranets continue to rise, there is

*Corresponding author: **Ubeeka Jain**

on the rise interest in assisting people enhanced find, filter, or accomplish these assets. Text categorization is the consignment of natural language copies to one or more than predefined classes based taking place their contented is vital component in many information society and organization tasks. It is most general application to period has been for passing on subject groupings to documents to support text recovery, direction discovery or clarifying (Nidhi *et al*, 2012)

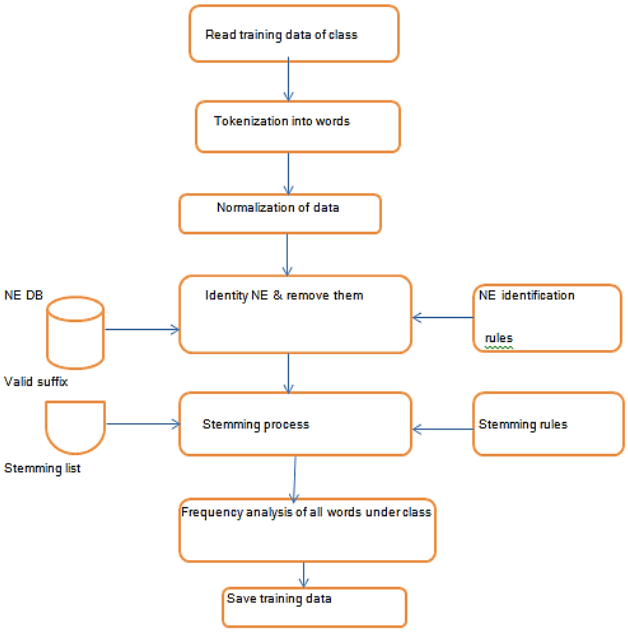
Automatic Text Categorisation

Automatic text classification can performs significant part in a wide variation of more flexible, active and modified information organization responsibilities as well real-time arranging of email or files into folder pyramids; topic documentation to support topic-specific processing actions; organized search or surfing; or finding text files that equal to long-term stand-up interests or more dynamic task-based interests. Grouping machineries should be capable to support category arrangements that are very common, consistent across those, and relatively static (e.g., Medical Subject Headings (Mesh), Dewey Decimal or Library of Congress classification systems, or Yahoo!'s topic hierarchy), as well as those that are more active and modified to specific interests or tasks (e.g., email about CIKM conference). In many contexts (Mesh, Dewey, Yahoo!, Cyber Patrol), skilled specialists are employed to classify fresh items. This procedure is very timewasting and expensive, thus limiting its applicability. So here is better interest in emerging machineries for automatic text classification (Bhumika *et al*, 2013)

News Classification

Every day editors at Dow Jones consign codes to hundreds of levels initiating from diverse resources such as newspapers, newswires, v or press releases. Each editor must master the 350 and so many different codes, gathered into seven classifications business, bazaar sector, product, subject, government action, and area. Due to the high volume of levels, classically several thousand per day, by finger coding all levels dependably and with high recall in a timely way is impractical. In general, different editors may code papers with variable levels of stability, accuracy, and wholeness. The coding assignment contains of conveying one or further codes to a text file. It shows the text of a classic story with codes (Shruti Bajaj *et al*, 2014) The codes look as if in the header are the ones allocated by editors or the codes subsequent "Proposed Codes" are those proposed by the automatic system. Every code is scoring in the left hand column, illustrative the influences of several near contests. By changing the score commencement, we can trade-off evoke and correctness.

Overview of System



System design and implementation

The system is divided into two portions. First is training and the second is testing.

Training Process

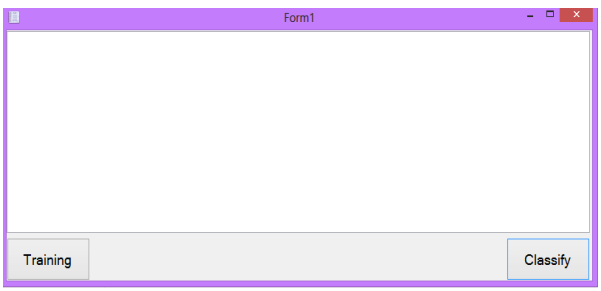
First of all we have collected the training data. The training data is raw text collected from various sources. This training data is saved in a separate file. For the training process of the system machine learning approach is used. The training file is saved in the memory as binary file.

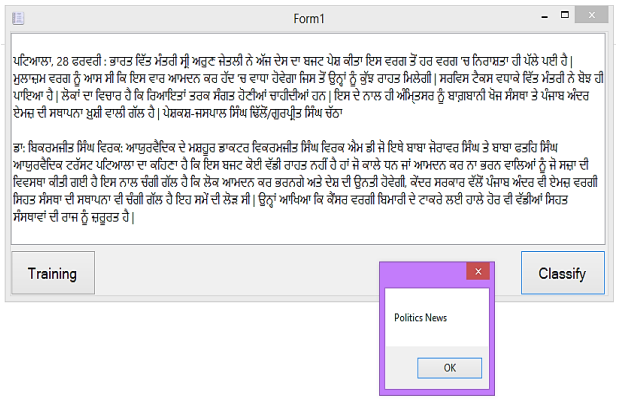
Testing Process

During the testing process Naive based algorithm is used. When we put some text into the system then system will be already trained. Clicking on the button classified and formed output.

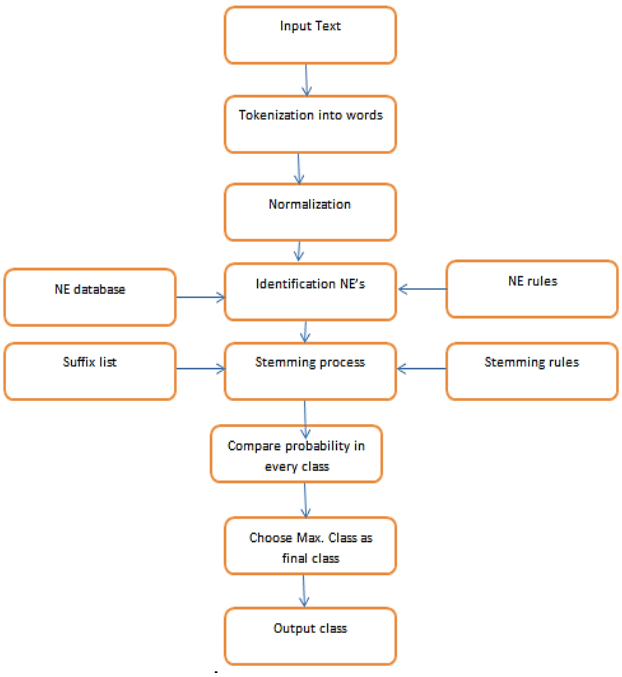
The system is implemented in Microsoft visual c#.

Some input and output





Testing and result



Firstly we have to put input text in the system and tokenization into words. After tokenization normalization will be performed. Then named entity rules will follows. It has been done in two parts, first is named entity database and named entity rules. After that stemming process will be done .It have two functions that are suffix list and stemming rules. Then compare the probability inverse class and choose maximum class as final class and output class has been performed.

Result and evaluation

To evaluation of the system we have been used testing data. Testing data has been collected from online news websites etc. Detail of testing data shown below:-

S. No.	Domain	No .of Files
1	Business	100
2	Cinema	100
3	Politics	100
4	Sports	100

Conclusion

Text classification is used to organise and manage the data from predefined data set. There is a little work has done in Punjabi text classification. We are using statistical approach to classify the text. To develop stemmer rules to get root words. Text classification system is vsital 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

Nidhi, Vishal Gupta (2012), University Institute of Engineering and Technology, Panjab University. Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach.

Vishal Gupta and Gurpreet Singh Lehal (2011), Department of Computer Science, Punjabi University Patiala, India. International Journal of Computer Applications.Named Entity Recognition for Punjabi Language Text Summarization.

Nidhi, Vishal Gupta (2012), University Institute of Engineering and Technology, Panjab University, Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach.

Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar (2013), International Journal of Application or Innovation in Engineering & Management (IJAIEM).

Shruti Bajaj Mangal, Dr. Vishal Goya Research Cell (2014), An International Journal of Engineering Sciences, Issue December, Vidya Publications. Authors are responsible for any plagiarism issues. Text News Classification System using Naïve Bayes Classifier.

Vishal Gupta and Gurpreet Singh Lehal (2011), Department of Computer Science, Punjabi University Patiala, India. International Journal of Computer Applications. Named Entity Recognition for Punjabi Language Text Summarization.

Surat C Namrata Mahender (March 2012), International Journal of Artificial Intelligence & Applications (IJAIA), Text classification and classifiers: a survey Vandana Korde Sardar Vallabhbhai National Institute of Technology,

Nidhi, Vishal Gupta (2012), University Institute of Engineering and Technology, Panjab University, Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach.

M Narayana Swamy ,M. Hanumanthappa, International Journal of Data Mining Techniques and Applications Indian Language Text Representation and Categorization Using Supervised Learning Algorithm.

Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar (2013), International Journal of Application or Innovation in Engineering & Management (IJAIEM).

Meera Patil and Pravin Game, ACEEE Int. J. on Information Technology (March 2014) Comparison of Marathi Text Classifier

Bijal Dalwadi Vishal Polara Chintan Mahant (March 2015), International Journal of Engineering, Technology, Management and Applied Sciences A Review: Text Categorization for Indian Language