ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

A Review on the Punjabi Text Classification using Natural Language Processing

Ubeeka Jain[1], Kavita Saini[2]

Assistant Professor, R.I.E.I.T, Railmajra, Punjab[1]

M. Tech research scholar, R.I.E.I.T, Railmajra, Punjab[2]

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 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.

I. 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. [1]

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

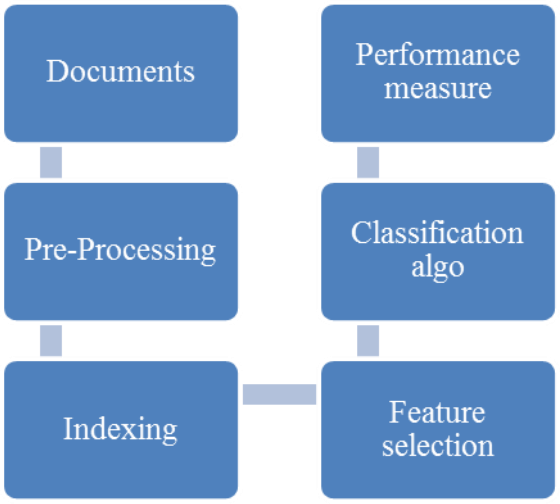
Now a day an automatic text classification becomes an important research issue in text mining. Manual text classification is very time consuming and an expensive. So automatic classification is much better than manual text classification. There are two machine learning methods to improve classification 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. [2]

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 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. [3]

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

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

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. [4]

II. 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. [5] 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.

III. CLASSIFICATION METHODS HAVE BEEN USEFUL TO

1. Spam filtering, a procedure which attempts to separate E-mail spam mails from authentic emails.
2. Email routing, transfer an email sent to all-purpose statement to a precise address or mailbox dependent on topic.
3. Language identification, mechanically influential the language of a text
4. Type classification, automatically influential the type of a text.
5. Readability assessment, mechanically responsible the degree of ability to read of a text, either to find appropriate resources for different age clusters or reader types or as part of a larger text to simplify the structure
6. Sentiment analysis, defining the approach of a speaker or a writer with respect to some topics or the overall related polarization of a text.Object triage, selecting objects that are related for manual collected works curation, for example as is being done as the first step to produce manually curated explanation catalogues in ecology.

IV. LITERATURE SURVEY

[6] In **2011**, Vishal Gupta and Gurpreet Singh Lehal used NER system for Punjabi news. It is generating precision =89.32% and recall = 83.4% and score is =86.25%. In this there are 50 news will be used to implement and check errors. There are 13.75% errors are occurred. Prefix rule generated no error and suffix rule generate 1% error. As such in name NER score is 2. Last name generate 10% error Exact name rule generated 0.25% error.

[7] In **2012**, Vandana Korde and C Namrata Mahender Used different type of algorithm the classification of text and text mining that is Rocchio's algorithm, K-nearest neighbours, Naïve Bayes, decision tree and neural networks etc. It was confirmed from study of information Grain and Chi square and statistics are most commonly used and well performed method to feature selection. Where existing system is compared to different type of parameters.

[8] In **2012**, Nidhi and Vishal Gupta has been work done in Punjabi text files using ontology and hybrid based approach. In the Punjabi text classification contain 180 Punjabi text documents and 45files used as training data. All type of text classifier based on Naïve Bayes and centroid based classification methods. All documents are consults to sports category.

[9] In **2013**, M. Narayana Swamy and M. Hanumanthappa used supervised learning algorithms for text representation and categorization. In this K-nearest neighbour, Naïve Bayes and decision tree C4.5 for south Indian language such as Telgu, Kannada and Tamil text have been calculated. In this 300 documents are including and divided into three categories. All documents are related to cinema. KNN gives 93% accuracy and decision tree give 97.66% accuracy. As result there is Naïve Bayes algorithm is best for text categorization.

[10] In **2013**, Bhumika, Sukhjit Singh Sehra and Anand Nayyar used different type of algorithms on classification. In this there are automatically sorting a set of document into categories that are from predefined set such as biological database, chat rooms, online forums, electronically, digital libraries and news articles etc. There are different types of algorithms are used such as Hunt method, sequential decision tree and parallel formulation etc.

[11] In **2014**, Meera Patil and Pravin Game compare the Marathi text classifiers. In this they compared Naïve Bayes, K-nearest neighbour, centroid classifier and modified K-nearest neighbour classifier. In this there are 4000 Marathi text documents being tested. As result Naïve Bayes is most efficient algorithm among the four classifications for time and accuracy.

[12] In **2015**, Bijal Dalwadi, Vishal polara and Chintan Mahant categorization of text for Indian languages. In this text categorization play important role in machine learning, text mining and information retrieval. It has been successful in handling wide varieties of world. In this they

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

have discussed various types of approaches and methods for text categorization of Indian languages.

V. PROBLEM FORMULATION

Now-a-days, text classification is very necessary for an every field to organise the text documents. Till now there are a few classifiers 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 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. 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 from objective sentences for Urdu language. Where Urdu is morphological rich language and it is very difficult to classify the text in Urdu.

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.

REFERENCES

[1] Nidhi, Vishal Gupta University Institute of Engineering and Technology, Panjab University. Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach(2012).
[2] Vishal Gupta and Gurpreet Singh Lehal Department of Computer Science, Punjabi University Patiala, India. International Journal of Computer Applications (2011) .Named Entity Recognition for Punjabi Language Text Summarization.
[3] Nidhi, Vishal Gupta University Institute of Engineering and Technology, Panjab University. Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach(2012).
[4] Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar. International Journal of Application or Innovation in Engineering & Management (IJAIEM)(2013).
[5] Shruti Bajaj Mangal, Dr. Vishal Goya Research Cell : An International Journal of Engineering Sciences, Issue December (2014), Vidya Publications. Authors are responsible for any plagiarism issues**.** Text News Classification System using Naïve Bayes Classifier.
[6] Vishal Gupta and Gurpreet Singh Lehal Department of Computer Science, Punjabi University Patiala, India. International Journal of Computer Applications(2011). Named Entity Recognition for Punjabi Language Text Summarization.
[7] International Journal of Artificial Intelligence & Application(IJAIA), March (2012) TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY Vandana Korde Sardar Vallabhbhai National Institute of Technology, Surat C Namrata Mahender Department of Computer Science& IT, Aurangabad.
[8] Nidhi, Vishal Gupta University Institute of Engineering and Technology, Panjab University. Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach(2012).
[9] International Journal of Data Mining Techniques and Applications Indian Language Text Representation and Categorization Using Supervised Learning Algorithm. M Narayana Swamy ,M. Hanumanthappa Department of Computer Applications, Presidency College ,Bangalore ,India, Department of Computer Science & Applications, Bangalore University, Bangalore, India.
[10] Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar. International Journal of Application or Innovation in Engineering & Management (IJAIEM)(2013).
[11] ACEEE Int. J. on Information Technology, March 2014 Comparison of Marathi Text Classifier Meera Patil and Pravin Game Pune Institute of Computer Technology, Computer Department, Pune, India.
[12] International Journal of Engineering Technology, Management and Applied Sciences March 2015 A Review: Text Categorization for Indian Language Bijal Dalwadi Vishal Polara Chintan Mahant.