Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach

Nidhi[1] and Vishal Gupta[2]

Department of Computer Science and Engineering,
Panjab University, Chandigarh, India
[1]naseeb.nidhi@gmail.com
[2]vishal@pu.ac.in

ABSTRACT

Punjabi Text Classification is the process of assigning predefined classes to the unlabelled text documents. Because of dramatic increase in the amount of content available in digital form, text classification becomes an urgent need to manage the digital data efficiently and accurately. Till now no Punjabi Text Classifier is available for Punjabi Text Documents. Therefore, in this paper, existing classification algorithm such as Naïve Bayes, Centroid Based techniques are used for Punjabi Text Classification. And one new approach is proposed for the Punjabi Text Documents which is the combination Naïve Bayes (to extract the relevant features so as to reduce the dimensionality) and Ontology Based Classification (that act as text classifier that used extracted features). These algorithms are performed over 184 Punjabi News Articles on Sports that classify the documents into 7 classes such as ਕ੍ਰਿਕਟ (krikaṭ), ਹਾਕੀ (hākī), ਕਬੱਡੀ (kabḍḍī), ਫੁਟਬਾਲ (phuṭbāl), ਟੈਨਿਸ (ṭainis), ਬੈਡਮਿੰਟਨ (baiḍminṭan), ਓਲੰਪਿਕ (ōlmpik).

KEYWORDS

Punjabi Text Classification, Hybrid Approach, Naïve Bayes, Centroid Based Classification, Ontology Based Classification (Domain Specific).

1. INTRODUCTION

Nowadays, most of the available contents are in digital form, and managing such data become difficult. Therefore, Text Classification is done to classify the documents into predefined classes, this result in increased access rate and efficiency of the search engine. Manual text classification is an expensive and time-consuming method, as it become difficult to classify millions of documents manually. Therefore, automatic text classifier is constructed using labeled documents and its accuracy is much better than manual text classification and it is less time consuming too [1] [2].

Text Classification task is two-step processes: 1) Training Phase: the set of documents in this phase is called training set. Each document in the training set belongs to particular class depends on their contents, called Labelled Documents. Training Phase helps text classifier to classify the unlabeled documents. 2) Testing Phase (also called Classification Phase): In this phase, unlabeled documents are classified using labelled documents. And to estimate the accuracy of the classifier, known class of the unlabeled document is compared with classification result [3] [4] [5].

Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, CRYPSIS, ICAIT, ICDIP, ITCSE, CS & IT 07,
pp. 245–252, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2421

1.1 Punjabi Text Classification Phases

Punjabi Text Classification consists of three phases:

1.1.1 Preprocessing Phase

This phase includes stopwords removal, stemming (e.g. ਖਿਡਾਰੀਆਂ (khiḍārīāṃ) → ਖਿਡਾਰੀ (khiḍārī), ਟੀਮਾਂ (ṭīmāṃà)→ ਟੀਮ (ṭīm)), punctuation marks removal and special symbols removal (<,>, :,{,},[,],^,&,*,(,) etc.), as they all are irrelevant to classification task [6]. Stemming is done using Punjabi Language Stemmer given in [7]. Table 1 shows lists of some stopwords that are removed from the document.

Table 1. Stopwords List

ਲਈ (laī)	ਨੇ (nē)	ਆਪਣੇ (āpaṇē)	ਨਹੀਂ (nahīṃ)	ਤਾਂ (tāṃ)
ਇਹ (ih)	ਹੀ (hī)	ਜਾਂ (jāṃ)	ਦਿੱਤਾ (dittā)	ਹੋ (hō)

1.1.2 Feature Extraction

For Punjabi Text Classification, TF*IDF is used as statistical approach to extract the relevant features, e.g. words having threshold value less than 2 are not considered as features [3][8]. And for linguistic approach, gazetteer lists are prepared especially for Punjabi Language e.g. list of middle and lastnames (ਸਿੰਘ, ਸੰਧੂ, ਗੁਪਤਾ, (siṅgh, sandhū, guptā)), places (ਬਠਿੰਡਾ, ਚੰਡੀਗੜ੍ਹ ਪੰਜਾਬ, (baṭhiṇḍā, caṇḍīgaṛh, pañjāb)), date/time (ਕੱਲ (kall), ਸਵੇਰ (savēr)), abbreviations (ਆਈ (āī), ਸੀ (sī), ਐਲ (ail), ਪੀ (pī), ਬੀ (bī)), designations (ਕਪਤਾਨ (kaptān), ਕੋਚ (kōc), ਕੈਪਟਨ (kaipṭan)) etc. to remove non-relevant features from the documents. Also certain rules are also constructed to extract location names and first names. E.g. If Punjabi word ਵਿਖੇ (vikhē) is found, its previous word is extracted as location name. If Punjabi word ਪਿੰਡ (piṇḍ) is found, its next word is extracted as location name. If Punjabi word ਜ਼ਿਲ੍ਹੇ (zilhē) is found, its previous word is extracted as location name. If lastname is found, its previous word is checked whether it is middle name or not, if it is, it's previous word is extracted from the document as first name. Else, the same is extracted as first name.

1.1.3 Processing Phase

In this phase, text classification techniques such as Centroid based Classification, Bayesian Classification are implemented to classify the Punjabi Text documents.

As no work has been done in this field for Punjabi Text Documents, in this paper, we presented three classification algorithms for Punjabi Text Documents, these are: Naïve Bayes Classification, Centroid Based Classification and one new approach is proposed for Punjabi text Classification for the first time which is the combination of Naïve Bayes and Ontology based Classification. This paper also highlights the text classification work done for Indian languages.

2. Related Work Done for Indian Languages

The development of internet led to an exponential increase in the amount of electronic documents not only in English but also other regional languages. So far very little work has been done for text classification with respect to Indian languages, due to the problems faced by many Indian Languages such as: No capitalization, non-availability of large gazetteer lists, lack of standardization and spelling, scarcity of resources and tools, free word order language. The only corpus available in most languages is an EMILLE/CIIL corpus that contains about 3 million words. These corpus documents are classified manually; hence they are used as Training Set. Indian Languages, especially Dravidian languages (Tamil, Telugu, Kannada, and Malayalam) are highly inflectional and derivational language, leading to a very large number of word forms for each root word. This makes the classification task more challenging [9].

Text Classification Techniques implemented for Southern Indian Languages are, e.g. Naive Bayes classifier has been applied to Telugu news articles in four major classes to about 800 documents. In this, normalized TFXIDF is used to extract the features from the document. Without any stopword removal and morphological analysis, at the threshold of 0.03, the classifier gives 93% precision. [10]; Semantic based classification using Sanskrit wordnet used to classify Sanskrit Text Document, this method is built on lexical chain of linking significant words that are about a particular topic with the help of hypernym relation in WordNet. [11]; statistical techniques using Naïve Bayes and Support Vector Machine used to classify subjective sentences from objective sentences for Urdu language, in this, language specific preprocessing used to extract the relevant features. As Urdu language is morphological rich language, this makes the classification task more difficult. The result of this implementation shows that accuracy, performance of Support Vector Machines is much better than Naïve Bayes Classification techniques. [12]; for Bangla Text Classification, n-gram based algorithm is used and to analyze the performance of the classifier Prothom-Alo news corpus is used. The result show that as we increase the value of n from 1 to 3, performance of the text classification also increases, but from value 3 to 4 performance decreases [13]; for morphologically rich Dravidian classical language Tamil, text classification is done using Vector Space Model and Artificial Neural network. The experimental results show that Artificial Neural network model achieves 93.33% which is better than the performance of Vector Space Model which yields 90.33% on Tamil document classification [14]. A new technique called Sentence level Classification is done for Kannada language; in this we analyze sentences to classify the Kannada documents as most user's comments, queries, opinions etc are expressed using sentences. This Technique extended further to sentiment classification, Question Answering, Text Summarization and also for customer reviews in Kannada Blogs [15].

But for Punjabi Text Document, not much work has been done to classify the documents due to lack of resources, annotated corpora, name dictionaries, good morphological analyzers, POS taggers are not yet available in the required measure. Therefore, in section 3, we present three classification algorithms for Punjabi Text Classification.

3. Punjabi Text Classification Algorithms

3.1 Naïve Bayes Classification

In this algorithm, we consider each Punjabi Text Document d as Bag of words i.e. d= (w_1, w_2,……w_n) where w_n is the n^{th} word in the document and then for classification calculate the posterior probability of the word of the document being annotated to a particular class [16].

Table 2. Naive Bayes Punjabi Text Classification

Training Set	Step1: Calculate total words in each class in the Training set. Step2: Calculate total words in Training set. Step3: Calculate P(c) the prior probability of a document occurring in each class c, P(c)= Total no. of words in class c/ Total No. of words in Training Set
Test set	Step4: For each class c, calculate the probability of a document d being in class c, $P(c|d)=P(c|w_1, w_2……w_n)/ n$ (1) P(c|d) = (Total Documents match in class c) * (P(c)/((Total words in class c)*(Total words in training set))
Input	ਰਾਜਵੰਤ ਸਿੰਘ ਹਾਕੀ ਜਗਤ ਦਾ ਉਹ ਹਸਤਾਖਰ ਹੈ, ਜਿਸ ਨੇ ਬਿਨਾਂ ਕਿਸੇ ਸਰਕਾਰੀ ਮਦਦ ਤੋਂ ਬਠਿੰਡਾ ਜ਼ਿਲ੍ਹੇ 'ਚ ਹਾਕੀ (rājvant siṅgh hākī jagat dā uh hastākhar hai, jis nē bināṃ kisē sarkārī madad tōṃ baṭhiṇḍā zilhē 'c hākī)
Preprocessing and Feature Extraction Phase	ਹਾਕੀ ਜਗਤ ਹਸਤਾਖਰ ਸਰਕਾਰੀ ਮਦਦ ਹਾਕੀ (hākī jagat hastākhar sarkārī madad hākī)
Output	Class: ਹਾਕੀ (hākī)

3.2 Centroid Based Classification

In this technique, each document d is represented by vector called Document Vector V(d) in the feature space and each component of the vector is represented by TF*IDF value i.e. V(d)=(tf*idf1, tf*idf2,......tf*idfn) where n is the nth word in the document. Also, compute centroid vector of each class and then calculate the Euclidean distance between document vector and centroid vector of each class. And assign class to the document that is having minimum distance from Centroid vector of particular class [17].

Table 3. Centroid Based Punjabi Text Classification

Training Set	Step1: Compute class vector for each class c in the training set using set of documents belongs to same class c that act as Centroid vector. E.g. $C_{cricket,}$ $C_{Hockey,}$ $C_{Badminton}$..... etc.

Test Set	Step2: Compute C_{doc} vector of the unlabelled document to be classified. Step3: Calculate Euclidean distance between document vector and each class vector. Step4: Assign class to the document that is having minimum distance from that class vector. E.g. let there are only two classes Hockey and Cricket. Euclidean distance between each class and unlabelled document is $C_{cricket,doc}$ = 2.33 and$C_{Hockey,doc}$ =3.15. As $C_{cricket,doc}$ has minimum distance, class Cricket is assigned to the unlabelled document.
Input	ਰਾਜਵੰਤ ਸਿੰਘ ਹਾਕੀ ਜਗਤ ਦਾ ਉਹ ਹਸਤਾਖਰ ਹੈ, ਜਿਸ ਨੇ ਬਿਨਾਂ ਕਿਸੇ ਸਰਕਾਰੀ ਮਦਦ ਤੋਂ ਬਠਿੰਡਾ ਜ਼ਿਲ੍ਹੇ 'ਚ ਹਾਕੀ (rājvant siṅgh hākī jagat dā uh hastākhar hai, jis nē bināṃ kisē sarkārī madad tōṃ baṭhiṇḍā zilhē 'c hākī)
Preprocessing and Feature Extraction Phase	ਹਾਕੀ ਜਗਤ ਹਸਤਾਖਰ ਸਰਕਾਰੀ ਮਦਦ ਹਾਕੀ (hākī jagat hastākhar sarkārī madad hākī)
Output	Class: ਹਾਕੀ (hākī)

3.3 Hybrid Approach

A new hybrid approach is proposed for Punjabi Text Classification which is a combination of Naïve Bayes and Ontology Based Classification techniques to classify the Punjabi text Documents more accurately and efficiently. Here, Naïve Bayes is used as Feature Extraction method for text classification and then Ontology based classification algorithm is performed on extracted features. Class Discriminating Measure (CDM), a feature evaluation metric using Naïve Bayes is used to extract relevant features from the document. This approach helps in reducing the dimensionality of the document and feature space, which helps in assigning class to the document in lesser time with less computation [18]. Features or words having CDM value less than threshold value is selected as feature and neglecting the others. And then Ontology Based Classifier is used to classify the documents. The main advantages of the Ontology Based Classification (Domain specific) is 1) Traditional statistics based text classification methods consider that terms are independent of each other and there are no semantic relations among them. Ontology makes up this drawback. 2) We do not need Training Data i.e. Labelled Documents to classify the documents, whereas other Classification Techniques such as Naïve Bayes Algorithm, Centroid Based Classification, Association Based Classification etc. need Training Set or Labeled Documents to train the classifier for classifying unlabelled documents. 3) Classification on the basis of Domain Specific Ontology is in the minority [19].

For Sports Ontology Based Classification, lists are prepared for each sports class that contains its related terms e.g. Class ਕ੍ਰਿਕਟ (Cricket) contain terms like ਬੱਲੇਬਾਜ਼ੀ (ballēbāzī), ਗੇਂਦਬਾਜ਼ੀ (gēndbāzī), ਫੀਲਡਿੰਗ (phīlḍiṅg), ਵਿਕਟ (vikaṭ), ਸਪਿਨ (sapin), ਆਊਟ (āūṭ), ਵਿਕਟਕੀਪਰ (vikṭakīpar) etc. Class ਫੁਟਬਾਲ (Football) contain ਗੋਲਕੀਪਰ (gōlkīpar),ਫਾਰਵਰਡ (phārvaraḍ), ਡਿਫੈਂਡਰ (ḍiphaiṇḍar),ਮਿਡਫੀਲਡਰ (miḍphīlḍar),ਲਾਈਨਮੈਨ (lāīnmain) etc. This is the first time such lists are prepared manually for Punjabi Language, as no such resources are available in electronic format for this language. These lists, if included in the training set, improve the performance of the classifier.

Table 4. Hybrid Approach for Punjabi Text Classification

Training Set	Step1: Compute class vector for each class i.e $C_{cricket,}$ $C_{Hockey,}$ $C_{Badminton}$….. etc.
Test Set	Step2: For each word in the document to be classified, calculate CDM value of it. CDM (w) = $(P(w \mid C_{Cricket})/1- P(w \mid C_{Cricket})) + (P(w \mid C_{Hockey})/1- P(w \mid C_{Hockey})) +$…….. where P(w|Ci)is the probability of the word occur in class Ci Step3: Each word that is having CDM value less than the Threshold value is ignored. And Remaining words are used by Ontology Based Classifier. Step4: Calculate similarity between Document vector and each class vector. Step5: Assign that class to the document that is having maximum similarity with document vector.
Input	ਰਾਜਵੰਤ ਸਿੰਘ ਹਾਕੀ ਜਗਤ ਦਾ ਉਹ ਹਸਤਾਖਰ ਹੈ, ਜਿਸ ਨੇ ਬਿਨਾਂ ਕਿਸੇ ਸਰਕਾਰੀ ਮਦਦ ਤੋਂ ਬਠਿੰਡਾ ਜ਼ਿਲ੍ਹੇ 'ਚ ਹਾਕੀ (rājvant siṅgh hākī jagat dā uh hastākhar hai, jis nē bināṃ kisē sarkārī madad tōṃ baṭhiṇḍā zilhē 'c hākī)
Preprocessing Phase	ਹਾਕੀ ਜਗਤ ਹਸਤਾਖਰ ਸਰਕਾਰੀ ਮਦਦ ਹਾਕੀ (hākī jagat hastākhar sarkārī madad hākī)
Feature Extraction using Naïve Bayes	ਹਾਕੀ ਹਾਕੀ (hākī hākī)
Output	Class: ਹਾਕੀ (hākī), means input text belongs to class Hockey

3.4 Dataset

The corpus used for Punjabi Text Classification contain 184 Punjabi text documents, from which 50 files are used as Training Data and rest of the files are used as Test Data. Training set contains total 3313 words that are used to train the Punjabi Text Classifier Naïve Bayes and Centroid Based Classification. All the documents are sports related and taken from the Punjabi News Web Sources such as likhari.org, jagbani.com, ajitweekly.com. The unlabelled documents are classified into 7 classes: ਕ੍ਰਿਕਟ (krikaṭ), ਹਾਕੀ (hākī), ਕਬੱਡੀ (kabḍḍī), ਫੁਟਬਾਲ (phuṭbāl), ਟੈਨਿਸ (ṭainis), ਬੈਡਮਿੰਟਨ (baiḍmiṇṭan), ਓਲੰਪਿਕ (ōlmpik).

4. EXPERIMENT

In this experiment, F-score for each class is calculated for each classifier.

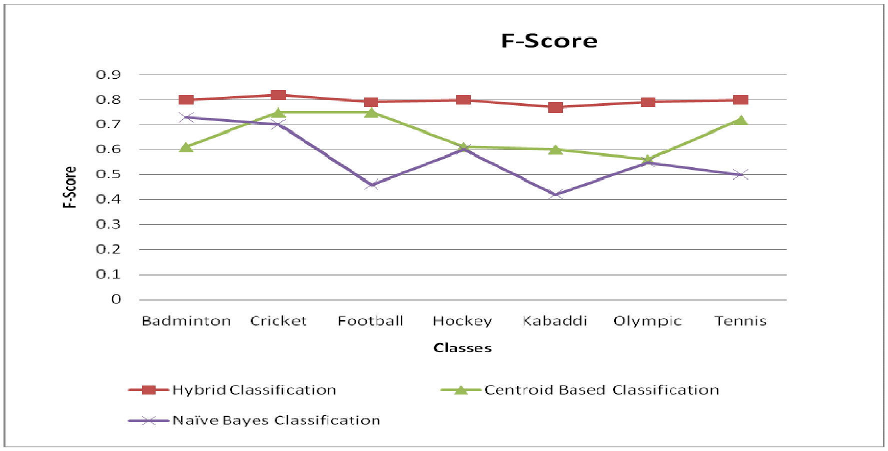

Fig.1. F-Score

From figure 1, it can be observed that Hybrid Classification gives better result in comparison to Centroid Based Classifier and Naïve Bayes Classifier that shows comparatively low results. This is due to the reason that after the removal of stopwords, system cannot find important features that increase the classification rate. And results can be improved by incorporating more linguistic features and rules to remove non-relevant features.

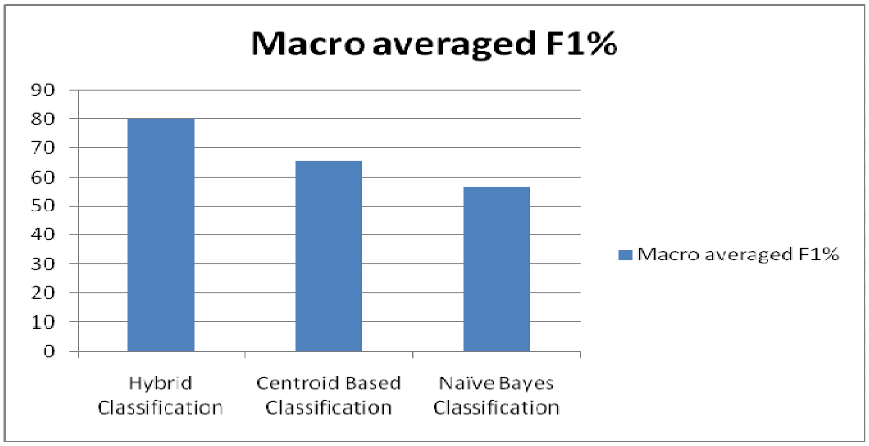

Fig.2. Macro-averaged F-score (F1) of the classifier

The results show that Hybrid Classification, achieves 80% F1, where Centroid Based Classifier and Naive Bayes achieves only 66% and 57% respectively.

5. CONCLUSION

No Punjabi Text Classifier is available in the world to classify the Punjabi Text Documents. This is the first time, three classification techniques presented for Punjabi Text Classification that are Naïve Bayes, Centroid Based and one new approach called Hybrid Approach which is the combination of Naïve Bayes and Ontology Based Classification Techniques proposed for this language. From these three algorithms, Hybrid Approach has better performance than others two, as features extracted with Naïve Bayes are less in count than others two which results in less computations and less time consuming. The Sports Based Ontology prepared for Punjabi Text Classification, contains sports related terms of each predefined classes. This Ontology can be further improved by including the terms of others sports, hence results in better performance of the classifier.

REFERENCES

[1] Nawei Chen and Dorothea Blostein (2006), "A survey of document image classification: problem statement, classifier architecture and performance evaluation", *Springer-Verlag*, DOI= 10.1007/s10032-006-0020-2.

[2] Christoph Goller, Joachim Löning, Thilo Will and Werner Wolf (2009), "Automatic Document Classification: A thorough Evaluation of various Methods", DOI=10.1.1.90.966.

[3] Gupta, Vishal and Lehal, Gurpreet S. (2009), "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence*, VOL. 1, NO. 1.

[4] Han, Jiawei and Kamber, Michelin (2001), "Data Mining Concepts and Techniques", *Morgan Kaufmann*, USA, 70-181.

[5] Gupta, Megha and Aggrawal Naveen (2010), "Classification Techniques Analysis", *NCCI 2010 - National Conference on Computational Instrumentation,* CSIO Chandigarh, INDIA, pp. 128-131.

[6] Gupta, Vishal and Lehal, Gurpreet Singh (2011)," Preprocessing Phase of Punjabi Language Text Summarization", *Information Systems for Indian Languages, Communications in Computer and Information Science, Vol. 139, Springer-Verlag*, pp. 250-253.

[7] Gupta, Vishal and Lehal, Gurpreet Singh (2011), "Punjabi Language Stemmer for nouns and proper name", *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP*, Chiang Mai, Thailand, pp. 35–39.

[8] Guoshi Wu, Kaiping Liu (2009), "Research on Text Classification Algorithm by Combining Statistical and Ontology Methods", *International Conference on Computational Intelligence and Software Engineering, IEEE.* DOI= 10.1109/CISE.2009.5363406

[9] kaur, Darvinder and Gupta, Vishal (2010), "A survey of Named Entity Recognition in English and other Indian Languages", *IJCSI International Journal of Computer Science Issues*, Vol. 7, Issue 6.

[10] Kavi Narayana Murthy (2003), "Automatic Categorization of Telugu News Articles", *Department of Computer and Information Sciences*, University of Hyderabad, Hyderabad, DOI= *202.41.85.68.*

[11] Mohanty, S., Santi, P. K. Mishra, Ranjeeta, Mohapatra, R. N. and Sabyasachi Swain (2006), "Semantic Based Text Classification Using WordNets: Indian Language Perspective", *The Third International Wordnet Conference (GWC 06).* DOI=10.1.1.134.866.

[12] Ali, Abbas Raza and Ijaz, Maliha (2009), "Urdu Text Classification", *FIT '09 Proceedings of the 7th International Conference on Frontiers of Information Technology, ACM* New York, USA. ISBN: 978-1-60558-642-7 DOI= 10.1145/1838002.1838025.

[13] Munirul Mansur, Naushad UzZaman, Mumit Khan, "Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus".

[14] Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., Palaniappan, B. (2009), "Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network" *Expert Systems with Applications, Elsevier*, Volume 36 Issue 8, October, DOI= 10.1016/j.eswa.2009.02.010.

[15] Jayashree, R. (2011) An analysis of sentence level text classification for the Kannada language, " *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*", pp. 147-151.

[16] Lam Hong Lee, Dino Isa (2010), "Automatically computed document dependent weighting factor facility for Naïve Bayes classification", E*xpert Systems with Applications, Elsevier*, DOI=10.1016/j.eswa.2010.05.030.

[17] Lifei Chen, Yanfang Ye, Yanfang Ye (2008), “A New Centroid-Based Classifier for Text Categorization”, *22nd International Conference on Advanced Information Networking and Applications, IEEE*, DOI= 10.1109/WAINA.2008.12.

[18] Jingnian Chen, Houkuan Huang, Shengfeng Tian and Youli Qu (2009), “Feature selection for text classification with Naïve Bayes”, *Expert Systems with Applications: An International Journal*, Volume 36 Issue 3, Elsevier.

[19] Guoshi Wu, Kaiping Liu (2009), “Research on Text Classification Algorithm by Combining Statistical and Ontology Methods”, *IEEE International Conference on Computational Intelligence and Software Engineering*, 11-13 Dec. 2009, Pages 1-4, DOI= 10.1109./CISE.2009.5363406.

[20] Verma, Rajesh Kumar and lehal, Gurpreet Singh. Gurmukhi-Roman Transliterator GTrans version 1.0, http://www.learnpunjabi.org/gtrans/index.asp