Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach

Nidhi, Vishal Gupta
University Institute of Engineering and Technology, Panjab University
naseeb.nidhi@gmail.com, vishal@pu.ac.in

ABSTRACT

Classification of text documents become a need in today's world due to increase in the availability of electronic data over internet. Till now, no text classifier is available for the classification of Punjabi documents. The objective of the work is to find best Punjabi Text Classifier for Punjabi language. Two new algorithms, Ontology Based Classification and Hybrid Approach (which is the combination of Naïve Bayes and Ontology Based Classification) are proposed for Punjabi Text Classification. A corpus of 180 Punjabi News Articles is used for training and testing purpose of the classifier. The experimental results conclude that Ontology Based Classification (85%) and Hybrid Approach (85%) provide better results in comparison to standard classification algorithms, Centroid Based Classification (71%) and Naïve Bayes Classification (64%).

KEYWORDS: Punjabi Text Classification, Ontology Based Classification, Naive Bayes Classification, Centroid Based Classification.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 109–122,
COLING 2012, Mumbai, December 2012.

1. Introduction

The amount of electronic data available such as digital libraries, blogs, electronic newspaper, electronic publications, emails, electronic books is increasing rapidly. However, as the volume of electronic data increases the challenge to manage that data also increases. Thus, automatic organization of text documents becomes an important research issue of Text Mining. Manual Text Classification is an expensive and time consuming task, as classifying millions of text documents efficiently and with accuracy is not an easy task. Therefore, automatic text classifier is constructed whose accuracy and time efficiency is much better than manual text classification. There are mainly two machine learning approaches to enhance classification task: supervised approach, where predefined classes are assigned to text documents with the help of labeled documents; and unsupervised approach, that does not require labeled documents to classify the text documents [Chen and Blostein 2006].

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. Punjabi is an Indo-Aryan Language, spoken in both western Punjab (Pakistan) and eastern Punjab (India). It is 10th most widely spoken language in the world. Also it is the official language of Indian state of Punjab. In comparison to English language, Punjabi language has rich inflectional morphology. E.g. English verb "Play" has 4 inflectional forms: play, played, playing, plays; whereas same word in Punjabi i.e. "ਖੇਡ" has 47 inflectional forms: ਖੇਡ, ਖੇਡਣਾ, ਖੇਡਣੇ, ਖੇਡਣ, ਖੇਡਦਾ, ਖੇਡਦੇ, ਖੇਡਦੀ, ਖੇਡਦੀਆਂ etc. This depends upon gender, number, person, tense, phase, transitivity values in a sentence. Therefore, using only statistical approaches to classify the Punjabi documents do not provide good classification results, there is a need of linguistic approaches too for the selection of the relevant features that increase the efficiency of the classification.

No work has been done to classify the Punjabi documents due to lack of resources available in electronic format. Therefore, two new approaches are proposed for Punjabi language, Ontology Based Classification and Hybrid Approach. And to compare the results of these classification algorithms, standard classification algorithms, Naïve Bayes and Centroid Based Classification are used.

2. Text Classification for Indian Languages

So far Text Classification Techniques implemented for Indian Languages are: Naïve Bayes classifier has been performed over Telugu News articles in four major classes: Politics, Sports, Business and Cinema; to about 800 documents. In this, normalized TFXIDF is used to extract the features. Without any stopword removal and morphological analysis, at the threshold of 0.03, the classifier gives 93% precision [Murthy 2003].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 [Mohanty et al. 2006]. 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 is done 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 classification results of Support Vector Machines is much better than Naïve Bayes [Ali and Ijaz 2009]. For Bangla Text Classification, n-gram based classification 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 using gram length more than 3 decreases the performance of the classifier [Mansur et al. 2006]. Text classification is done using Vector Space Model and Artificial Neural network on morphological rich Dravidian classical language Tamil. 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 [Rajan et al. 2009]. A new technique called Sentence level Classification is used for Kannada language; in this, main focus is on sentences as most users' comments, queries, opinions etc are expressed in sentences. This Technique extended further to sentiment classification, Question Answering, Text Summarization and also for customer reviews in Kannada Blogs [Jayashree R. 2011].

3. Problem Description

Manually classifying millions of documents require lots of effort and time. Therefore, automatic Punjabi Text Classifier is constructed to manage these documents efficiently in less time with minimum efforts. In literature review, it has been observed that no work has been done in this context for Punjabi Language.

The problem is to classify large collection of Punjabi Text Documents into one of the predefined classes. These classes are: ਕ੍ਰਿਕਟ (krikaṭ) (Cricket), ਹਾਕੀ (hākī) (Hockey), ਕਬੱਡੀ (kabḍḍī) (Kabaddi), ਫੁਟਬਾਲ (phuṭbāl) (Football), ਟੈਨਿਸ (ṭainis) (Tennis), ਬੈਡਮਿੰਟਨ (baiḍmiṇṭan) (Badminton), ਉਲੰਪਿਕ (ōlmpik) (Olympic).

The tasks that need to do to achieve the objectives are following:

- To develop NER system and language dependent rules to extract names, locations, time/date, designation, abbreviation, numbers/counting from the document.
- To extract the relevant features from the document for better classification results.
- To prepare gazetteer lists for the classification task.
- To create Domain (sports) specific Ontology for the Punjabi language that includes terms related to class.

4. Punjabi Text Classification

For Punjabi Text Classification, initial steps that need to do are following:

- Prepare Training Set for Naïve Bayes and Centroid Based Classifier. The documents in the training set are tokenized and preprocessed. Stopwords, punctuations, special symbols, name entities are extracted from the document.
- For each class, centroid vectors are created so as to calculate the similarity between centroid vectors and unlabelled document vector.

- Create Sports specific Ontology in Punjabi language for Ontology based Classifier that contains terms related to the class e.g. class ਕ੍ਰਿਕਟ (Cricket) contains terms such as ਬੱਲੇਬਾਜ਼ੀ (ballēbāzī) (batting), ਗੇਂਦਬਾਜ਼ੀ (gēndbāzī) (bowling), ਫੀਲਡਿੰਗ (phīlḍiṅg) (fielding), ਵਿਕਟ (vikaṭ) (wicket), ਸਪਿਨ (sapin) (spin), ਆਊਟ (āūṭ), ਵਿਕਟਕੀਪਰ (vikṭakīpar) (wicket-keeper) etc. For the first time such lists are created for Punjabi Language.
- Prepare Gazetteer lists such as 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.

After initial steps, Punjabi Text Classification is implemented into three main phases:

- Preprocessing Phase
- Feature Extraction Phase
- Processing Phase

4.1 Pre-processing Phase

Each Unlabelled Punjabi Text Documents are represented as “Bag of Words”. Before processing, stopwords, special symbols, punctuations (<,>, :,{,},[,],^,&,*,(,) etc.) are removed from the documents, as they are irrelevant to the classification task. Stopwords list is manually prepared by analysing the punjabi corpus. Table 1 shows lists of some stopwords that are removed from the document.

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

TABLE 1- Stopwords List

4.2 Feature Extraction Phase

Input document may contain redundant or non-relevant data that increases the computations. Therefore, to reduce the feature space by extracting the relevant features from the document, along with statistical approaches, language dependent rules, gazetteer lists are created by analyzing the Punjabi documents.

4.2.1 Statistical Approaches

For classification of Punjabi Documents TFXIDF is used as statistical approach [Yanbo et al. 2009; Han and Kamber 2001]. TFXIDF weighting is the most common method used for term weighting that takes into account this property. In this approach, the TFXIDF weight of term i in document d assigned proportionally to the number of times

the term appears in the document, and in inverse proportion to the number of documents in the corpus in which the term appears.

$$W(i) = tf(i)*\log(N/N_i)$$

TFXIDF weighting approach weights the frequency of a term in a document with a factor that discounts its importance if it appears in most of the documents, as in this case the term is assumed to have little discriminating power.

4.2.2 Linguistic Features

Statistical approach fails to extract language dependent features. Therefore, Hybrid Approach is used which is the combination of Rule Based Approach and List Lookup approach. A number of rules specific for Punjabi language used to extract the language dependent features are following:

1. Name Rule
   a. if word is found, its previous word is checked for middle name.
   b. if middle name is found, its previous word is extracted as first name from the document.
   c. Otherwise, word is extracted from the document.
2. Location Rules
   a. if word is found, it is extracted from the document.
   b. if Punjabi word ਵਿਖੇ (vikhē) is found, its previous word is extracted as location name.
   c. if Punjabi word ਪਿੰਡ (piṇḍ) is found, its next word is extracted as location name.
   d. if Punjabi word ਜ਼ਿਲ੍ਹੇ (zilhē) is found, its previous word is extracted as location name.
3. Date/Time Rules
   a. if month is found, it is extracted.
   b. if week day is found, it is extracted.
   c. if Punjabi words ਅੱਜ (ajj), ਕੱਲ (kall), ਸਵੇਰ (savēr), ਸ਼ਾਮ (shāmm), ਦੁਪਹਿਰ (duphir) etc. are found, they are extracted.
4. Numbers/Counting
   a. if any numeric character is found, it is extracted.
   b. if Punjabi words ਇੱਕ (ikk), ਦੂਜਾ (dūjā), ਦੋ (dō), ਪਹਿਲਾ (pahilā), ਛੇਵੀਂ (chēvīṃ) etc. are found, they are extracted.
5. Designation Rule
   c. if designation found e.g. ਕਪਤਾਨ (kaptān), ਕੋਚ (kōc), ਕੈਪਟਨ (kaipṭan), it is extracted.
6. Abbreviation
   d. if words like ਆਈ (āī), ਸੀ (sī), ਐਲ (ail), ਪੀ (pī), ਬੀ (bī) etc. are found, they are extracted.

4.2.3 Gazetteer Lists

Due to limited resources available in electronic format for Punjabi language, gazetteer lists are prepared manually by analyzing the Punjabi Text documents. Each gazetteer list contains 100-150 words in it. These lists are following:

- Middle Names
- Last names
- Location Names
- Month Names
- Day Names
- Designation names
- Number/Counting
- Abbreviations
- Stop words
- Sports Specific Ontology (e.g. preparing list for class ਹਾਕੀ (Hockey) that contain all of its related terms like ਸਟਰਾਈਕਰ (Striker), ਡਰਿਬਲਰ (Dribbler), ਪੈਨਲਟੀ (Penalty) etc.

4.3 Processing Phase

At this phase, apply classification algorithm such as Naïve Bayes, Centroid Based, Ontology Based and Hybrid Approach to relevant features extracted from feature extraction phase in order to classify the unlabelled document into predefined classes.

4.3.1 Naïve Bayes Classification

It is simple probabilistic classifier that considers each term independent of each other. The two common Naïve Bayes Models used for text classification are: Multinomial Event Model and Multi-variate Bernoulli Event Model. For Punjabi Text Classification, Multinomial Event Model is used as it performs better than Multi-variate Bernoulli event model [McCallum and Nigam 1998; Chen et al. 2009]. In this, each document is represented as "bag of words". Following steps are taken to classify the Punjabi text documents using Naïve Bayes Classifier:

Step 1: Training Set
Prepare training set for the classifier in which folders represent class and each folder contains set of documents called labeled documents. Punctuations, special symbols are removed from the document. Then, documents are segmented into meaningful units called words. Stopwords, name entities such as names, locations, date/time, counting etc. are removed from the document as they are irrelevant to the classification task. Calculate probability of each class P(Ci), using equation (1)

$$P(Ci) = (\text{Total docs in Ci}) / (\text{total docs in training set}) \quad (1)$$

Step 2: Test Set
After preprocessing and feature extraction steps, each unlabelled document are represented as list of words i.e. w1, w2 ....wn, where wn is the nth word of the document. Calculate probability of the document to belong to the particular class using equation (2).

$$P(Ci|document) = (P(Ci|w1, w2.....wn)/ n \quad (2)$$

Where n is the total word in the input document.
Assign class Ci to the document if it has maximum posterior probability with that class.

$$P(Ci|document) = \max (P(Ci)*P(wj|Ci))/ n$$

Where

$P(wj|Ci) = (1+\text{freq. of wj in class Ci})/(\text{total words in Ci + total words in training set})$

4.3.2 Centroid Based Classification

Centroid based classifier is simple and efficient method for the classification task. Its basic idea is to construct Centroid vector per class using training set and document vector. And then calculate the distance between each Centroid vector and document vector; assign that class to the document that is having minimum distance from the Centroid vector [Chen et al. 2008]. Following are steps done for classifying Punjabi Text Documents are:

Step 1: Training Set: After preprocessing and feature extraction phase, Centroid vector for each class is calculated using labeled documents in that class. These vectors are: $C_{cricket}$, C_{hockey}, $C_{badminton}$.....etc.
Step 2: Test Set: For each unlabelled document, calculate document vector C_{doc}, where each component of the vector is represented by TFXIDF value i.e.C_{doc}=[$tfidf_1$, $tfidf_2$.....$tfidf_{|V|}$] ;|V| is total dimensions of the collection.
Step 3: Euclidean Distance: Calculate Euclidean distance between C_{doc} and Centroid vector of each class. And assign that class to the document that is having minimum distance from Centroid vector of that class. If d and c are two vectors, Euclidean distance is calculated as in equation (3)

$$\text{Distance}(d,c) = \text{sqrt}((di- ci) \text{ sup } 2) \quad \text{where } i=1, 2.....7 \quad (3)$$

E.g. assume 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 [Chen and Ye 2008].

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 & Feature Extraction	ਹਾਕੀ ਜਗਤ ਹਸਤਾਖਰ ਸਰਕਾਰੀ ਮਦਦ ਹਾਕੀ (hākī jagat hastākhar sarkārī madad hākī)
Output	Class: ਹਾਕੀ (hākī)

TABLE 2- An example of Centroid Based Classification

4.3.3 Ontology Based Classification

Traditional Classification methods ignore relationship between words, they consider each term independent of each result. But, in fact, there exist a semantic relation between terms such as synonym, hyponymy etc. [Wu and Liu 2009]. Therefore, for

better classification results, there is need to understand the context of the text document. The Ontology has different meaning for different users, in this classification task, Ontology stores words that are related to particular sport. Therefore, with the use of domain specific ontology, it becomes easy to classify the documents even if the document does not contain the class name in it. The ontology is created manually that contains Badminton, Cricket, Football, Hockey, Kabaddi, Olympic and Tennis as their top classes. To create ontology, transliteration and English to Hindi, Hindi to Punjabi translators are used. Each ontology class contain 80-90 terms related to that class. The advantage of using Ontology is that there is no requirement of training set i.e. labeled documents. Hence, no human input is required to create training set. Also this ontology can be used by other researchers for easily classifying Punjabi sports documents with accuracy.

Step 1: Create Domain (i.e. sports) specific ontology, represented as “bag of words”.
Step 2: For each unlabeled document, remove stopwords, punctuations, special symbols, and name entities from the document and represent document as “bag of words”.
Step 3: To determine in which class unlabelled document belongs, calculate the frequency of document terms matched with class ontology. Assign class cricket to the unlabelled document, if frequency of matching terms with class cricket ontology is maximum.
Step 4: If no match is found or a document shows same results for two or more classes then that document is not classified into any class, and left for manual classification.

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 & Feature Extraction	ਹਾਕੀ ਜਗਤ ਹਸਤਾਖਰ ਸਰਕਾਰੀ ਮਦਦ ਹਾਕੀ (hākī jagat hastākhar sarkārī madad hākī)
Output	Class: ਹਾਕੀ (hākī)

TABLE 3- An example of Ontology Based Classification

4.3.4 Hybrid Approach

In hybrid approach, the two algorithms Naïve Bayes and Ontology based Classifier are combined for better results of classification. Using TFXIDF, Information Gain (IG) as feature selection method, results in some features that are still irrelevant. Therefore, Class Discriminating Measure (CDM), a feature evaluation metric for Naïve Bayes that calculates the effectiveness of the feature using probabilities, is used. The results shown in [Chen et al. 2009], indicate that CDM is best feature selection approach than IG. Therefore, instead of using TFXIDF as feature selection method, CDM is used. The term having CDM value less than defined threshold value is ignored. It has been observed that fewer features are left for the computations, this simplifies and speedup the classification task with accuracy. And the remaining terms are used to represent the input unlabelled document; and to match the terms with domain specific ontology, to determine the class of the unlabelled document.

Step 1: For each unlabelled document, remove stopwords, punctuations, special symbols, and name entities from the document and represent document as “bag of words”.
Step 2: For each term in the unlabelled document, calculate CDM for that term using equation (4).

$$CDM(w) = |\log P(w|Ci) - \log P(w|\overline{Ci})| \quad (4)$$

Where P(w|Ci) = probability that word w occurs if class value is i
P(w|$\overline{Ci}$) = probability that word w occurs when class value is not i
i=1 to 7
Step 3: Term having CDM value less than threshold value is ignored. Remaining terms are represented as input document, are used to determine the class of the document.
Step 4: Calculate the frequency of document terms matched with class ontology. Assign class cricket to the unlabelled document, if frequency of matching terms with class cricket ontology is maximum.
Step 5: If no match is found or a document shows same results for two or more classes then that document is not classified into any class, and left for manual classification.

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 (Naive Bayes)	ਹਾਕੀ ਹਾਕੀ (hākī hākī)
Output	Class: ਹਾਕੀ (hākī), means input text belongs to class Hockey

TABLE 4- An example of Hybrid Based Classification

4.4 Performance Measure

In this phase, the performances of the each classifier is evaluated using standard evaluation measures such as Precision, Recall, F-Score, Fallout, Macro averaged Precision, Recall and F-score.

5. Results and Discussions

5.1 Dataset

The corpus used for Punjabi Text Classification contains 180 Punjabi text documents, 45 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 based on Naïve Bayes and Centroid Based Classification techniques. All the documents in the corpus are sports related and taken from the Punjabi News Web Sources such as likhari.org, jagbani.com, ajitweekly.com. Then, classify unlabelled documents into seven classes: ਕ੍ਰਿਕਟ (krikaṭ), ਹਾਕੀ (hākī), ਕਬੱਡੀ (kabḍḍī), ਫੁਟਬਾਲ (phuṭbāl), ਟੈਨਿਸ (ṭainis), ਬੈਡਮਿੰਟਨ (baiḍmiṇṭan), ਉਲੰਪਿਕ (ōlmpik). The system has been implemented using C#.net platform. The stopword list that is prepared manually for the classification task contains 2319 words.

The data structures used for Punjabi Text Classification are files and arrays. Stopwords list, gazetteer lists and sports specific ontology is stored in text file. During the implementation, these lists are stored in arrays to access the contents fast, otherwise, accessing contents directly from the files increase computational time. Even training set documents and test set document are also stored into files. Test Set contains 154 documents that are distributed among seven classes as shown in figure 1.

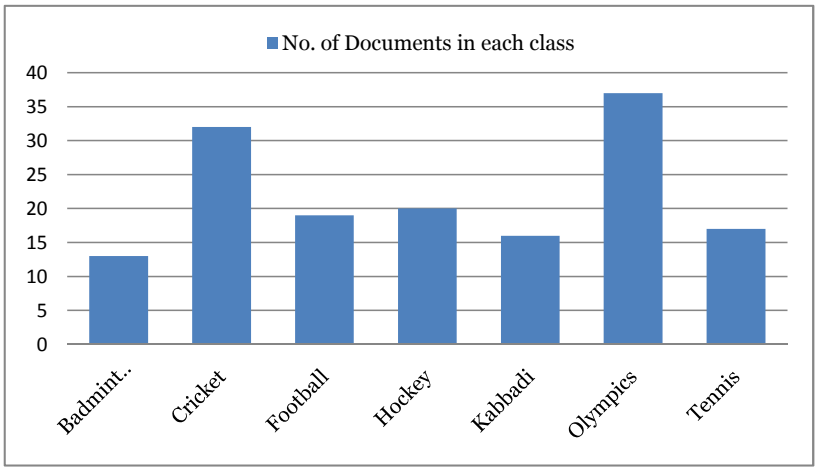

FIGURE 1- Distribution of test set documents

5.2 Screenshot of the system

For classification of Punjabi Text Documents, the dataset of 180 documents are arranged in four sets: Set 1, Set 2, Set 3 and Set 4. Each set contains 40 documents on average. Figure 2 shows main page of the classification system that consists two menu bar items "HELP" and "ABOUT US". In this, there are two buttons "BROWSE" and "APPLY".

"BROWSE" button is used for browsing Punjabi Text Documents that is to be classified. "APPLY" button is used to implement classification algorithm chosen from "COMBOBOX". "COMBOBOX" consists of following items:

1. Ontology Based Classification
2. Naïve Bayes Classification
3. Centroid Based Classification
4. Hybrid Approach

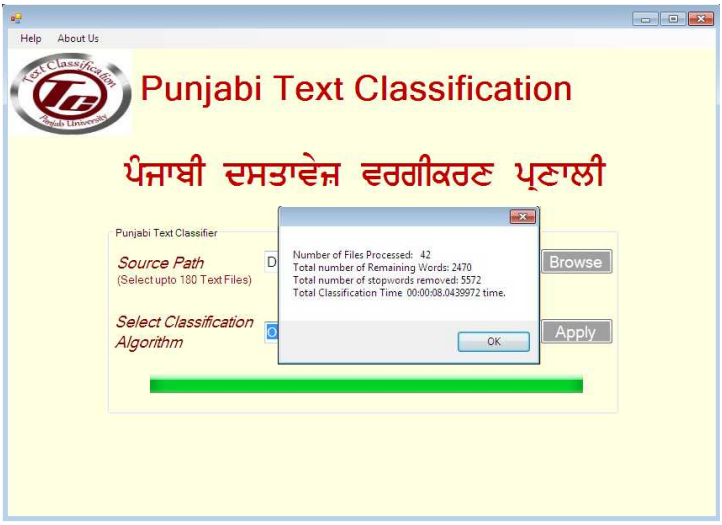

FIGURE 2- Punjabi Text Classifier System

Figure 2 shows the system takes 8 secs 04 ms to classify 42 Punjabi Text Documents. It also gives information about number of stopwords removed and number of words that are left after preprocessing phase.

5.3 Experimental Results

5.3.1 Experiment 1

In experiment 1, F-score for each class is calculated for each classifier using equation (5)

F-Score = (2*Precision*Recall)/ (Precision + Recall) (5)

Precision = (docs correctly classified in class Ci)/ (total docs retrieved in class Ci)

Recall = (docs correctly classified in class Ci)/ (total relevant docs in test set that belong to class Ci)

	Badminton	Cricket	Football	Hockey	Kabaddi	Olympic	Tennis
Ontology Based Classification	0.84	0.89	0.89	0.81	0.88	0.82	0.8
Hybrid	0.83	0.91	0.88	0.84	0.8	0.82	0.88

Classification							
Centroid Based Classification	0.64	0.85	0.8	0.64	0.67	0.56	0.81
Naïve Bayes Classification	0.87	0.77	0.46	0.63	0.42	0.55	0.75

TABLE 5- F-Score of each class using different classification techniques

5.3.2 Experiment 2

In Experiment 2, comparison between classifiers is done, based on non-relevant documents retrieved by each classifier from the total non-relevant documents in the collection as shown in figure 4.

The results show that only 2% from the retrieved documents are non-relevant if Ontology based and Hybrid approach is used to classify the Punjabi Text Documents. Where, in case of Centroid based classifier and Naïve Bayes, 5% and 6% from the retrieved documents are irrelevant, respectively.

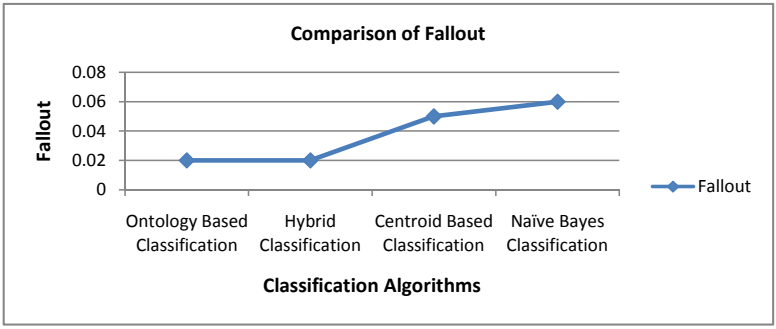

FIGURE 3- Fallout results of each classifier

5.3.3 Experiment 3

The experiment 3 shows the average value of Precision, Recall and F1 for each classifier. From table 6, it can be observed that in comparison with others classifiers, Ontology Based Classification has better averaged Precision (89%) and Recall (85%).

	Ontology Based Classification	Hybrid Classification	Centroid Based Classification	Naïve Bayes Classification
Macro-averaged Precision	0.89	0.88	0.73	0.77
Macro-averaged Recall	0.85	0.84	0.76	0.66

TABLE 6- Macro-averaged Precision and Recall of each classifier

Conclusions

- As of our knowledge, it is first time that we have proposed and implemented two new algorithms for classification of Punjabi documents as previously no other Punjabi document classifier is available in the world.
- Two new algorithms proposed by us are Ontology Based Classification and Hybrid Approach (which is the combination of Naïve Bayes and Ontology Based Classification) for Punjabi documents classification.
- The experimental results conclude that Ontology Based Classification and Hybrid Classification provide better results in comparison to standard classification algorithms Naïve Bayes and Centroid Based for Punjabi documents.
- It is for the first time that sports specific ontology for Punjabi has been developed manually by us, as no such ontology was previously available and it can be beneficial for developing other NLP applications in Punjabi.
- An in depth analysis of Punjabi Corpus is done to prepare gazetteer lists such as middle names, last names, abbreviations, numbers/counting etc. and language dependent rules for Punjabi NER.

References

ALI, ABBAS RAZA AND IJAZ MALIHA (2009). Urdu Text Classification. In: *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.

CHEN JINGNIAN, HUANG HOUKUAN, TIAN SHENGFENG AND QU YOULI (2009). Feature selection for text classification with Naïve Bayes. In: *Expert Systems with Applications: An International Journal*, Volume 36 Issue 3, Elsevier.

CHEN LIFEI, YE YANFANG AND JIANG QINGSHAN (2008). A New Centroid-Based Classifier for Text Categorization. In: *Proceedings of IEEE 22nd International Conference on Advanced Information Networking and Applications*, DOI= 10.1109/WAINA.2008.12.

CHEN NAWEI AND BLOSTEIN DOROTHEA (2006). A survey of document image classification: problem statement, classifier architecture and performance evaluation. In *Springer-Verlag*, DOI=10.1007/s10032-006-0020-2.

HAN JIAWEI AND KAMBER MICHELIN (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2nd edition, USA, 70-181.

JAYASHREE, R. (2011). An analysis of sentence level text classification for the Kannada language. In: *Proceedings of IEEE International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 147-151, DOI=10.1109/SoCPaR.2011.6089130.

MANSUR MUNIRUL, UZZAMAN NAUSHAD AND KHAN MUMIT. Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus. *Center for Research on Bangla Language Processing*, BRAC University, Dhaka, Bangladesh.

McCALLUM, A. AND NIGAM, K. (1998). A comparison of event models for naive Bayes text classification. In: *AAAI-98 workshop on learning for text categorization*. 41-48. Technical Report WS-98-05. AAAI Press.

MOHANTY, S., SANTI, P.K., MISHRA RAJNEETA, MOHAPATRA, R.N. AND SWAIN SABYASACHI . Semantic Based Text Classification Using WordNets: Indian Language Perspective. In: *Proceedings of 3rd International Global Wordnet Conference (GWC 06)*, 321-324, DOI=10.1.1.134.866.

MURTHY, KAVI NARAYANA (2003). Automatic Categorization of Telugu News Articles. In: *Department of Computer and Information Sciences*, University of Hyderabad, Hyderabad, DOI= 202.41.85.68.

PUNJABI LANGUAGE (2012). In: http://en.wikipedia.org/wiki/Punjabi_language.

PUNJABI NEWS CORPUS

RAJAN, K., RAMALINGAM, V., GANESAN, M., PALANIVEL, S. AND PALANIAPPAN, B. (2009). Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network. In: *Expert Systems with Applications*, Elsevier, Volume 36 Issue 8, DOI= 10.1016/j.eswa.2009.02.010.

SUN AIXIN AND LIM Ee-PENG (2001). Hierarchical Text Classification and Evaluation. In: *Proceedings of the 2001 IEEE International Conference on Data Mining(ICDM 2001)*, Pages 521-528, California, USA, November 2001.

VERMA, RAJESH KUMAR AND LEHAL, GURPREET SINGH. Gurmukhi-Roman Transliterator GTrans version 1.0, http://www.learnpunjabi.org/gtrans/index.asp.

WANG, YANBO J., COENEN FRANS AND SANDERSON ROBERT (2009). A Hybrid Statistical Data Pre-processing Approach for Language-Independent Text Classification. In: Proceedings of ADMA 5th International Conference on Advanced Data Mining and Applications. 338-349, DOI=10.1.1.157.6558.

WU GUOSHI AND LIU KAIPING (2009). Research on Text Classification Algorithm by Combining Statistical and Ontology Methods. In: *International Conference on Computational Intelligence and Software Engineering*, IEEE. DOI= 10.1109/CISE.2009.5363406.