# Domain Based Classification of Punjabi Text Documents

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

ABSTRACT

With the dramatic increase in the amount of content available in digital forms gives rise to a problem to manage this online textual data. As a result, it has become a necessary to classify large texts (documents) into specific classes. And Text Classification is a text mining technique which is used to classify the text documents into predefined classes. Most text classification techniques work on the principle of probabilities or matching terms with class name, in order to classify the documents into classes. The objective of this work is to consider the relationship among terms. And for this, Sports Specific Ontology is manually created for the first time. Two new algorithms, Ontology Based Classification and Hybrid Approach are proposed for Punjabi Text Classification. The experimental results conclude that Ontology Based Classification (85%) and Hybrid Approach (85%) provide better results.

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

*Proceedings of COLING 2012: Demonstration Papers*, pages 297–304,
COLING 2012, Mumbai, December 2012.

## 1. Introduction

A review of the literature shows that no prior work has been done to classify the Punjabi documents. Therefore the objective of the work is to develop a system that takes Punjabi text documents as input and classify them into its corresponding classes using classification algorithm selected by user. These classes are: ਕ੍ਰਿਕਟ (krikaṭ) (Cricket), ਹਾਕੀ (hākī) (Hockey), ਕਬੱਡੀ (kabḍḍī) (Kabaddi), ਫੁਟਬਾਲ (phuṭbāl) (Football), ਟੈਨਿਸ (ṭainis) (Tennis), ਬੈਡਮਿੰਟਨ (baiḍminṭan) (Badminton). These classes are defined by analyzing the Punjabi Corpus used for the classification task. And for classifying these documents, two new approaches are proposed for Punjabi language, Ontology Based Classification and Hybrid Approach. For Ontology Based Classification, Sports specific Ontology is manually created in Punjabi Language for the first time that consist of terms related to the class. E.g. Cricket class consists of following terms ਬੱਲੇਬਾਜ਼ੀ (batting), ਗੇਂਦਬਾਜ਼ੀ (bowling), ਫੀਲਡਿੰਗ (fielding), ਵਿਕਟ (wicket), Football class consist of ਗੋਲਕੀਪਰ (Goalkeeper), ਗੋਲ (Goal), ਫਾਰਵਰਡ (Forward), ਮਿਡਫੀਲਡਰ (Mid-fielder), ਡਿਫੈਂਡਰ (Defender) etc. An advantage of using Ontology is, there is no requirement of training set i.e. labeled documents and it can also be beneficial for developing other NLP applications in Punjabi. And to compare the efficiency of proposed algorithms, standard classification algorithms results, Naïve Bayes and Centroid Based Classification are compared using F-score and Fallout.

## 2. Proposed algorithm for Punjabi Text Classification

For Punjabi Text Classification, initials 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 using training set.

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

- Preprocessing Phase
- Feature Extraction Phase
- Processing Phase

### 2.1 Pre-processing Phase

Each Unlabelled Punjabi Text Documents are represented as "Bag of Words". Before classifying, stopwords, special symbols, punctuations (<,>, :,{,},[,],^,&,*,(,) etc.) are removed from the documents, as they are irrelevant to the classification task.  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

## 2.2 Feature Extraction Phase

After pre-processing, input documents still contain redundant or non-relevant features that increase the computations. Therefore, to reduce the feature space, along with statistical approaches, language dependent rules and gazetteer lists are also used by analyzing the documents. TFXIDF weighting is the most common statistical method used for feature extraction [Han J. and Kamber M. 2006] using equation (1).

$$W(i) = tf(i)*\log(N/N_i) \qquad (1)$$

### 2.2.1 Linguistic Features

And to extract the language dependent features, Hybrid Approach is used that include Rule Based Approach and List Lookup approach. A number of rules specific for Punjabi language is formed 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. 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) or ਜ਼ਿਲ੍ਹੇ (zilhē) 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.
3. Date/Time Rules
   a. if month or week day is found, it is extracted.
   b. if Punjabi words ਅੱਜ (ajj), ਕੱਲ (kall), ਸਵੇਰ (savēr), ਸ਼ਾਂਮ (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
   a. if designation found e.g. ਕਪਤਾਨ (kaptān), ਕੋਚ (kōc), ਕੈਪਟਨ (kaipṭan), it is extracted.
6. Abbreviation
   a. if words like ਆਈ (āī), ਸੀ (sī), ਐਲ (ail), ਪੀ (pī), ਬੀ (bī) etc. are found, they are extracted.

### 2.2.2 Gazetteer Lists

Lists prepared for classifying Punjabi Text Documents 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.

## 2.3 Processing Phase

At this phase, classification algorithms are applied as following:

### 2.3.1 Naive Bayes Classification

Multinomial Event Model of Naive Bayes is used [McCallum and Nigam 1998; Chen et al. 2009] and for classification assign class Ci to the document if it has maximum posterior probability with that class.

### 2.3.2 Centroid Based Classification

Calculate the distance between each Centroid vector (c) and document vector (d); assign that class to the document that is having minimum Euclidean distance from the Centroid vector [Chen et al. 2008].

### 2.3.3 Ontology Based Classification

Traditional Classification methods ignore relationship between words. But, in fact, there exist a semantic relation between terms such as synonym, hyponymy etc. [Wu and Liu 2009]. 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. After feature extraction phase, for classification, calculate the frequency of extracted terms that are matched with terms in ontology. E.g. assign class cricket to the unlabelled document, if frequency of matching terms with class cricket ontology is maximum. 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.

### 2.3.4 Hybrid Approach

In hybrid approach, the two algorithms Naïve Bayes and Ontology based Classifier are combined for better results of classification. Using TF, TFXIDF or Information Gain (IG) as feature selection method sometimes results in features that are irrelevant. Therefore, Class Discriminating Measure (CDM), a feature evaluation metric for Naïve Bayes calculates the effectiveness of the feature using probabilities, is used for feature extraction. The results shown in [Chen et al. 2009], indicate that CDM is best feature selection approach than IG. The term having CDM value less than defined threshold value is ignored. And the remaining terms are used to represent the input unlabelled document. CDM for each term is calculated using (2)

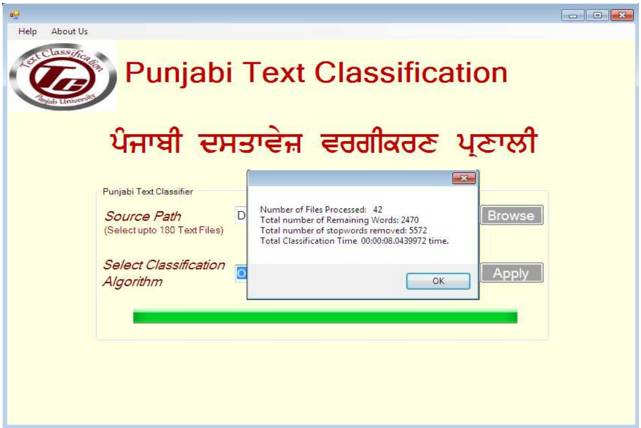
$$CDM(w) = |\log P(w|Ci) - \log P(w|\overline{Ci})| \tag{2}$$

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=1,2,.....6

Calculate the frequency of extracted terms that are matched with terms in ontology. Assign class Badminton to the unlabelled document, if the frequency of matching terms is maximum with class badminton. 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.

### 2.4 Classified Documents

After processing phase, unlabelled documents are classifies into classes.


Help
About Us
Text Classification
Tc
Punjab University
Punjabi Text Classification
ਪੰਜਾਬੀ ਦਸਤਾਵੇਜ਼ ਵਰਗੀਕਰਣ ਪ੍ਰਣਾਲੀ
Punjabi Text Classifier
Source Path
(Select upto 180 Text Files)
Select Classification Algorithm
D
O
Browse
Apply
OK

Number of Files Processed: 42
Total number of Remaining Words: 2470
Total number of stopwords removed: 5572
Total Classification Time 00:00:08.0439972 time.

FIGURE 1- Punjabi Text Classifier System

Figure 1 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.

## 3 Experimental Evaluations

### 3.1 Dataset

The corpus used for Punjabi Text Classification contains 180 Punjabi text documents, 45 files are used as Training Data. Training set contains total 3313 words. All the documents in the corpus are sports related and taken from the Punjabi News Web

Sources such as jagbani.com. The system has been implemented using C#.net platform. The stopword list is prepared manually contains 2319 words. The data structures used are files and arrays. Stopwords list, gazetteer lists and ontology are stored in text file. During the implementation, these files are stored into arrays to read the contents fast.

## 3.2 Experimental Results

F-score [Sun and Lim 2001] for each class is calculated for each classifier using equation (3)

F-Score = (2*Precision*Recall)/ (Precision + Recall) (3)
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 | Tennis |
|---|---|---|---|---|---|---|
| Ontology Based Classification | 0.84 | 0.89 | 0.89 | 0.81 | 0.88 | 0.8 |
| Hybrid Classification | 0.83 | 0.91 | 0.88 | 0.84 | 0.8 | 0.88 |
| Centroid Based Classification | 0.64 | 0.85 | 0.8 | 0.64 | 0.67 | 0.81 |
| Naïve Bayes Classification | 0.87 | 0.77 | 0.46 | 0.63 | 0.42 | 0.75 |

TABLE 2- F-Score of each class using different classification techniques

From Table 2, it is concluded that on average Ontology (85%) and Hybrid Based Classification (85%) shows better results than standard algorithms, Naive Bayes (64%) and Centroid Based Classification (71%) for Punjabi Language. Even the fallout results shows that 2% of the documents retrieved by system are irrelevant in case of Ontology and Hybrid Based Classification where as 5% and 6% non-relevant documents are retrieved if Centroid and Naive Bayes Algorithm are chosen respectively.

## Conclusion

It is first time that two new algorithms Ontology and Hybrid Based Approach are proposed and implemented for classification of Punjabi documents as previously no other Punjabi document classifier is available in the world. The experimental results conclude that Ontology and Hybrid Classification provide better results in comparison to Naïve Bayes and Centroid Based for Punjabi documents.

## References

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.

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

McCALLUM, A. AND NIGAM, K. (1998). A comparison of event models for naive Bayes text classification. In: AAAI98 workshop on learning for text categorization. 41-48. Technical Report WS-98-05. AAAI Press.

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

PUNJABI NEWS CORPUS

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.

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.