# Cross Language Training and Classification


| *Nishant Kumar* | *Anirban Mishra* | *Sayali Degaonkar* | *Pooja Bhandari* |
|---|---|---|---|
| *7305-8548-36* | *6287-3467-22* | *3434-3196-94* | *4146-0109-87* |
| *nishant@usc.edu* | *anirban@usc.edu* | *degaonka@usc.edu* | *poojab@usc.edu* |


## Introduction

Although there has been a lot of research work done in monolingual text classification in various domains, bi-lingual or cross language classification remains to be a less explored area of Natural Language processing. Cross language classification is defined as the task of training the model in one language and then classifying the test data in other language with similar script.

For our research we decided to work on the cross-language news classification for the languages belonging to Devanagari script, which is the basis for over 120 languages. To present a simplified model for the idea we propose to start with Hindi & Marathi. We would be working on real world news archives as the corpus for our project which we will be collecting and annotating manually and classifying into four categories Politics, Sports, Entertainment and Business.

### *Challenges*

The biggest challenge in text classification using cross language training is the wide difference between the words having same or similar meaning. For e.g the word "disappointment" is spelled as "Nirasha" in Hindi and "Nairashye" in Marathi. Also several words in Marathi language are represented by two words in Hindi. For e.g. "from the policy" can be expressed as one word in Marathi ("Yojananche") whereas we need two words to represent it in Hindi ("Yojna se").

This makes it difficult to define the feature used for classification since we can't directly use the word from one language as the feature to be tested on a model trained on the other language.

The other challenge is the availability of corpus and bi-lingual dictionary for these languages which are very hard to find.

### *Related Work*

There has been some work done in cross language categorization using "Expectation Maximizing " [1] and using bilingual dictionary [2] but these works are done on English, Chinese and Italian languages. There has been some work done in Cross language information retrieval in Hindi and Marathi but that has been done using translating both to English [3]

We are planning to learn and build upon these concepts to train and define our own stemming and K-NN algorithm for classifying our data.

## Method

### *Materials*

Since we could not figure out any significant classified corpus related to our work, we are planning to collect data manually. We will start with collecting 250 news articles from each news category for both

Hindi and Marathi. Also we will build our own bi-lingual dictionary for high frequency words in Hindi and Marathi which will be used for classification.

***Features***

Since there is a wide dissimilarly between the two languages under consideration, we will be choosing our features dynamically depending on the scenario. The feature might be exact word from the test data or we might derive our feature by applying K-NN algorithm for each word.

***Procedure***

- Tokenization, removing punctuations and stop words for pre-processing on Hindi news corpus.
- Train the model based on annotated corpus in Hindi.
- Preprocess the Marathi data using the similar approach as Hindi data.
- Try to map the feature exactly as it appears since some of the words in both the languages might be in common.
- Search for the test feature in the dictionary, if not found above.
- If still not found, build the feature cluster using K-NN algorithm and try classifying with the cluster, where k would be dynamically decided based on the feature length
- Apply suitable smoothing as a last resort.
- Apply Naive Bayes classifier to find the most probable classification

***Tools Used*** NLTK and scikit-learn

***Evaluation***

- By Computing Precision, Recall and F-measure evaluate our classifier's performance against the annotated Marathi corpus.
- Performance achieved by the classifiers on other languages as some sort of baseline as no exact statistics available for Hindi and Marathi.

## References

[1] LeiShi, Rada, Mihalcea and Mingjun Tian. 2010, Cross Language Text Classification.
[2] Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting Comparable Corpora and Bilingual Dictionaries for Cross-Language Text Categorization
[3] Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani. 2007. Hindi and Marathi to English Cross Language Information Retrieval.

## Division of Work

***Corpus Annotation****:* Anirban & Nishant (Hindi), Sayali & Pooja (Marathi)
***Background Research:*** Anirban, Sayali
***Algorithm & Feasibility Analysis:*** Nishant, Pooja
***Bilingual-Dictionary formation:*** Pooja, Sayali
***Implementation:*** Anirban, Sayali, Pooja, Nishant
**Documentation*:*** Nishant -25%, Anirban-25%, Sayali- 25%, Pooja- 25%

## Word Count 744