Classification of Hindi authors on the basis of author writing style

Dhruv Anand, 11251[*1] and Srijan R. Shetty,11727[†1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur

April 24, 2014

Abstract

This project is an attempt to study well established methods of author attribution in the context of Hindi Literature. While sophisticated approaches to author attribution have been tested in English, the terrain of Indian languages remains untouched. This is partly because of the morphologically rich structure of the languages and partly because of the absence of an authoritative dataset in Hindi. Our project explorers supervised and unsupervised methods of author attribution in Hindi Literature. The differences in accuracy in results can be related to variance in effectiveness of the common stylistic features in indicating identity of the author. These could vary uniformly from Hindi to English Texts and this project tries to find such differences.

[*]adhruv@iitk.ac.in
[†]srijans@iitk.ac.in

Contents

List of Figures

List of Tables

1 Introduction

Authorship Attribution is a research area in Natural Language Processing and Information Retrieval which attempts to identify the key indicators that set apart various authors from each other. This kind of information helps linguists and literary scholars in understanding the possible influences in a person's writing style and can help improve the overall understanding of human-generated language. It could also help construct systems that generate literature in accordance with a particular person's writing style.

2 Motivation

The motivation to undertake such an analysis is mostly literary in nature, an analysis of the author writing style can give a number of insights on possible reasons of the allure of particular authors. Such insights can help in automated generation of stories in various flavours depending on the chosen author.
That being said, classification of literature on the basis of writing style can also be used to classify stories by unknown authors and in the possible uncovering of authors who write under a different pen name.

3 Corpus

One of the most vital component of our project was a high quality dataset preferably in Unicode. Unfortunately, none of the datasets that we came across met our requirements, hence we decide to build a corpus of Hindi Literature by ourselves.
The data source of our corpus if www.hindisamay.com, which allows all of its content to be downloaded in doc format. For the data to be useful, we had to convert it into text format using LibreOffice in headless mode. The created corpus consists of novels from the following authors: Rabindranath Tagore, Vibhuti Narayan, Premchand, Sarat Chandra Chattopadhyay and Dhamarvir Bharati. The details of the novels have been given in Table 1. (Note: due to the lack of hindi novels by Rabindranath Tagore, we had to resort to using short stories written by him.)

Author	**Novels**	**Tokens**
Premchand	Mangalsootra, Karmabhumi, Alankar, Vardaan, Gaban	328181
Sarat	Srikant, Path ke Daavedar, Dehati Samjh	350598
Vibhuti	Tabadla, Loktantra, Ghar, Shehar mein Curfew	238297
Dharamvir	Gunhao ke Devta, Suraj ka Saatva Ghoda	124461
Tagore	Aankh ki Kirkiri (and stories)	144950

Table 1: Hindisamay Corpus

4 Methodology

4.1 Preprocessing

The documents were normalized by removing all Hindi punctuations. To vectorize the authors, we concatenated the work of each author and divided them into snippets of 500 words each.

4.2 Feature Vector Formation

Unigram

For Unigram analysis, the top 4500 frequent words were used to create a Bag of Words model for each vector created in the pre-processing stage.

Bigram and Trigram Analysis

For the other ngram analysis, the top 2000 frequent ngrams were used to create a Bag of Words model for each vector created in the pre-processing stage.

Multiple Discriminant Analysis

For MDA, we only considered the top 1000 frequency bigrams and top 1000 frequent trigrams which were concatenated in order to create feature vectors.

4.2.1 Dimensionality Reduction using PCA

The high dimensional vectors obtained from the feature vector creation stage were reduced to manageable dimensions, using Principal Component Analysis. This ensured optimum usage of time in the later stages. The following caveats are in order:

- Unigram analysis was done the raw feature vectors. This decision was undertaken because, Unigram merely served to eliminate the authors whose works were translated by multiple translators (Rabindranath Tagore in our case).
- Results have been reported for the top 20 dimensions, albeit analysis was carried out for 5, 10 and 40 dimensions. It was seen that the results remained fairly static until 20 dimensions.

4.3 Classification

4.3.1 Support Vector Machine Classification

We trained an SVM for each of the author separately using the Radial Basis Kernel Function.

4.3.2 K-Means Clustering

The dimensionally reduced feature vectors were clustered using K-Means clustering with the number of clusters set to four. The cluster containing maximum number from a particular author was assumed to be that author's cluster.

4.3.3 Multivariate Discriminant Analysis

As an attempt to improve results further, two features - word bigrams and trigrams - were combined together and then clustered using K-Means clustering.

4.4 Evaluation

4.4.1 Supervised Learning

A test set for each author was kept (such that training:test snippets were in a 5:2 ratio). Evaluation on each one v/s all classifier was done separately and the results were tabulated. Out of the total 2089 snippets, the number of snippets which were used for training and testing have been given in Table 2 for reference.

Author	Training +ve	Training -ve	Test +ve	Test -ve
Premchand	300	1000	358	431
Sarat	500	700	202	687
Vibhuti	300	900	179	710
Dharamvir	200	1000	50	839

Table 2: Test and Training Sets for SVM Classifier

4.4.2 Unsupervised Learning

The cluster with the most snippets actually labelled as written by author A was assigned A. After the assignment of clusters to authors, the misclassfication was computed using the groundtruth. Using these values, various statistics for each clustering was computed.

5 Results

5.1 Supervised Learning

Author	Precision	Recall	F-score
Premchand	0.9325	0.9651	0.9485
Sarat	0.9643	0.9356	0.9497
Vibhuti	0.9933	0.8268	0.9024
Dharamvir	1.0	0.64	0.7805

Table 3: SVM Statistics for Bigrams

5.2 Unigram Analysis

Author	Precision	Recall	F-score
Premchand	0.7177	0.5301	0.6098
Tagore	0.2265	0.735	0.3462
Sarat	0.8219	0.6125	0.7018
Vibhuti	0.7446	0.6626	0.7012
Dharamvir	1.0000	0.7460	0.8545

Table 4: K-Means Statistics for Unigrams

5.3 Bigram Analysis

Author	Precision	Recall	F-score
Premchand	0.9939	0.9894	0.9916
Sarat	0.9982	0.9544	0.9758
Vibhuti	0.8764	0.9916	0.9304
Dharamvir	0.9906	0.848	0.9137

Table 5: K-Means Statistics for Bigrams

5.4 Trigram Analysis

Author	Precision	Recall	F-score
Premchand	0.8715	0.9696	0.9179
Sarat	0.9676	0.8504	0.9052
Vibhuti	0.9102	0.9727	0.9404
Dharamvir	0.8947	0.816	0.8535

Table 6: K-Means Statistics for Trigrams

5.5 Multiple Discriminant Analysis

Author	Precision	Recall	F-score
Premchand	0.9969	0.9878	0.9923
Sarat	0.9935	0.8732	0.9295
Vibhuti	0.9464	0.9958	0.9705
Dharamvir	0.7151	0.904	0.7985

Table 7: K-Means Statistics for MDA

6 Clusters

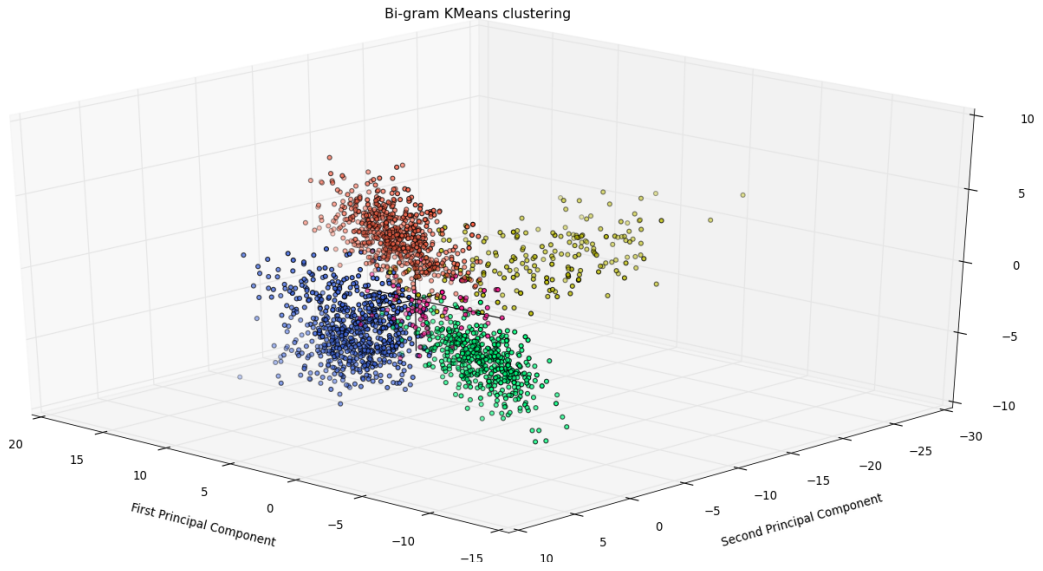

Figure 1: K-Means: Bigrams

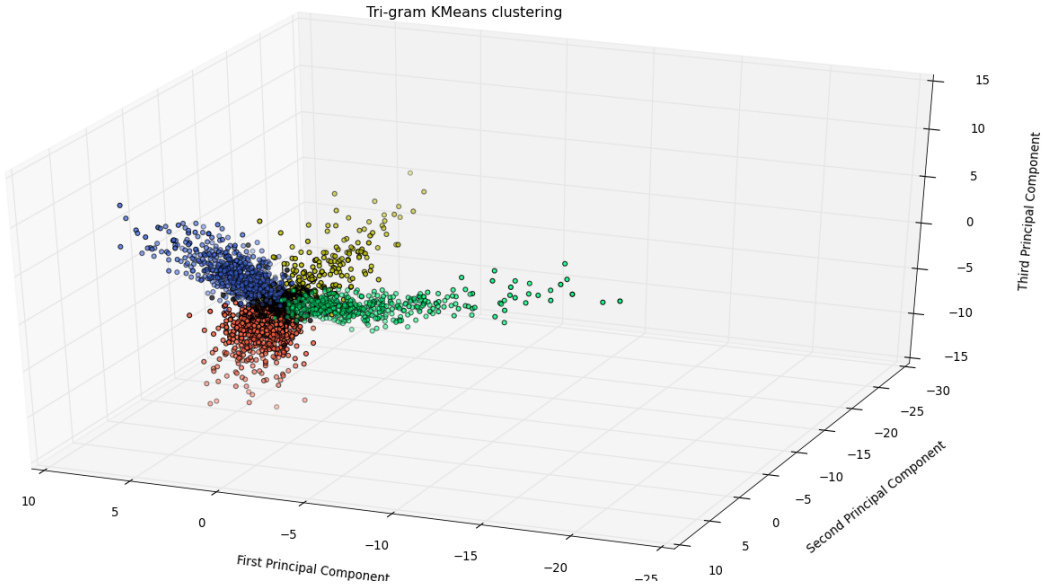

Figure 2: K-Means: Trigrams

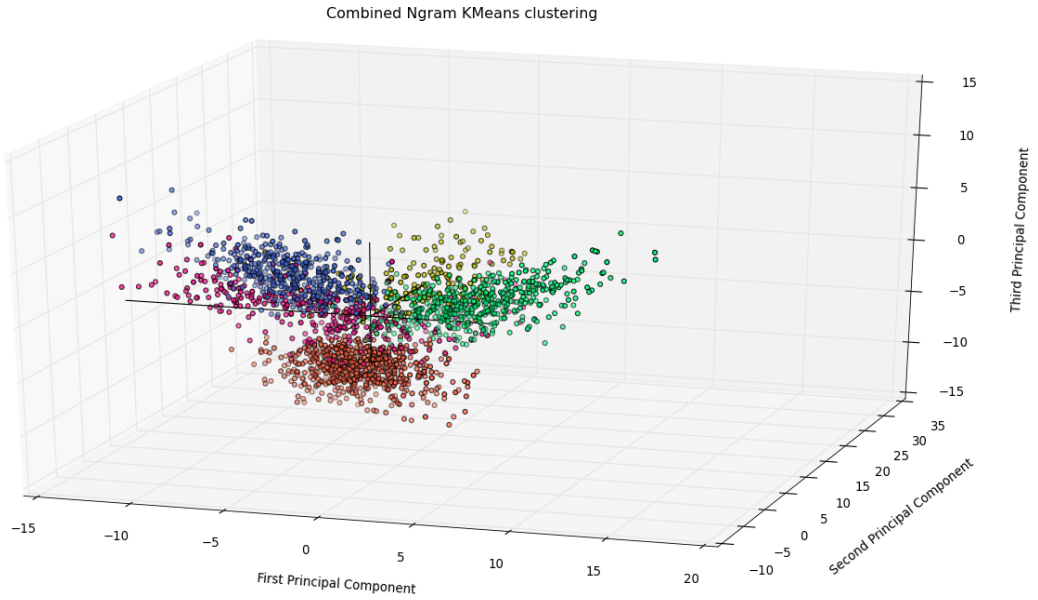

Figure 3: K-Means: MDA

7 Insights

1. Since works of Rabindranath Tagore were translations, their removal from the analysis improved the results since the multiple translators made his work heterogeneous. (The fact that most of his collected works were translations is bolstered by the results Unigram Analysis.)

2. The corpus contained many essays by Vibhuti Narayan Rai. Ergo, his works included many domain specific content words, leading to good results.

3. The Corpus contained only novels for Premchand and so both recall and precision for him were high $> 70\%$

4. The increase in number of classes (authors) will gradually lead to reduction in accuracy of the results as the amount of distinction in the feature vector space will be too less for spherical clusters to be found.

5. The sparsity of the trigram vectors led to a lower F-score for it compared to Bigrams. This indicates that a larger dataset will result in better Precision and Recall for trigrams.

6. MDA did not prove to be too beneficial as the F-score for both features individually was already saturated at more than 90-95%. Thus, the application of MDA did not improve our study very substantially.

8 Conclusion

Our analysis has given us valuable insights on the nature of and on the viability of using standard methods on Indian Literature datasets for Authorship Attribution. Further work on Authorship Attribution in Indian Language datasets can be pursued by augmenting the dataset we have constructed and by running the methods we have used on it. Since unsupervised methods have also proven to work well in our analysis, our approach could work even if the ground truth of authorship is not known. This could greatly benefit people working on plagiarism detection and literary analysts who want to study writing style of Hindi Authors.

9 Code

The code used in this project can be viewed at GitHub.

References

[1] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.

[2] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.

[3] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.