Classification of Hindi Literature based on Author Writing Style

Srijan R. Shetty, 11727 Dhruv Anand, 11251

Department of Computer Science and Engineering,
Indian Institute of Technology, Kanpur

Motivation

- Document Fraud Detection
- Classifying works from unknown authors
- From a Literary perspective
  - Repeating trends of authors
  - Adopting styles of popular authors

Features Used

- Stemmed/non-stemmed unigrams
- Collocations (word bigrams)
- Word Trigrams

Corpus Creation

Due to the dearth of a good Hindi Literature corpus, we had to create a corpus of our own. The corpus was created by scraping hindisamay.com, which contains a sizeable collection of works from Rabindranath Tagore, Premchand, Sarat Chandra, Dharamvir Bharati and Vibhuti Narayan. For each author we had two to three novels of about 90, 000 words each.

Combined Ngram analysis

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 1: SVM Statistics

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 2: Combined Ngram Statistics

Bigram Analysis

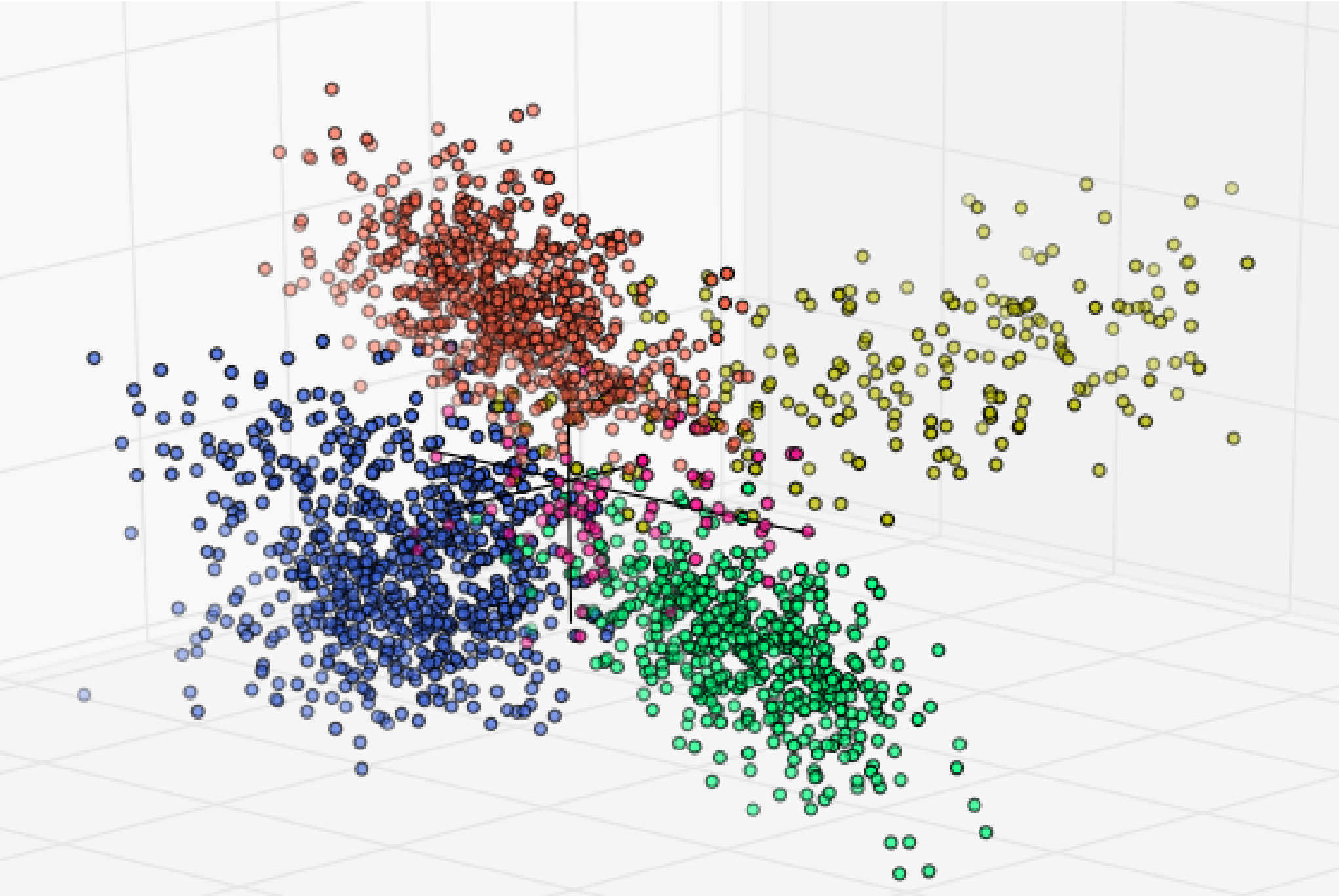

Figure 1: KMeans - bigrams

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 3: Bigram Statistics

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 4: Trigram Statistics

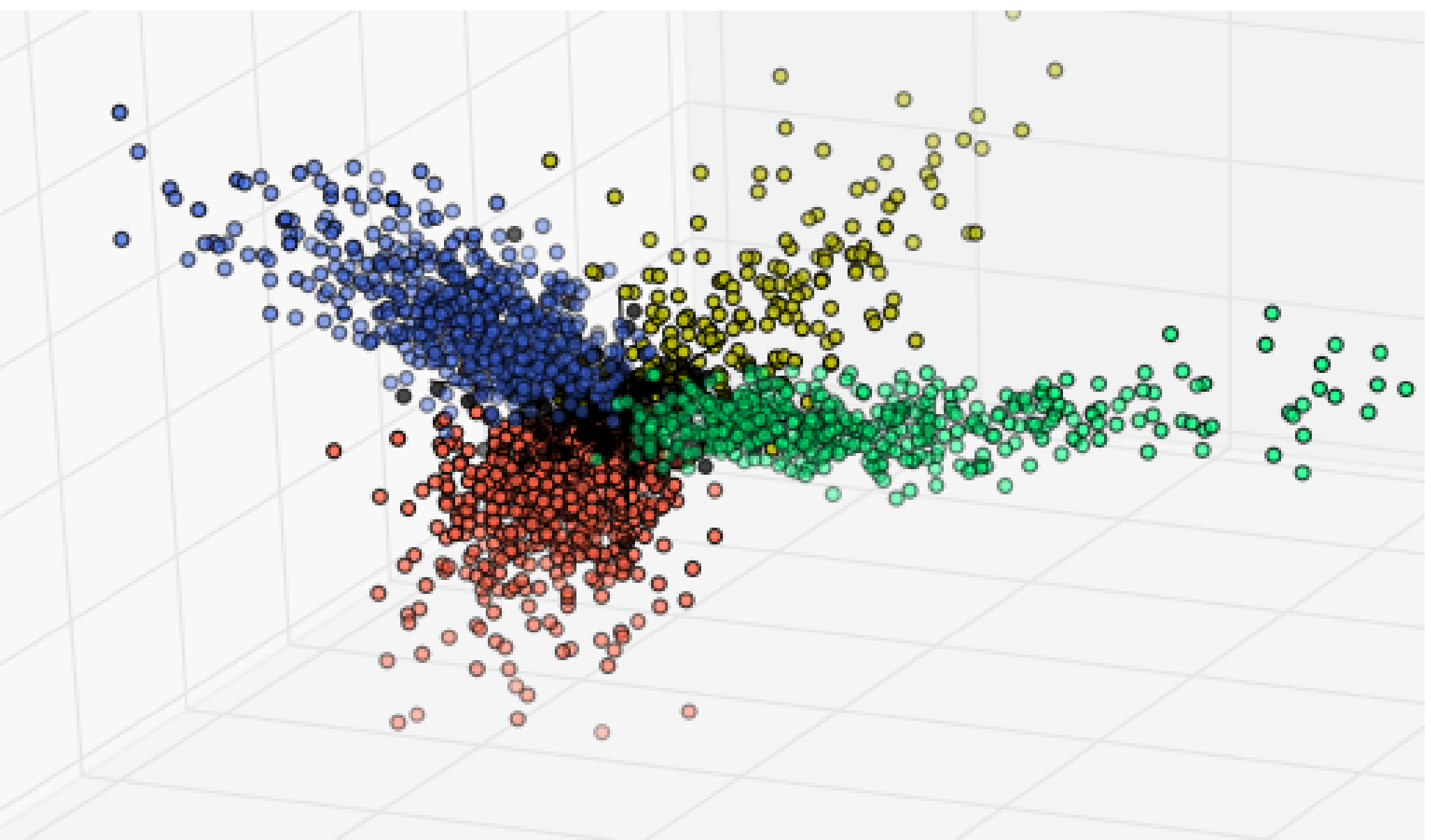

Figure 2: KMeans: trigrams

Methodology

Preprocessing: The document was purged of all non-Hindi letters and each author's work is divided into snippets of 500 words each. (For supervised learning, a random set of snippets from each author's corpus is separated for testing purposes.)

Feature Vector Formation: Frequency vectors for the top 2000 most used collocations constructed for each snippet. Additionally, to carry out Multiple Discriminant Analysis, the top 1000 components of both vectors previously created are concatenated and stored separately for each snippet.

Dimensionality Reduction using PCA: To reduce time spent in clustering, we applied the method of Principal Component Analysis. We took the top 20 components obtained from this process in the next step - Classification.

K-Means Clustering: The obtained feature vectors were put through a generic K-means clustering module to get back 4 classes of snippets. The cluster containing maximum number from a particular author were post

Multivariate Discriminant Analysis: As an attempt to improve results further, we took two features - word bigrams and trigrams together and tried learning the classes based on their combined vectors. This 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.

Support Vector Machine Classification: Supervised learning was done for each author sepaarately. A linear classifier was constructed using Radial Basis Function Kernel. A random set of positive and negative examples (with respect to a particular author) were used for training and the rest were used as test cases.

Insights

- Corpus contained only novels for Premchand and so both recall and precision for him were high > 70
- The corpus contained essays by V.N.Rai, indicating high amount of content words.
- Since works of Rabindranath Tagore were translations, their removal from the analysis improved the results since the multiple translators made his work heterogenous.
- 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.

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.

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.