A Survey of Authorship Attribution Methods for the Hindi Language

Shwetank Shrey
IIIT Delhi
shwetank16095@iiitd.ac.in

Rohan Chhokra
IIIT Delhi
rohan16080@iiitd.ac.in

Abstract

Authorship attribution is one of the oldest problems in linguistics which also became one of the most important ones with the rise of modern statistics. We define Authorship Attribution as an attempt to identify if the testing corpus has been written by the aforementioned author or not using their stylometric fingerprint. Sophisticated techniques exist for the English language however the same cannot be said for Indic languages. Differing in their morphological structure, quite less stylometric analysis has been done for these languages. In this paper, we apply various stylometric techniques for Hindi authors and compare their performances for the following problem: *given a particular sample of text believed to be by one of a set of authors, determine which one is the most likely to write it*.

1 Introduction

Authorship attribution is the science of inferring the author of a document using one or more features in the documents written by the author, called stylometric features. It is an important problem in computational linguistics and information retrieval as well as in applied areas like law and journalism where knowing the author of a document such as a ransom note, for example, may enable law agencies to save lives. The most vanilla form for testing candidate algorithms for authorship attribution is the closed-set authorship attribution task: given a sample of reference documents believed to be written by someone from a finite set of candidate authors, the task is to determine the most likely author of a previously unseen document of unknown authorship.

With the increase in the number of publicly available documents on the internet, an ample amount of work has been done in authorship attribution for European languages such as English, French and Spanish. However, this progress is not shown by other languages where a huge amount of data is not available digitally. In its simplest form, authorship attribution is shown to be a text classification problem in which we simply model the algorithm to recognize the text content. However, authorship attribution also has to take into consideration other factors like the style of the author and the context of the text. Thus, other than the bag of words models traditionally used to train these classifiers, we also have more reliable features like functional lexical features, parts-of-speech n-grams, etc. With such a varied list of features, we aim to perform authorship attribution for languages such as Hindi, which does not have a dataset as accessible as others.

In this paper, we apply popular authorship attribution techniques in the Hindi language and measure if these techniques are relevant for Indic languages as well or not. We use a combination of popular features as training data for our classifiers and survey some classification algorithms trained on this data. We measure the accuracy of these permutations and try to identify an ideal combination for the authorship attribution problem in the Hindi language. In the process, we also provide a standard dataset for literary works labelled by their respective authors.

2 Related Work

Authorship attribution is one of the oldest and newest problems in text classification. Disputes over ownership of words have been as prevalent as the words themselves. The arrival of modern statistics enabled more sophisticated techniques

to verify the authorship of the text. The arrival of computers and the availability of large corpora made it even more feasible to apply various algorithms.

In the domain of authorship attribution, we try to tackle three main problems: (a) closed set authorship attribution, (b) open set authorship attribution, and (c) author profiling. The first problem is, given a set of authors one of whom is known to have written a given document, identify which one wrote it. The second problem is a variation of the first problem, except that here our set of authors isn't closed. Even, if there is a given set of authors, it is possible that none of them wrote it. Thus, it is important to distinguish people outside the set as well. The final problem is a well-established domain in itself, called stylometry or profiling. Here, we determine the properties of the author of a document, given a document such as gender, nationality, age, etc.

The assumption that researchers take while dealing with the problem of authorship attribution is that people have a sort of authorial fingerprint that can be detected in their style of writing. (Van Halteren) terms this as the human stylome, defined as a specific set of traits that can be measured and used to uniquely identify the authors.

The earliest recorded attempt at authorship attribution proposes word length to be a characteristic of authors. The origin is traced to the letters of De Morgan to a clergyman suggesting a way to solve the problem of Gospel authorship. He states that the latter text does not contain words longer than the former text and this can be used to detect spurious writings. Mosteller and Wallace analyzed the distribution of function words extracted from the text of the Federalist papers, thus leading to one of the most cited works in authorship attribution. Their work was based on Bayesian statistical analysis on the frequencies of the function words and produced significant distinguishing between the candidate authors.

Since then, there have been numerous features proposed and discarded such as the average sentence length, average word length, average number of syllables, distribution of parts of speech, and other measures of vocabulary richness. Rudman estimated that nearly 1000 different measures had been proposed. Research during this period lacked an open standard dataset, ground truth labels of texts (i.e. mis-or-unattributed works), failure to control results for a topic, and small datasets and sets of candidate authors. These fundamental problems lead to a general lack of progress in the field other than the case of the Federalist papers.

After the advent of large-scale web-based datasets with objective evaluation metrics as well as a new set of statistical models which work on large and sparse corpora more effectively, authorship attribution has changed for the good. The focus shifted from strict literary analysis to more relevant and immediate goals like plagiarism detection. This trend has led to well-known competitions with well-structured datasets, a large number of participants, and a better comparison of successful and unsuccessful methods. Further support by law enforcement and government institutions helped rekindle interest and innovation in the field, leading to improved results on open benchmarks. The two most cited authorship competitions are the Ad-hoc Authorship Attribution Competition and the PAN (International Workshop on Plagiarism Detection, Author Identification, and Near-Duplicate Detection), which have been run in 2004 and annually from 2007 respectively.

However, researchers are yet to formally evaluate the capabilities of the modern authorship attribution systems on the Hindi language. Information Retrieval and Natural Language Processing in Hindi is still at a very nascent stage. Implementing popular techniques in Hindi lead to a lot of problems due to having a drastically different morphology than the English language, word ambiguity and the phonetic nature of the Hindi language.

3 Steps of Authorship Attribution

We model the problem of authorship attribution as a text categorization task in which we label documents according to a set of predefined categories. We approach this text categorization task in three steps as follows: (a) normalize the text to create uniform snippets of text, (b) achieve a selection of features to represent the text, having high predictive value for the categories (ie. authors), and (c) use machine learning algorithms to learn to categorize new documents as per the extracted features.

3.1 Text Normalization

The first step in the process of authorship attribution is pre-processing to convert our text into a single canonical form. To do so, we first concate-

nate all the literary works by one author into one document and then divide this document into uniform snippets. These snippets have an equal number of sentences. Every single snippet represents a data point, with the label for each snippet being the name of the author who wrote the document from which the snippet was derived. Thus, we reduce all documents into equal-size representation profiles. Further normalization includes transformations such as eliminating all punctuation marks, except in the snippet where we calculate the punctuation mark frequency itself.

3.2 Text Representation

The features used to represent text fall under the one of following different categories: (a) character features, (b) lexical features, (c) syntactic features, and (d) morphological features. We use one or more features at one time to provide a representation of the text to train classification models.

3.2.1 Character Features

These features are accounted for, at the character level. Examples of character features are letter frequencies, punctuation mark frequencies, and character n-grams.

3.2.2 Lexical Features

These features are defined from the word usage in the texts. Examples of lexical features are frequencies of the unigrams and n-grams. Also in this category are relative frequencies of function words or stopwords and a specific class of words such as pronouns, verbs, slangs, etc.

3.2.3 Syntactic Features

These features are syntax related features such as parts-of-speech tags and their n-grams or dependency links.

3.2.4 Morphological Features

These features relate to the form of the text, such as the length related features such as lengths of lines, words and sentences, the overall formatting of the text such as the fraction of empty lines, and other orthographic features such as spelling, hyphens, etc.

3.3 Classification Models

Classification models mainly lie in two main categories: intrinsic and extrinsic verification models. Intrinsic models are based on the set of documents of known authorship to make their decision about the document with unknown authorship. On the other hand, extrinsic models use external resources, that is additional documents by other authors taken from the training corpus.

Both intrinsic and extrinsic models use ensemble classification methods as their classifiers. We can use a variety of machine learning based classification algorithms: linear classifiers such as the Naive Bayes classifier, support vector machines, decision trees, random forests, k nearest neighbours and neural networks. Other popular classification models are modifications of the CNG method, the unmasking method and compression based approaches.

4 Corpus

For authorship attribution, we need a labelled dataset of literary works by authors in Hindi - a language whose script is represented using Unicode characters. Due to the dearth of a good dataset which satisfied the above criterion, we collected and built our dataset. The website hindisamay.com is a collection of novels, poems, essays, among other literary works. However, the documents are present in proprietary formats. We downloaded novels by all the available authors on the website and converted them to the standard plaintext format. Each plaintext file represents a novel or its subparts, in case a single novel is too big. These subparts are stored in a common folder, which represents a single author. The resulting dataset consists of 714 novels by 48 authors. Among these authors, we choose five authors to perform the closed set authorship attribution on Bhairav Prasad Gupta, Dharmveer Bharti, Munshi Premchand, Sharatchandra Chattopadhyaya and Vibhooti Narayana Rai.

5 Experimental Setup

This paper focuses on obtaining and comparing the accuracies of different permutations of text representation features as well as classification models to achieve closed set authorship attribution over the Hindi language. Our experiments involve the following features to construct three different classes of stylometric fingerprints:

1. Morphological Features: This feature involves the mean length and standard deviation of the length of the author's sentences and the vocabulary strength of an author's

Author	Measure	Accuracy
Premchand	Precision	0.972
	Recall	0.957
	Accuracy	97.001 %
Sharatchandra	Precision	0.961
	Recall	0.896
	Accuracy	94.352 %
Dharamveer	Precision	0.975
	Recall	0.778
	Accuracy	95.349 %
Bhairav	Precision	0.971
	Recall	0.816
	Accuracy	94.684 %
Vibhooti	Precision	0.949
	Recall	0.903
	Accuracy	95.183 %
Multi Class	Precision	0.922
	Recall	0.858
	Accuracy	89.369 %

Table 1: Performance score by a support vector machine for bigrams.

work. Every datapoint of this kind is calculated by normalizing the author's document chapter wise.

2. Lexical Features: This feature involves the bag-of-words models and modelling our feature vector for every datapoint using the respective frequency of the 250 most common n-grams. The datapoints are calculated by normalizing through the splits in the concatenated novels of 25 sentences each.

3. Character Features: This feature involves modelling the character n-grams and punctuation frequencies for every datapoint. Again the frequency of the 250 most common character n-grams is used. The datapoints are calculated using the aforementioned 25 sentence splits.

We used an intrinsic approach for our classification model, using only the feature vectors to train our model. Our datapoints were split into the training and testing datapoints in the 4:1 ratio. Upon the test data, we trained both supervised and unsupervised classification models. We use k-means and agglomerative clustering algorithms as our unsupervised clustering methods. For a supervised approach, we trained a support vector machine classifer with a radial basis function kernel.

To analyse our results we are using different metrics. To analyse our clustering we are using the adjusted rand index, homogeneity score, silhoutte score and contingency matrix. To measure the effectiveness of our trained SVM we are measuring accuracy, precision and recall.

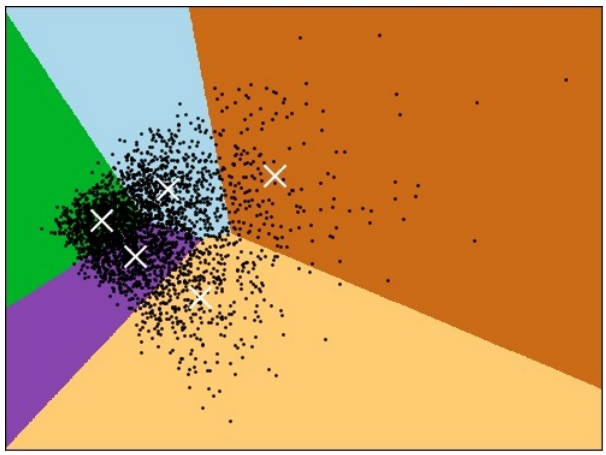

Figure 1: K means clustering on the SVD reduced bigram model. Centroids are marked with a white cross. (ARI: 0.126, HOMO: 0.156. SIL: 0.024)

6 Results

As mentioned above, we use an SVM for our supervised appraoch. Our lexical and character features gave a higher score overall than our morphological features by our SVM. Among lexical features, we received high scores from the bigram model. Our one versus all models over bigram frequencies gave accuracies ranging from about 94% to 97%, with our overall multi class model giving an accuracy of 89%. Table 1 shows the detailed performance scores for the bigram model. The trigram model had fairly high one versus all scores, however the multi class model had a low accuracy of 44%. Among character features, we got the highest scores overall from the character hexagram model. Individual accuracies ranged from 96% to 98.5% and the multi class model had an accuracy of nearly 96% as mentioned in Table 2. The punctuation frequencies too had a fairly high accuracy of nearly 63%, but nowhere close to the character hexagrams. Finally, the morphological model, which is otherwise known to give good scores for the English language failed for Hindi giving an accuracy of less that 40%.

Meanwhile, the results from the clustering algorithm weren't impressive. The possible reason for this is that k means clustering favours circular data and our dataset isn't circular enough. Moreover, there are a lot of outliers. The best results were again given by the character hexagrams and the bigrams with an ARI of 0.278 and 0.126 re-

Author	Measure	Accuracy
Premchand	Precision	0.973
	Recall	0.964
	Accuracy	97.342 %
Sharatchandra	Precision	0.977
	Recall	0.931
	Accuracy	96.512 %
Dharamveer	Precision	0.986
	Recall	0.867
	Accuracy	97.342 %
Bhairav	Precision	0.972
	Recall	0.882
	Accuracy	96.013 %
Vibhooti	Precision	0.977
	Recall	0.973
	Accuracy	98.505 %
Multi Class	Precision	0.966
	Recall	0.949
	Accuracy	95.847 %

Table 2: Performance score by a support vector machine for character hexagrams.

Figure 2: K means clustering on the SVD reduced character hexagram model. Centroids are marked with a white cross. (ARI: 0.278, HOMO: 0.347. SIL: 0.029)

spectively. The clusters are plotted in the Figures 1 and 2.

7 Discussion

Our experiments showed fairly positive values for an initial research in authorship attribution for the Hindi language. While the traditional bag of words model gave a high score as was expected, we found out that character hexagrams are an important feature for constructing a stylometric fingerprint for a Hindi author. As bag-of-words models are usually prone to overfitting, we made sure our model was unbiased by randomizing the datapoint while training. The results also solidify the choice of a support vector machine model as the most common model for authorship attribution.

A major problem faced by our model is the lack of a more elaborate dataset, which could be what is regarded as a reliable minimum for an authorial set. When only limited data is available for a specific author, the authorship attribution using both intrinsic and extrinsic models becomes much more difficult. However, our dataset represents the actual state of the world dataset and more studies could be conducted parallely to the active limited data authorship attribution research.

We were not able to train a proper syntactic model on our dataset, due to the lack of an accurate parts-of-speech tagger in Hindi, and the ones we were trying to run were prone to errors especially on a low power machine which we used for our experiments. With the proper infrastructure and libraries, we might be able to try and extract syntactic features and validate if they are of our use or not.

Using our lexical and character features, we can further train more complex classification models like neural networks which have recently have shown to give better results than other machine learning based models. Till now, we have restricted ourselves from delving right into these other techniques. However, experiments could be conducted to check their performances out.

As an initial experiment, we stuck to the most vanilla form of authorship attribution which is the closed set authorship attribution. However, this is a relatively simpler task compared to other complex tasks mentioned above such as the open set authorship attribution as well as profiling. We further avoid cross genre data, for which research has been actively taking place. Our experiments act as a decent starting ground for the above tasks.

8 Conclusion

In this paper, we implement popular authorship attribution techniques to a Hindi dataset and predict the author of a given text. From our results, we conclude that supervised learning methods like support vector machines which are trained on lexical and character features like word bigrams and character hexagrams, are able to identify the author of text document from a set of authors with a high accuracr of over 90%.

Acknowledgements

This research was conducted as part of the Natural Language Processing course at Indraprastha Institute of Information Technology, Delhi under the guidance of Dr. Tanmoy Chakraborty and his able band of Teaching Assistants.

References

1. Juola, Patrick (2007). Authorship Attribution. Foundations and Trends in Information Retrieval (2007), 1(3):233

2. Stamatatos, Efstathios. (2009). A Survey of Modern Authorship Attribution Methods. JASIST. 60. 538-556. 10.1002/asi.21001

3. Nadi Bozkurt, Ilker Baghoglu, O Uyar, Erkan. (2007). Authorship attribution. Foundations and Trends in Information Retrieval - FTIR. 1. 1 - 5. 10.1109/ISCIS.2007.4456854

4. Koppel, Moshe Schler, Jonathan Argamon, Shlomo. (2011). Authorship attribution in the wild. Language Resources and Evaluation. 45. 83-94. 10.1007/s10579-009-9111-2

5. Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 (COLING '08), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 513-520.

6. Rhodes, Dylan. Author Attribution with CNNs. (2015).

7. Yadav, Sudesh. Natural Language Processing and Hindi Language. (2015).

8. https://www.nltk.org/

9. https://scikit-learn.org/