ICRITCSA -2019

Authorship Identification using supervised learning and *n*-grams for Hindi Language

Jagadish S Kallimani, Chandrika C P, Aniket Singh, Zaifa Khan,
Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India

Email: *jagadish.k@msrit.edu*

Insert Submission Date Here

Submitted/Received: ……....……………..…   Accepted: ………….....…………...

Abstract

Authorship Identification pertains to establishing the author of a particular document, currently unknown, based on the documents previously available. The field of authorship identification has been explored so far primarily in the English language, using several supervised and unsupervised machine learning models along with usage of NLP techniques, but work on regional languages is highly limited. This may be due to the lack of collection of proper datasets and preprocessing techniques attributed to the rich morphological and stylistic features in these languages. In this paper we apply some supervised machine learning models, namely SVM and Naïve Bayes to Hindi literature to perform authorship analysis by picking four Hindi authors. We compare and analyze the accuracy which is so obtained using different models and bag of words approach.

Keywords: Authorship Attribution, Support Vector Machines, Stylometry, *n*-grams, Supervised Learning

1. Introduction

Authorship identification refers to identifying different stylistics features in the article or published works of a certain writer and determining the author in a different unidentified text. Authorship Identification can be termed as a classification problem however each author may have a different way of writing and mere analysis of words and there categories may result in wrong outcomes. Authorship Identification has several applications including but not limited to identifying authorship of an anonymous article, preventing fraud and plagiarism, allowing only authorized writer to perform classified tasks.

With the advent of the internet there is a huge increase in textual data and several systems and attempts have been made to classify this data, based on different parameters. However one very important attribute remains authorship, as it enables the benefits and responsibilities of the content to fall in the right place. Plagiarism and fraud cases have gained momentum with the increase in data available online and with the right authorship identification tools and models we can prevent any such theft of intellectual property and gains associated with them. Authorship identification can also be used to resolve conflicted categorization of historical documents, mine sentimental data from the features extracted from the text.

Research work has been done in this regard but mostly on the English language whereas the vast field of regional languages, especially Hindi, remains mostly untouched. The proposed work includes two supervised learning models to identify the authorship of anonymous Hindi language articles. Our results can further serve as a base to future research work pertaining Hindi and as motivation for local and regional languages.

2. Related Works

Support vector machine (SVMs) is a supervised algorithm, which can solve classification problems of large and higher dimension datasets[1]. However they require manual labeling of data which may become difficult if performed for all the data available on the digital platforms. In [2] several strategies were presented to solve the problem of imbalanced classification for real-world text and a comparative study was completed with respect to Support Vector Machines and it was found that the best hyper plane for classification was generated when standard SVM was used, which implicitly pointed to paying attention towards the thresholding strategies.

In [3] it explains that stylometry is identifying the style of an author as a measurable quantity. It involves tasks like authorship attribution, authorship verification, and authorship profiling. However it identifies an existing challenge in the field which is the lack of analysis to identify authorship from a large number of writers having very few samples per author, which is recognized in [4] as well, wherein methods of machine learning are implemented to handle the problems of verifying, profiling and needle-in-a-haystack.

According to [5] authorship identification, which is a branch of digital forensics, is a problem where we have a set of documents by a particular writer and a document which is allegedly written by the same writer. For different languages and different authors, selection of specific segmentation parameters and threshold is the most important factor. In [6] different information retrieval methods like *tf-idf* structure with SVM and clustering were tested for the authorship attribution of Turkish text from a local newspaper and found best results with support vector classifier using bag of words.

The paper [7] explores the use of n-grams for segmentation and better understanding of Chinese texts which provides an insight into feature selection methods for local and regional languages where the structure is more complex than English. The paper [8] also identifies how internet has caused an increase in the need of authorship identification especially for smaller data samples. Using word *n*-gram and character *n*-gram features and supervised learning methods they showed a promising result using SVM and Naïve Bayes whereas Conditional Tree and Random Forests underperformed in comparison.

3. Implementation

The proposed idea is implemented in the following sequence as shown in the Figure 1:

- Text processing
- Feature extraction
- Authorship identification

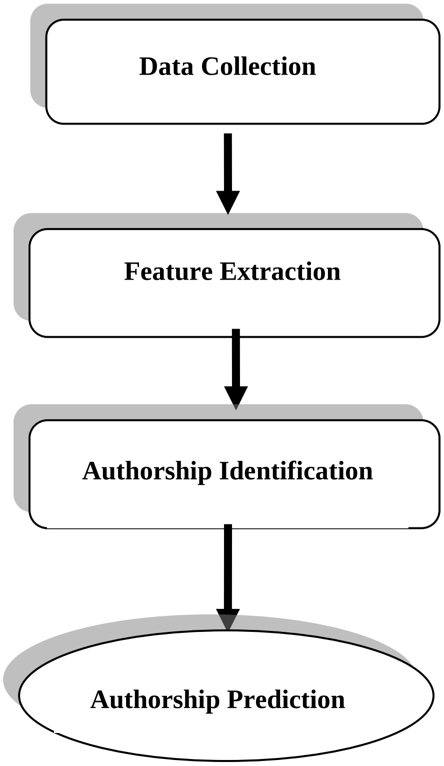

Figure 1: Flow sequence of the proposed work

3.1 Text Preprocessing:

The dataset for the proposed work is taken from Hindi text corpus which contains stories written by four different authors, totally 2089 different files are considered.

In this module the punctuations including Hindi *danda* (the vertical stroke used to distinguish between sentences, as period in English), quotation marks, and other punctuations are pruned out. Followed by splitting the processed text corpus into different files so that they can be tokenized for the creation of *n*-grams, with each file contains 500 words separated into folders according to their respective authors. As machine learning algorithms cannot directly process the text corpus, the raw data is converted into vectors of numbers using *n*-grams technique, particularly bi-gram and tri-gram.

3.2 Feature extraction:

N-grams refer to the grouping of *n* words together. In comparison to single words, *n*-grams offer more insight (uni-grams, where *n*=1) as features, therefore we generated bi-grams and tri-grams and considered top 3000 most frequent groups from those, as the features in the input matrix. For each of the 2089 documents words we generated tokens in a group of two and three. The value of input matrix for that particular feature is 1 for the features it matches; otherwise it is marked as 0. Hence an input feature matrix ‘X’ of dimension (2089 x 3000) is obtained. The dependent variable ‘Y’ or the output vector is different for every author. If a particular attribute belongs to a particular author then it is marked as 1 otherwise it is marked as 0, for that author. This results in 4 dependent vector of dimension (2089 x 1), one for each author.

3.3 Authorship identification:

Since we have both, dependent or input matrix ‘X’ and independent or output vector ‘Y’, we feed them to two machine learning algorithms, namely SVM (support vector machine) and NB(Naïve-Bayes) to learn the weights for all the 3000 features and make predictions on new text provided to the algorithm.

3.3.1 Training Using Support Vector Machine:

Application of Support Vector Machine (SVM) includes solution of regression problems as well as classification tasks, but it’s more frequent application is for classification problems. Here, each feature set is plotted in a vector space of *n*-dimensions (where *n* is number of total features in the data set). The value of a particular point in the co-ordinate system is derived from all of the *n*-features. By creating a plane, known as the hyper-plane we classify the input parameters with this plane that creates a boundary between different classes. Hyper-plane is characterized by the weights that are learnt by the machine learning algorithm. So for best fitted hyper-plane optimal set of weights has to be determined. The mathematical insight for determining the optimal set of weights is as follows.

Let us assume that the weights associated with each of the features that is the 3000 most frequently occurring bigrams or trigrams, generated from the text corpus, are represented through ‘Θ’. Our main objective in the implementation of SVM, is to minimize the squared sum of weights (depicted in equation 1), such that on multiplication of weight and input attribute (each of the 2089 documents), the value obtained should be greater than or equal to 1 if that document is the work of the particular author, or value obtained should be less than -1 if that document is not the work of that particular author. Mathematically,

Objective function:

$$\frac{1}{2} minimize(\sum \theta^2) \qquad \text{.......(1)}$$

Such that,

$$\theta^T x^{(i)} \geq 1 \qquad \text{if } y^{(i)} = 1$$

$$\theta^T x^{(i)} \leq -1 \qquad \text{if } y^{(i)} = 0$$

To upgrade the value of weights, we use convex optimization technique, which gives high probability of providing globally minimum value of weights. To predict the authorship of a new document, $x^{(p)}$, we calculate the value $\Theta^T x^{(p)}$, i.e. product of input value with the learnt weights. Let this value be $y^{(p)}$. If $y^{(p)} \geq 1$, the input document is the work of this specific author, otherwise not.

3.3.2 Training Using Naive-Bayes algorithm:

Another supervised learning algorithm we used was the Naïve-Bayes classification which is a machine learning technique used for classification based on the mathematical concept of Bayes' Theorem. It assumes that predictors are independent among themselves. In simple words, this algorithm assumes that the presence of any specific feature in a class is completely independent and is unrelated to any other feature's presence. Posterior probability can be calculated using Bayes' Theorem with the help of prior probability. The mathematical equation involved is as follows:

$$\boldsymbol{P}\left(\frac{\boldsymbol{y}}{\boldsymbol{x}}\right) = \frac{\boldsymbol{P}\left(\frac{\boldsymbol{x}}{\boldsymbol{y}}\right) * \boldsymbol{P}(\boldsymbol{y})}{\boldsymbol{P}(\boldsymbol{x})} \quad \ldots.. \qquad (2)$$

where, $\boldsymbol{P}\left(\frac{\boldsymbol{y}}{\boldsymbol{x}}\right)$ is the probability that given a particular document, x, it is authored by y.

$\boldsymbol{P}(\boldsymbol{y})$ is the probability of authorship by y.

$\boldsymbol{P}\left(\frac{\boldsymbol{x}}{\boldsymbol{y}}\right)$is the probability author y, has written document x.

$\boldsymbol{P}(\boldsymbol{x})$signifies the probability of showing up of a particular document, x, among the complete dataset of documents.

In our case x is the value of a particular document from the input matrix X and y is the corresponding value of dependent variable from the output vector Y, which represents the truth value of authorship.

4. Results

After training the model with SVM and NB using bigram and trigram we obtained the following results, shown in the Table 1 and graph below:

Table 1. Accuracy measures for different authors

S.No	**Authors**	**Accuracy from Different Machine Learning Algorithms**			
		Support Vector Machine (%)		**Naïve-Bayes (%)**	
		Bi-gram	Tri-gram	Bi-gram	Tri-gram
1.	Author 1	94.48	94.37	87.96	88.30
2.	Author 2	64.91	73.91	74.11	67.84
3.	Author 3	75.28	74.39	83.01	82.25
4.	Author 4	79.86	79.86	65.24	76.82

From the table it can be inferred that SVM gives maximum accuracy on the given dataset than NB. We obtained minimum error of 5.52% for the Author 4 by SVM model using bigrams and the maximum error of 35.09% for the Author 2 by SVM model using bigrams.

The graph is plotted for the obtained results. It shows that NB has less variance in accuracy in comparison to SVM.

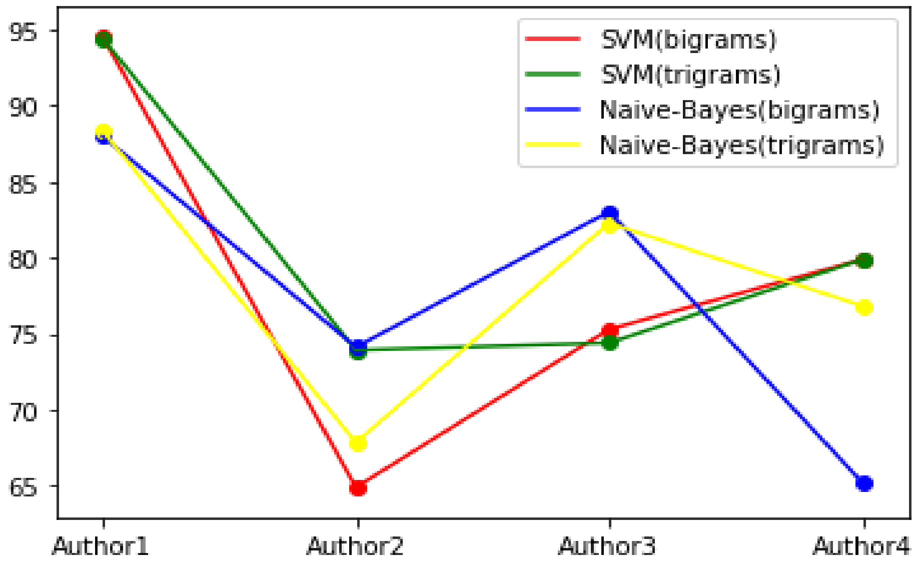

Figure2: Accuracy rate of SVM and NB

5. Conclusion and Future Scope

Different authors and their work was selected from Hindi literature and four machine learning approaches were trained and tested on it. This analysis provided the valuable insight on the feasibility of using standard machine learning methods on Hindi language as well as Indian Literature for Authorship Attribution.

Enhancements to the working of our model can be provided using hyper parameter tuning techniques like Grid Search. Different segmentation methodologies can be implemented and dataset can be made more comprehensive spreading over different domains besides literature like finance or governance and can be used to find correlations in market and strategy analysis. We can deploy other deep learning models like Artificial Neural Networks or unsupervised learning models like K-Means Clustering and attempt to boost the accuracy hence obtained.

References and Notes

1. Mohamed Goudjil, Mouloud Koudil, Mouldi Bedda, Noureddine Ghoggali., **2018**A Novel Active Learning Method Using SVM for Text Classification.*International Journal of Automation and Computing*, Volume 15, Issue 3, pp 290–29.

2. AixinSun, Ee-PengLim, YingLiu., **2009**On strategies for imbalanced text classification using SVM.*A comparative study, Decision Support Systems*, Volume 48, Issue 1, Pages 191-201.
3. Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, Yingfei Xiang, Damon Woodard., **2018**Surveying Stylometry Techniques and Applications – Journal

ACM Computing Surveys (CSUR) Surveys Homepagearchive, Volume 50 Issue 6, , Article No. 86.

4. Moshe Koppel, Jonathan Schler, Shlomo Argamon.,**2009** Computational Methods in Authorship Attribution.*Journal of the American Society for Information Science and Technology.*

5. Oren Halvani, Christian Winter, Anika P flug., **2016**Authorship verification for different languages, genres and topics.*DFRWS Europe- Proceedings of the Third Annual DFRWS Europe.*

6. I. N. Bozkurt, O. Baglioglu, and E. Uyar.,**2007**Authorship attribution. *22nd International Symposium on Computer and Information Sciences (1—5)***.**

7. ZhihuaWEI , Duoqian MIAO, Jean-Hugues CHAUCHAT, Rui ZHAO1, Wen LI.,**2009** N-grams based feature selection and text representation for Chinese Text Classification.*International Journal of Computational Intelligence Systems*, Vol.2, No. 4 ,pp 365-374.

8. Edwin Dauber, Rebekah Overdorf, Rachel Greenstadt.,**2017**StylometricAuthorship Attribution of Collaborative Documents.*International Conference on Cyber Security Cryptography and Machine Learning,CSCML* ,Cyber Security Cryptography and Machine Learning pp 115-135.