Author Identification using Stylometry

Mentor, Mrs. Sujata Khedkar
Associate Professor
Computer Engineering,
VESIT, Chembur
sujata.khedkar@ves.ac.in

Shashank Agnihotri
B.E. Computer Engineering,
VESIT, Chembur
shashank.agnihotri@ves.ac.in

Anshul Agarwal
B.E. Computer Engineering,
VESIT, Chembur
anshul.agarwal@ves.ac.in

Mahak Pancholi
B.E. Computer Engineering,
VESIT, Chembur
mahak.pancholi@ves.ac.in

Pooja hande
B.E. Computer Engineering,
VESIT, Chembur
pooja.hande@ves.ac.in

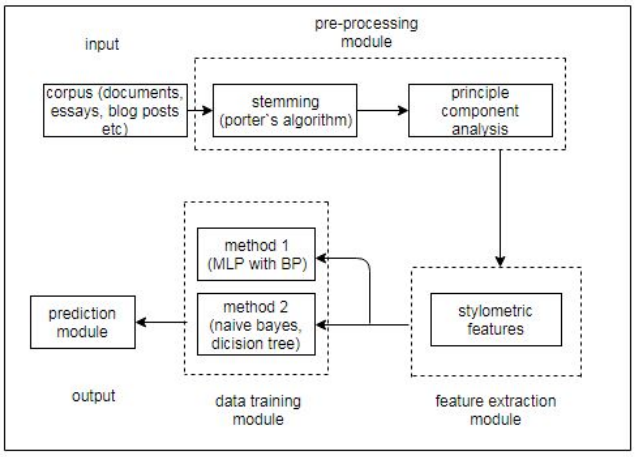
***Abstract*—"Every person is unique", we have been hearing this since ages. Every person has a unique identity, a unique fingerprint, a unique retina and a lot more. These features play a vital role in identification of individuals for security purposes. Unfortunately, when it comes to security of written pieces or words from an individual, these primary unique identities are futile. One cannot identify a writer from a written piece of text on the basis of retina or fingerprint scans, sometimes even the signature can be forged, in such situations for security purposes and intellectual property rights it becomes very important to identify the true author. Stylometry plays an important role in this. Every author has a unique style of writing, measure of this style of writing is called Stylometry. This paper proposes to identify authors from text based on their style of writing. First a data set consisting of articles, short stories and emails will be used to train the system for multiple authors, then a random text would be given to the system to identify the author correctly, if the author predicted by the system is similar to the author claimed then the information is authentic otherwise the author claiming to be the writer is a fraud. For stylometry, over the ages, many features have been focused on, but this paper proposes new features to be used for this purpose. While writing, there are many unconscious styles that are incorporated by the author, these features have been unnoticed till date, but can play a vital role in accurate and fast identification of authors. These features include: 'intellectual property right', 'chapter length' and frequency of particular words per thousand words. The algorithms used to train the system can be Decision tree, Naive Bayesian or Multilayer Perceptron.**

Keywords—feature extraction, data set, Decision tree, artificial intelligence, machine learning, supervised learning.

I. Introduction

Various attempts have been made to identify author using stylometry. Most of the attempts made use of similar feature extractions but different data sets and algorithms. Every system had a drawback that couldn't be overlooked. Jose Hurtado, Napat Taweewitchakreeya, and Xingquan Zhu in their paper[1] used multilayer perceptron, random forest, SVM and k-nearest neighbour for training the data. Here the MLP learner, combined with the six categories of stylometric features, provides better performance over other classifiers and baseline approaches however Random forest and k-nearest neighbours give low accuracy and only few authors can be identified accurately. While in [2] Kohonen Self Organising Maps and backpropagation is used which is suitable to capture an intangible concept like style and in this fewer input variables are required as compared to the traditional statistics but this can be implemented only for small number of authors. [3] seems to cover all the drawbacks of [1] and [2] and other related works. [3] uses LDA and Naive Bayes for classification which enables it to do semantic analysis of corpus however it brings in a

new drawback with it: to classify a new unknown document it would be necessary to reprocess all documents including new ones, this is an onerous and time consuming task. Thus this paper proposes a new methodology that encompasses almost all the benefits of [1], [2], [3] and [4] as it overcomes their drawbacks. Stemming and Principal Component Analysis will provide a sharp edge in cutting down the processing time while increasing the efficiency of the new proposed system. Moreover focus on a new set of features will provide better accuracy and including a 'Pre-Processing stage' in the system will



tremendously decrease the payload on the system when adding new data sets to the already trained system.

Fig. 1. Block diagram of the proposed system

II. Preprocessing

A. Stemming

Stemming, a crude heuristic process is used to chop the end of the words in order to achieve the goal correctly and more easily. It focuses on removing the derivational affixes as well.

Porter's Algorithm as mentioned in [6] can be used. 5 phases of word reductions are applied sequentially in Porter's algorithm. Each phase consists of various conventions to select the rules which are suitable.The example of the same can be that a rule can be selected from a particular rule group and hence applying it to the suffix with the largest length.

B. Principal Component Analysis

Principal Component Analysis refers to analysis of data which will be responsible to identify the patterns and then finding the patterns to reduce the dimensions of the dataset drastically, taking into consideration the minimal loss of the information. One of the way for performing Principal Component Analysis is by choosing a subset of Principal Components and Variables as mentioned in section 6 of [7].

III. Feature Extraction

A. Mainstream Methods

1. **type-token ratio:** The type- token ratio denotes the tendency of the author to repeat certain words in a sentence. It is used to define the richness of the vocabulary of the author. Higher type-token ratio denotes varied vocabulary.
2. **mean word length:** Longer words tend to have more formal styles and pedantic as compared to the shorter words which are typically used for informal spoken language.
3. **mean sentence length:** Sentence length is a vital factor wherein the longer sentences are usually associated with careful planned writing, whereas shorter sentences symbolize more of a spoken language characteristic.
4. **standard deviation of sentence length:** It is one of the important marker of style. Standard deviation denotes the variation of a sentence length.
5. **mean paragraph length:** The paragraph length is used to determine the occurrences of dialogues.
6. **number of commas per thousand tokens:** Commas play a crucial role, which denote the ongoing flow of ideas within a sentence.
7. **number of ands per thousand tokens:** Ands are the markers used to represent coordination. It is frequently used in spoken production.
8. **number of buts per thousand tokens:** Buts are the markers of coordination, used to represent the contrastive linking.

9. **vocabulary:** Every authors selected vocabulary was chosen.

B. New Methods

1. **chapter length:** It will denote the length of the sample chapter.
2. **number of colons per thousand tokens:** colons indicate the reluctance of an author to stop a sentence where(s) he could.
3. **number of quotation marks per thousand tokens**: The frequent use of quotations marks denote the involvement feature.
4. **number of exclamation marks per thousand tokens:** Exclamation mark is a marker of strong emotions.
5. **number of hyphens per thousand tokens**: Hyphens play an important signal as some of the authors use hyphenated words more than other authors.
6. **number of howevers per thousand tokens:** "However" forms a contrastive pair with "but" in a sentence structure which can act as an important marker as well.
7. **number of 'if's per thousand tokens:** Ifs will denote the samples of subordination.
8. **number of 'that's per thousand tokens:** 'That's in a sentence usually denote the subordination and also can be used as demonstratives.
9. **number of 'more's per thousand tokens:** More is used as an indicator of author's liking for a comparative structure.
10. **number of 'must's per thousand tokens:** Modal verbs are potential candidates for expressing tentativeness. Musts are more often used non-epistemically.
11. **number of 'might's per thousand tokens:** 'Might's are a marker which are usually used epistemically.
12. **number of 'this's per thousand tokens:** This is a marker which is used in the case of anaphoric reference.
13. **number of 'very's per thousand tokens:** Verys are significant because of it's emphasis on it's modifies.
14. **part-of-speech tagging (PoS tagging):** Penn Treebank PoS tagging denotes annotations. (ex. CC for coordinating conjunction, SYM for symbol)

IV. Intended Methodology

A. Method 1

Author identification would be done using a mixture of two algorithms. The primary algorithm would be Decision tree but instead of using values of features in the Decision tree, their conditional probabilities would be used. Multinomial Naive Bayes algorithm would be used for this purpose. While the training the system, on the data set, the system will calculate the priors for every tag word (focused features), then on the basis of the documents calculate the conditional probability of these features to the documents. Then on the basis of conditional probabilities the Decision Tree would be formed to classify each document to its respective author, in the process giving the system a measure of style of writing for every author.

Then when an author needs to be identified for a given random document, the priors and conditional probabilities would be calculated using the Multinomial Naive Bayes algorithm for that document and the values of the conditional probabilities would be used as an input in the trained Decision Tree. This would ensure high accuracy and fast computation when compared to that of both algorithms individually.

B. Method 2

Multilayer Perceptron would be used for training the system on the dataset.The Multilayer Perceptron for better accuracy will have backpropagation with it. The features to be focused on by the system, as discussed in section III of this paper, would be the factors of the hidden layers in the Multilevel Perceptron.The target state would be grouping all documents from the same author in one category or class. This would be partially supervised learning. As the system would initially not know the true authors of the documents but just have the documents. On the basis of the features, the documents will pass several layers of the perceptron to be distinguished on the basis of stylometry and then the output would be compared with the desired output. The errors would be back propagated and the weights would be adjusted to finally categorize all the documents correctly under their respective authors. This will give us a unique identity and

stylometry for every author. This would be tagged to every author to help us in author identification and verification in case of new documents.
Then when a new document would be tested against the MLP the system would already be trained and the weights adjusted, thus the system would accurately be able to identify the true author and plagiarism, if any.

V. Specifications and Tools

A. *Hardware Specifications*

1. *DDR4 RAM*
2. *4 GB Graphic card*
3. *6th generation intel processor or above*

B. *Software Specifications*

1. Data Processing : Anaconda IPython notebook, NLTK, GraphLab, Matlab, Octaves
2. User Interface : Xampp, Html, CSS, JavaScript, Bootstrap, Php
3. Analytical Tools : Knime, Weka
4. Data Visualization : Tableau,D3.js

C. *Tools*

1. Jetbrains Webstorm
2. *R Studio*
3. *Jetbrains Pycharm*
4. *Scikit*
5. *nltk library*
6. *MatLab*
7. *Octave*

VI. Results

The system will calculate the chances of each author having the chances of writing the document, and the author which has the highest percentage would be identified by the system to be the true author. If the author claimed and the author identified by the system are same then the claim is validated, if not then the author has falsely claimed to be the author of that document.

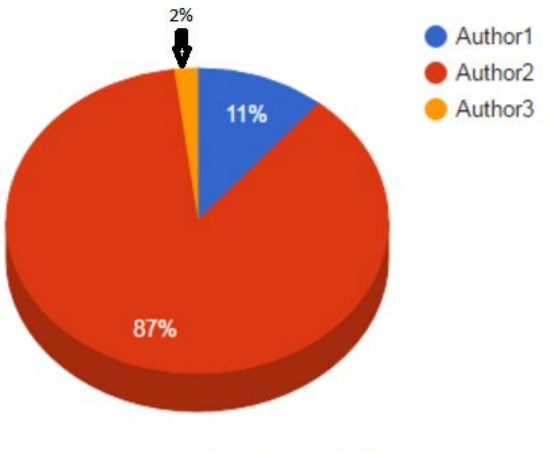


Fig. 2. Sample output

This is how the result will look. Author 1 has 11% chances of being the author of the given document, Author 2 has 87% chances and Author 3 has 2% chances of being the author of the given document. Thus clearly Author 2 is identified by the system as the true author of the document.

VII. Constraints

The techniques used to determine the authors will be requiring relatively large dataset of novels, to train and test our system. Large dataset is required to accurately determine author in a large pool of authors.
The proposed algorithm might not work, with high precision, for a smaller dataset.

VIII. Future Scope

Determining the most favorable algorithm or developing a new algorithm for training the data-set is something in sight as future scope of this project. Moreover the proposed system only takes into account syntax for feature extraction in stylometry, extending the feature extraction to semantic features would develop the system and increase its accuracy many folds. However, mechanisms for this purpose haven't been developed yet, so incorporating this feature in the system remains as a future scope waiting for developments of suitable and feasible mechanisms to make this a reality.

Stylometry will play an important role in identification of potential social media hazards and in cracking cyber crime cases. Being able to incorporate short messages like tweets, Facebook

posts or WhatsApp messages to train data and identify the author would be helpful and play an instrumental role in this field. This use of stylometry is also something that has to be left for future development as currently for accurately identifying authors there is substantial amount to written text that is needed to train the system.

IX. Conclusion

Thus this paper addresses an old but unsolved problem of accurate and reliable author identification using stylometry. If successful, stylometry would a vital role in cybercrime forensics and would help the world solve ages of mysteries regarding ownership of various writing pieces by authors. It would be used to identify anonymous works and saying by comparing it the the style of authors and famous personalities of those days.

For this very purpose, the paper proposed new features to be extracted from the document which hopefully would assist in increasing accuracy and reducing a few redundant dimensions. The paper even proposes a new algorithm that could be identified for faster computation and better accuracy. This new proposed algorithm is basically a combination of two well known and used algorithm. This paper certainly proposes methods that would overcome the known drawbacks of previous works in this field.

References

[1] Hurtado, Jose, Napat Taweewitchakreeya, and Xingquan Zhu. "Who wrote this paper? learning for authorship de-identification using stylometric featuress." Information reuse and integration (IRI), 2014 IEEE 15th international conference on. IEEE, 2014.

[2] Ramyaa, Congzhou He, and Khaled Rasheed. "Using machine learning techniques for stylometry." Proceedings of International Conference on Machine Learning. 2004.

[3] Hernández-Castañeda, Ángel, and Hiram Calvo. "Author Verification Using a Semantic Space Model." Computación y Sistemas 21.2 (2017).

[4] Brocardo, Marcelo Luiz, et al. "Authorship verification for short messages using stylometry." Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on. IEEE, 2013.

[5] Crawford, Michael, et al. "Survey of review spam detection using machine learning techniques." Journal of Big Data 2.1 (2015): 23.

[6] Willett, Peter. "The Porter stemming algorithm: then and now." Program 40.3 (2006): 219-223.

[7] Jolliffe, Ian T. "Principal Component Analysis and Factor Analysis." Principal component analysis. Springer New York, 1986. 115-128