# Reclaiming Individuality of Mysterious Passage

M. Chaurasia
Member, ACM

Dr. Sushil Kumar
Bhilai Institute of Technology, Durg, INDIA

***Abstract*— Authorship attribution, the science of inferring characteristics of author from characteristics of documents written by that author become an urgent need to find the original author of anonymous text. In this paper, a novel approach is proposed that attempts to measure the style variation of author using character n-gram profiles. This proposed method is a different approach to identify the author using initial character n-gram whereas prior research has shown the identification on total character n-gram. This approach will prove to be quite stable. With the help of small experiment, we attempt to prove it. The results acquired from the mention technique are quite accurate and it hikes to 100% in identifying the author from an anonymous text. Using N-gram frequency profiles, it provides a simple and reliable way to categorize documents in a wide range of classification tasks.**

**Keywords- Author identification, Character n-gram, Dis-similarity measure, Natural Language Processing.**

## I. LITERATURE RIVIEW

Organizations (business, academia, government, etc.) are facing risks resulting from their ever-increasing reliance on the information infrastructure. Decision and policy makers managing these risks are challenged by a lack of information intelligence concerning the risks and consequences of cyber events .They need to understand the implications of cyber security risks and solutions related to their information infrastructure and business. The combination of increased vulnerability, increased stakes and increased threats make cyber security and information intelligence (CSII) one of the most important emerging challenges in the evolution of modern cyberspace mechanism.

In some criminal, civil, and security matters, language can be evidence. A suicide note, a threatening letter, anonymous communications, business, emails, blog posts, trademarks—all of these can help investigators, attorneys, human resource executives and private individuals understand the heart of an incident. When you are faced with a suspicious document, whether you need to know who wrote it, or if it is a real threat or a real suicide note, or if it is too close for comfort to some other document, you need reliable, validated methods [8].

Author attribution is a well-studied area of artificial intelligence. Formalized methods for determining authorship have even older roots. The field of Stylometry began even before the turn of the twentieth century with several documented methods of analyzing documents to settle disputed authorship. Linguistic idiosyncrasies that have been identified and are currently being exploited include methods ranging from counting keywords to analyzing punctuation [1]. Reference [2] introduced a new recognition technology to identify authors with their synonyms choice. Author's choice of synonyms is peculiar and can be used in determining the identity of an author. In the same year, R.Williams [3] explores a biometric tool based on frequency analysis which is used for identifying the author or can say digital signature for authorship attribution.

Stylometry attempts to capture an author's style using quantitative measurements of various features in the text such as word length or vocabulary distributions. Many stylometric studies have measured word dependencies as a feature of an author's style using language models that restrict what words a given word can depend upon. Furthermore, another approach [4] was based on building a character-level n-gram model of an author's writing, in which one Greek dataset performs 18% accuracy improvement over deeper NLP techniques. They [4] measures the classifier performance by using *precision*, *recall,* and *macro-average F-measure* scores whereas,

*Precision* = number of correctly classified texts in, divided by the number of all texts classified to be in, *Recall* = number of correctly classified texts in, divided by the number of all texts that truly belong to,

F-measure = (2*precision*recall) / (precision + recall)

The character n-gram approach has been proven to be quite useful to computing the writing style [5]. In 1994 [6], Kjell first used character bigrams and trigrams to distinguish the Federalist Papers. Keselj [10] and Stamatatos [7] described character n-gram information with very fine results. Moreover, one of the best performing algorithms in an authorship attribution competition organized in 2004 was also based on a character n-gram representation [11]. An N-gram is a fill in succession of k-items in any given string of length m, where grams can be anything, from characters to words. In computational linguistics n-gram models are used most commonly in predicting words (in word level n-gram) or predicting characters (in character level n-gram) for the purpose of various applications [12].

## II. OUR METHODOLOGY AND ALGORITHM

An N-gram is an N-character slit of a longer string. Although in the literature, the terms have an impression of any co-occurring set of characters in a string. The key feature of this paper is we do not use the term for conterminous slits. Typically, our paper strategy based on one start slit of every term (initial slot). Thus, for e.g. sentence “Play makes child healthy and fit.” would be composed of following initial n-gram.

TABLE I. POSSIBLE INITIAL CHARACTER N-GRAM

| | | | | | | |
|---|---|---|---|---|---|---|
| **Bi-gram** | pl | ma | ch | he | an | fi |
| **Tri-gram** | pla | mak | chi | hea | and | fit |
| **Quad-gram** | play | make | chil | heal | -- | -- |

In general, a string of length m, will have m bi-grams, m tri-grams, m quad grams and so on. A possible strategy for character n-gram is to use all positional n-grams including initial n-gram. However, such a strategy often leads to less feature space i.e. easy identification from an existing machine learning methods.

### *A. Representation of Profiles of N-gram Frequency*

As in [9], working atmosphere of our approach is shown in bubbles Fig. 1. Pre-processed data generate the n-gram profiles of author which then fed to evaluate the value of dis-similaity measure which must be unique for identifying an author. However, pre-processing required steps are as follows:

- Remove numerals from the text.
- Take away all punctuations marks from text.
- Complete course of action is case-insensitive.

Creating the character n-gram profile includes the following steps:

- Divide the text into separate tokens based on preprocessing steps.
- Compute all possible character n-gram for N = 2, 3and so on.
- Form it into table to ascertain that each output N-gram has its own frequency.
- Sort the N-grams order by their frequencies from most frequent to least frequent.
- Assemble the profile size of each author like 100, 200 and so on for bi-grams, tri-grams, quad-grams and so on.

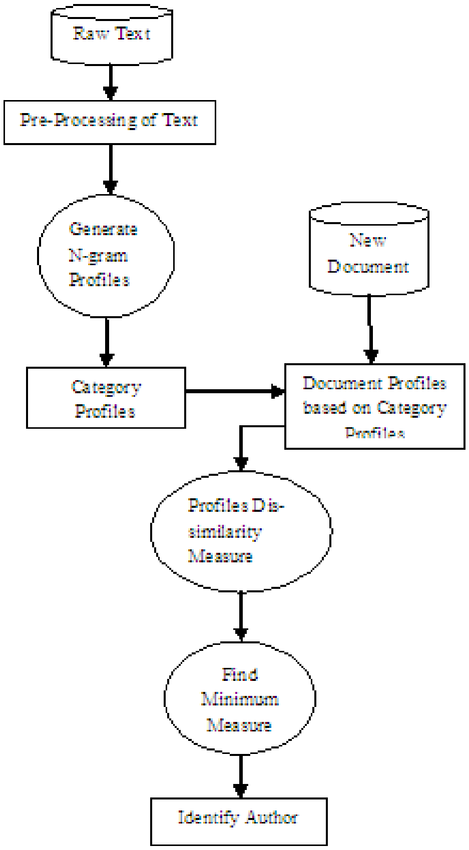

Figure 1. Rendering N-Grams Profiles

Each document to be classified went through the text preprocessing phase, and then the N-gram profile was generated as described above. The N-gram profile of each text document (document profile) was compared against the profiles of all documents in the training classes (class profile) in terms of dis-similarity algorithm [10] used in this paper.

### B. Algorithm and Equations

This algorithm used for computing dissimilarity is the one presented in [10] .We generates the bi- & tri-grams from author's original sample text called Author's Profile. A similar profile is generated for the test data. Now, let *f1 (n)* be the frequency of the *nth* bi- or tri-gram in the Author's Profile. Let *f2 (n)* be the frequency of the *nth* bi- or tri-gram in the test data. We calculate the dissimilarity between the two using the following formula:

$$\sum_{n\text{€ } profiles} (f1\ (n) - f2\ (n))^2$$

In order to normalize these differences, we divide them by the average frequency for a given n-gram.

$$Sum = \sum_{n \in profile} \left( \frac{f1(n) - f2(n)}{\frac{f1(n) + f2(n)}{2}} \right)^2$$

$$Sum = \sum_{n \in profile} \left( \frac{2(f1(n) - f2(n))}{f1(n) + f2(n)} \right)^2$$

Given two profiles the algorithm returns a positive number, which is a measure of dissimilarity. For identical texts, and more generally, for texts that has identical L most frequent n-grams, the dissimilarity are 0. Using this measure and a set of author profiles, we can easily assign a text to an author by generating a text profiles and assigning the text to the category with which the calculated dissimilarity is minimal. Moreover, in order to get better result text size data should be separated and of same size approximately whereas train size can varies and use to train the system.

In experimental analysis, take profiles of different length as reference to compare with the complete profile of the unknown text for the same author & also for the unknown text for the different author .Then the profile of each author is being compared with the profiles of all authors exist in the same data set. If the value of dis-similarity measure of an anonymous text is close to the value of dis-similarity of author specified in data set, then author could be identified.

## III. EXPERIMENTAL RESULTS

In this section, we discuss a small corpus which includes train size text and test size text for our problem province. In continuation, we compute the results of accuracy to identify an unknown text which varies with type of n-gram.

In experimental analysis, capture the profiles of authors of variable size for training and capture the profiles of different authors along with the same author profiles for testing. For our experimental purpose, we used data-set of four authors mention in TABLE II.

TABLE II. AUTHORS IN OUR DATA-SET

| | | **AUTHOR NAME** | **TRAIN SIZE (Words)** | **TEST SIZE (Words)** |
|---|---|---|---|---|
| **DATA SET** | P0 | EVA -GALE | 22068 | 22068 |
| | P1 | PAYTON LEE | 278303 | 275585 |
| | P2 | RUTH ANN NORDIN | 235807 | 109254 |
| | P3 | ROSS BECKMANN | 124047 | 124047 |

In TABLE II, it is mention that the total number of words for train size and total number of words for test size of respective authors. It includes three novels for P0, six novels for P1, five novels for P2 and two novels for P3. A profile is generated from each novel. Furthermore, minimally all the authors has two novels. So, the training set profiles is generated from the combination of two novels from each author and the accuracy is measured on attribution of rest of novels. We experimented our approach with the small data-set and compared the accuracy of verifying the author from different authors' profiles. In the following tables, we demonstrated medial n-gram, final n-gram, total n-gram and the initial n-gram.

TABLE III. INITIAL BI-GRAM

| **Profile Length** | 50 | 100 | 200 |
|---|---|---|---|
| **Accuracy Measure in percentage** | 68.75 | 87.5 | 100 |

TABLE IV. INITIAL TRI-GRAM

| **Profile Length** | 100 | 200 | 500 | 700 |
|---|---|---|---|---|
| **Accuracy Measure in percentage** | 81.25 | 87.5 | 100 | 100 |

TABLE V. TOTAL BI-GRAM

| **Profile Length** | 100 | 200 | 430 |
|---|---|---|---|
| **Accuracy Measure in percentage** | 56.25 | 75 | 81.25 |

TABLE VI. TOTAL TRI-GRAM

| **Profile Length** | 100 | 200 | 400 | 700 |
|---|---|---|---|---|
| **Accuracy Measure in percentage** | 75 | 75 | 81.25 | 87.5 |

TABLE VII. MEDIAL BI-GRAM

| **Profile Length** | 50 | 100 | 200 |
|---|---|---|---|
| **Accuracy Measure in percentage** | 62.5 | 81.25 | 81.25 |

TABLE VIII. MEDIAL TRI-GRAM

| **Profile Length** | 500 | 700 | 1000 |
|---|---|---|---|
| **Accuracy Measure in percentage** | 81.25 | 87.5 | 93.75 |

TABLE IX. FINAL BI-GRAM

| **Profile Length** | 50 | 100 | 200 |
|---|---|---|---|
| **Accuracy Measure in percentage** | 56.25 | 62.5 | 62.5 |

TABLE X. FINAL TRI-GRAM

| **Profile Length** | 100 | 200 | 500 | 700 |
|---|---|---|---|---|
| **Accuracy Measure in percentage** | 75 | 56.25 | 75 | 62.5 |

From all above experimental results, it shows an unambiguous picture that in assessment with other types of positional n-grams, initial n-gram approach declares good result to identify the identity of an anonymous text. Results of initial n-gram show the utmost correctness of 100%; on the other hand supplementary n-grams achieve lower accuracy level. The method is very successful on this data-set.

## CONCLUSION

In this paper, a small experiment demonstrates the effectiveness of initial n-grams in reclaiming the mortal is more accurate than other positional n-grams extracted from words. This experiment utters first-class eminence result than medial n-gram, final n-gram and total n-gram. Since, we established our approach in small data-set & obtained a good piece of state so, our future potential will be to enlarge our corpus with a high anticipate to get valuable and accurate results. This can be enhanced to higher scale and also it varies with different sizes of n-gram.

## REFERENCES

[1] H. Baayen, H. V. Halteren, A. Neijt, and F. Tweedie, "An experiment in authorship attribution", 6es Journees internationales d'Analyse statistique des Donnees Textuelles, St Malo, France, March 2002.

[2] J. H. Clark, C. J. Hannon, 2007, "An Algorithm for Identifying Authors Using Synonyms", Proceeding ENC'07 Proceedings of the Eighth Mexican International Conference on Current Trends in Computer Science IEEE Computer Society Washington, DC, USA.

[3] R. Williams, S. Gunasekaran, W. Patterson, "On the Development of Digital Signatures for Author Identification ",First IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS 2007, pp 1-5.

[4] F. Peng, F. Shuurmans, V. Keselj, S. Wang," Language Independent Authorship Attribution Using Character Level Language Models", In Proc. of the 10th European Association for Computational Linguistics (2003).

[5] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," Literary and Linguistic Computing, 22(3), pp 251-270, 2007.

[6] Kjell, B., "Discrimination of authorship using visualization," Information Processing and Management, 30(1), 141-150, 1994.

[7] Stamatatos, E., "Ensemble-based author identification using character n-grams," In Proceedings of the 3rd International Workshop on Text-based Information Retrieval, pp. 41-46, 2006.

[8] Institute for Linguistic Evidence, I. 2008. Institute for linguistic evidence - mission & philosophy. http://www.fermentas.com/techinfo/nucleicacids/maplambda.htm

[9] M. A. Chaurasia, Dr. S. Kumar, "An Empirical Study on Author Affirmation" in International Journal of Electrical & Computer Science (IJECS/IJENS), Vol. 11, Issue 1,UK, in press.

[10] V. Keˇselj, F. Peng, N. Cercone, and C. Thomas. "N-Gram-Based Author Profiles For Authorship Attribution". In PACLING'03, August 2003, pages 255–264.

[11] P. Juola, "Ad-hoc authorship attribution competition," In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, pp. 175-176, 2004.

[12] M. Mansur, N. UzZaman, M. Khan, "Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus "Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh, 2005