# Design and Implementation of ASCII based Method for Author Attribution

**Monali P. Mohite, S. Renuka Devi**

***Abstract:*** ***With the presence of computer and internet, a developing variety of hoodlums are utilizing the web to spread a wide extend of illicit materials and wrong information universally in mysterious manner, making criminal personality following troublesome in the cybercrime examination handle. The virtual world provides criminals with an anonymous environment to conduct malicious activities such as malware, sending random messages, spamming, stealing intellectual property and sending ransom e-mails. All of these activities are text in somehow. Therefore, there is a need for a tool in order to identify the author or creator of this criminality by analyzing the text. Text-based Authorship Attribution techniques are used to identify the most possible author from a bunch of potential suspects of text. In this paper, the novel approach is presented for authorship attribution in English text using ASCII based processing approach Using this ASCII based method for authorship attribution help us to obtain better result in terms of accuracy and computational efficiency. The result is based on the text which is posted on social media considering real world data set.***

***Index Terms: Computer Forensic, Social Media Forensics, Digital Evidence, Forensic Investigation, Authorship Attribution***

## I. INTRODUCTION

Cybercrime, especially through the Internet, has grown in importance as the computer has become central to commerce, entertainment, and government. Senders can hide their identities by forging sender's address; Routed through an anonymous server and by using multiple usernames to distribute online messages via different anonymous channel. The researchers in the field of information security are interested in finding automated methods in order to determine the author of anonymous texts based on detecting some textual features. The problem of anonymity in a text is addressed by applying Authorship Analysis (AA) techniques [1]. It could be useful in many applications such as online data security and forensic analysis and assuring that the text data follows the guidelines and rules of styling. [1] (e.g. the text is a source code). Authorship analysis and its attribution have various societal applications and considered as the potential areas of research in the field of Stylometry which deal with identification of the author(s) of any given text by using several different approaches or techniques of text data mining. Obviously, longer the text, always better is the identification accuracy expected [2]. In paper [11], a framework for authorship detection is given. Authors have compared three different types of techniques for the identification strategies. Decision tree, feedback neural network and support vector machine is considered. Writing style features extraction is main technique followed. In paper [12], authorship detection task is elaborated for detection cross style of writing. In paper [13], authors describe methods used for pre-processing of the input text. The feature extraction Along with machine learning methods are shown. F1 scores based comparative analysis is given. In paper [14], authors have focused on classifiers such as random forest, SVM for authorship detection with respect to class. In paper [15], authors have given n-grams based stylometric features extraction technique. In paper [16], authors have given neural network based Short message authorship detection with training and testing platform and performance evaluation with respect to count of users. In paper [17], authorship detection with respect to meaning of words as degree of measure is used. In paper [18], recurrent interconnection patterns amongst words features are used for authorship detection. In paper [19], a survey of applications of Stylometry technique is given in authorship detection point of view. The subject area of Stylometry or authorship recognition/detection has several related fields of research such as:

1. Authorship Attribution (AA) or identification, i.e. identifying the author(s) of a set of different texts;
2. Authorship Characterization (AC), i.e. collection of data of author details such as age, gender, occupation etc;
3. Authorship Verification (AV), i.e. checking whether a text or set of similar texts is written or not written by an author;
4. Authorship Discrimination (AD), i.e. checking of two different texts whether those are written by the same author;
5. Plagiarism Detection (PD), i.e. identify the partial or full text length for copied text from other authors text materials;
6. Text Indexing and Segmentation (TIS), i.e. to generate the global book type document by concatenating variety of text segments from various authors text data and
7. Authorship De-Identification (ADI), i.e. to identify features which can properly capture an author's writing style.

Some of the commonly used Stylometry features to capture an author's writing style are namely,

1. Lexical features, i.e. text is viewed as a sequence of tokens grouped in to sentences;
2. Syntactic features, i.e. patterns used to form sentences;
3. Structural features, i.e. how an author organizes the structure of the document;
4. Content-specific features, i.e. characterization of certain activities, discussion forum or interest groups with few key words and so on.[4]

With reference to above mentioned authorship detection types, variety of methods are available which

**Revised Manuscript Received on July 05, 2019.**

**Monali P. Mohite**, School of computing science and engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India.

**S. Renuka Devi,** School of computing science and engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India.

constitute main part of feature vector formation. In most of the techniques, pre-processing part is introduced to create text vectors with removal of special symbols, spaces and creation of vectors with specific sizes. The main approach to perform pre-processing can be seen that to have better feature extraction perspectives. The features of the text can be of type syntactical, structural, textural, sequential, and many more which are mainly focusing to process text data and extracting features. Almost all the approaches are responsible for text data processing and hence show less scope for mathematical approaches as most of them are applicable for numerical data. The idea behind addressing this perspective consideration is to test the numeric conversion of text using some fixed numbering standard such as ASCII and applying mathematical approaches to have faster processing with less computational requirements and better feature extraction strategies for improved performances.

## II. ARCHITECTURE OF TEXT-BASED FORENSIC ANALYSIS APPROACH FOR AUTHORSHIP ATTRIBUTION

Figure 1 shows classic approach to model authorship attribution problem [3].
Step1: Input Text/ Data Collection:- Collect online messages written by potential authors from online communication.
Step2: Data Pre-Processing:- Pre-process the input text and leave the useful texts of information to be analyzed.
Step 3: Feature Extraction:-Features are extracted and presented as a vector as a measure of writing style.
Step 4: Model Generation:- Splitting of data for testing and training for various classification techniques with desired count of iterations.
Step 5: Frequency based occurrence and then finding similarity between two documents can be done by representing text as a vector.
Step 6: Classification: a learning process to measure the similarity index with distance between the two documents for known dataset.
Step 7: Author Identification:- Using developed model from now dataset for applying on new dataset.

## III. A NOVEL APPROACH FOR AUTHORSHIP ATTRIBUTION

### A. *Preprocessing*

To extract features of the text data, we fragment all the sentences from a set of sentences (paragraph) into three words based segments. As three words is easiest way of identifying meaningful sequence of words compared to less meaningful two words fragments and more complex for four words combination.

### B. *Converting text information to numeric data for establishing mathematical processing*

The processing of any features requires features to be in numeric format, we make use of ASCII based values of each character in text information including special symbols in the words and excluding space in the words to obtain numeric data.

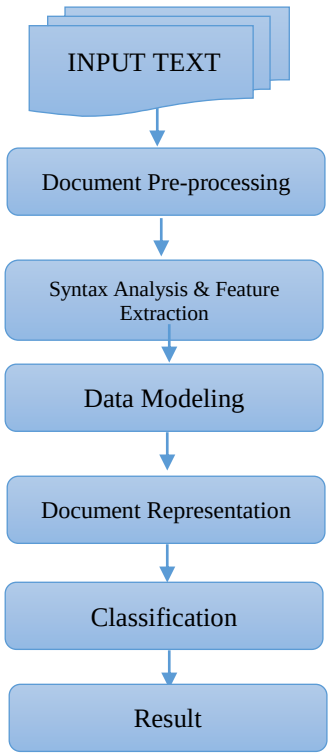


**Figure 1: Classic approach to model authorship attribution problem**

### C. *The feature vector formation*

The feature vector for any text with sequence of words (in our case three words) can be formed with following steps and respective mentioned presumptions. ASCII values for each word and each character in a word can be bunched together. The drawback in (1) is that, the words with different lengths with different number of characters will have different ASCII value lengths when bunched together. This leads to some ambiguity while establishing relation amongst multiple vectors with different sizes. To overcome the drawback mentioned above in (2), we add ASCII value of each character to form a single numeric value for single word. Each word in English will have its own value which is not being expected to be unique.

e.g.
Consider the set of three words, ‘parrot is green’
Consider the word ‘parrot’
ASCII values vector M= {112 97 114 114 111 116}
Single numbered value,

$$s = \text{sum}(M) = 564$$

Here 564 will not be unique value if observed and compared with all other values obtained. The exclusion of spaces in words is not having any impact as fixed number of spaces is present in each set of three words. After calculation of sum of ASCII vector for all the three words in each vector of words, we add a length of characters in that word after each sum vector to form vector of feature for set of three words.

e.g.
Consider the set of three words, ‘parrot is green’
Sum based vector T = {564 220 529}
Feature vector F = {564 6 220 2 529 5}

Where 6, 2, 5 numbers in feature vector F, shows number of characters in each word.
In generalized form we can deduce the feature vector F as,

$$F = \{s1\ ls1\ s2\ ls2\ s3\ ls3\}...(2)$$

Where, s1, s2, s3 represent sum of ASCII of each word and ls1, ls2 and ls3 represent number of characters in each word.

### *D. Feature matrix formation*

While establishing the relation between two sets of text (paragraphs) irrespective of the length of each text set, there is need of forming the two vectors to be formed in this method. When feature vector of each text set is to be formed, it has to be fixed with particular dimensionality for all text sets to reduce the computational complexities. For fixed dimensional approach, we consider that there will be 10 feature vectors, in which each vector is of three words sets features, obtained from each text set with presumption that, the text set is in the category of short information. To form the matrix of feature vector, we concatenate each feature vector of three words as a each row of matrix. The dimension of each feature vector obtained in (2) is 1X6. By concatenating 6 feature vectors, the obtained matrix will have dimensions, 6X6 which is square matrix.
The generalized form of this feature matrix can be given as,

$$Fm = \begin{bmatrix} f1 \\ f2 \\ f10 \end{bmatrix} ...(3)$$

Where, Fm represents feature matrix of feature vectors f1,f2,…,f10.

We perform conditional analysis of the matrix obtained in (3) for fixed dimensions,
Condition1: The word count in the text set can form exact number of feature vectors to form exact 10X6 size of matrix.
Condition 2: The word count in the text set has less number of words which leads to shortfall in formation of 6X6 sized matrixes.
Condition 3: The word count is greater than desired which is responsible to form matrix exceeding the dimension 6X6.
At first we consider first two conditions and later on we will evaluate strategy for third condition. As only second condition requires additional processing while in first phase work, we pad zeroes at the end of the matrix elements to form exact 6X6 dimension of the matrix. e.g. consider there are 6 feature vectors obtainable as if only 18 words are present in the text set. Therefore for feature vectors f7 to f10 we consider zeroes in the respective places. The condition occurring above shows the requirement of solving the matrix to form unique solution. After solving the matrix it will deduce the requirement of minimum number of words in particular text set to have optimal solution of the matrix along with method of obtaining the solution.

### *E. Solving the matrix to find the unique solution.*

To establish the relation between two matrices, we will solve the matrix generated in section 'D' using two approaches.

### *F. Estimating Determinant Value of Matrices*

For the authorship detection of entire text structure using entire matrix can be done by comparing two matrices. The partial text authorship detection problem still persists. The matrix fragmentation can result in number of matrices and resulting in more computational requirements. The fragmented matrices thus further be solved to estimate the determinant values and the choice of comparison these values can result in selective comparison process thereby reducing the computational requirements. The matrix fragmentation thus can be structured commonly for all matrices which result in fixed structure of processing. For the sake of simplicity in development we consider 42x6 as a dimension of original matrix developed by using equation (3). The matrix fragmentation thus we consider for generating 7 different small matrices with dimensions 6x6 to keep dimension of each matrix as square. Each of these matrices will constitute 18 words as per equation (2). These matrices can be then used to estimate the matching which will provide unique match estimation for the bunch of 30 words. The set of matrix fragmentation can be given as,

$$Fn = \{m1,m2,...,m10\}...(4)$$

Where, m1,m2,…,m10 represent fragmented matrices of matrix Fm such a that each matrix is squared matrix and can be used to estimate the determinant value easily.
The sets obtained from equation (4) for two text sets then can be used to calculate the determinant values and then estimating the distance between the two.
The formula for obtaining determinant value of 2x2 matrix can be given as,

$$D = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc...(5)$$

Equation (5) gives idea of finding determinant value of matrix size 2x2, the idea can then be extended to find determinant value for matrix with dimensions 6x6. The matrix which is not having features count such as ending or last matrices, such matrices can be converted into square matrices by padding 0's at the end.

***Algorithm:***

1. *Get input text vector F1(s) and F2(s) for two text sets*
2. *Convert Given text to Ascii format matrixes such that F1m=[f1;f2;f3;…] and F2m=[f1;f2;f3;…]*
3. *Generate fragmented matrix vectorsF1n = {m1,m2,…,m10} and F2n = {m1,m2,…,m10}*
4. *Calculate the determinant value of each matrix in F1n and F2n to get D1n and D2n vectors.*
5. *Estimate the Bhattacharya distance between two vectors (D1n and D2n) to get distance vector BC*
6. *If BC(n) is zero then there is exact match where n=1,2,3…*

## IV. ESTIMATING THE DISTANCE FOR SIMILARITY ESTIMATION

The distance estimation process will provide the similarity estimation between two input vectors. The distance estimation should be reliable in terms of effects of standard deviation and distribution effects and hence Bhattacharya distance estimation of two vectors for similarity estimation. In case of measuring the similarity index there are variety of possible techniques such as Euclidean distance, Manhattan distance, cosine similarity. Bhattacharya distance not only measures the similarity but also is capable of showing the overlapping regions between two vectors. This overlap in our application can provide sufficient understanding that whichpart of the text is similar between the two.
Therefore, the two distance estimations can be given by,

$BC(D_1(n), D_2(n)) = \sum\sqrt{D_1(n), D_2(n)}$ ...(6)

Where D1(n) and D2(n) are matrices sets of two sets of text to be compared. Here we have established the mathematical illustration of identifying authorship of text sets thereby establishing relationships and identification of similarity between them using ASCII based processing approach which will reduce computational complexities and time consumption compared to text processing based approaches. The BC provides the vector with distance of each member of D1(n) and D2(n) vectors from each other. The larger the value less is the similarity and lesser or zero the value similarity is nearer or exact.

## V. RESULTS

The proposed work is implemented in python for processing. The two text sets considered are shown in figure 2 and 3. The part of text set 1 is included at the start of text set 2 as can be seen from figure 2 and 3. The text data is read and converted into ASCII matrix the part of matrix can be seen in figure 4. From figure 2, first word is 'The' whose ASCII values are '84 104 101' and count of letters in this word is '3'. Hence, part of the matrix shown figure 4 has starting value '289' and '3' where 289 is summation of Ascii values and 3 is count of letters. The remaining values follow the similar justification.

289 3 627 6 1340 12

363 3 457 4 259 2

353 3 547 5 653 6

1294 12 245 2 353 3

688 6 379 4 215 2

353 3 479 4 378 3

**Figure 4: Ascii-sum Features matrix with 6x6 dimension**

The Indian subcontinent was home to the Indus Valley Civilisation of the bronze age. In the next two millennia, the oldest scriptures of Hinduism were composed, social stratification based on caste emerged, and Buddhism and Jainism arose. Political consolidations took place under the Maurya and Gupta Empires. The peninsular Middle Kingdoms influenced the cultures of Southeast Asia. In India's medieval era, Judaism, Zoroastrianism, Christianity, and Islam arrived, and Sikhism emerged, adding to a diverse culture. North India fell to the Delhi Sultanate; south India was united under the Vijayanagara Empire. In the early modern era, the expansive Mughal Empire was followed by British East India Company rule. India's modern age was marked by British Crown rule and a nationalist movement which, under Mahatma Gandhi, was noted for nonviolence and led to India's independence in 1947.

**Figure 2: Text set 1**

The Indian subcontinent was home to the Indus Valley Civilisation of the bronze age. In the next two millennia, the oldest scriptures of Hinduism were composed, social stratification based on caste emerged, and Buddhism and Jainism arose. Political consolidations took place under the Maurya and Gupta Empires. The peninsular Middle Kingdoms influenced the cultures of Southeast Asia. In India's medieval era, The Godavari and the Krishna are the two major rivers in the state. The Narmada and Tapi Rivers flow near the border between Maharashtra and Madhya Pradesh and Gujarat. Maharashtra is the third-most urbanized state of India. Prior to Indian independence, Maharashtra was chronologically ruled by the Satavahana dynasty, Rashtrakuta dynasty, Western Chalukyas, Deccan sultanates, Mughals and Marathas, and the British. Ruins, monuments, tombs, forts, and places of worship left by these rulers are dotted around the state. They include the UNESCO World Heritage Sites of the Ajanta and Ellora.

**Figure 3: Text set 2**

Two matrices obtained similar to figure 4 for two text sets are processed to find values of Bhattacharya distances between all elements of two matrices. As part of text is matching, the value of distance will be zero for this matching size of text feature matrix. Table I shows the Bhattacharya distance values for elements of two feature matrices. From TABLE I, it can be observed that, first 3 elements show distance as zero which are most matching words from two text sets.

**Table I: Bhattacharya Distance for two text sets**

| Determinant value of feature set 1 | Determinant value of feature set 2 | Bhattacharya Distance |
|---|---|---|
| 227292 | 227292 | 0 |
| -46011727 | -46011727 | 0 |
| -62147488 | -62147488 | 0 |
| 204721863 | 34367071 | 170354792 |
| -86368767 | 15986796 | 70381971 |
| 20341278 | 362030473 | -341689195 |

As each matrix determinant value is for 6x6 feature vectors, it means for the value is for 18 words. As such three matrices match in determinant values, total 54 words are similar in two text sets. These text sets are thus compared in very simplistic computations and similarity is estimated which satisfies the objective with outstanding performance.

## VI. CONCLUSION

In this work, we have proposed a novel approach for processing text data in numeric format for authorship attribution. The main perspective of the proposed work is to reduce the computational complexities compared to text data processing and achieving maximum performance in terms of accuracy and computational efficiency. The results obtained through implementation in python

based processing, provides satisfactory results. The approach can be helpful for further development for the researchers in this era and can form the platform for upcoming intuitions and inventions.

## REFERENCES


1. BushraAlhijawi_, SafaaHriezy, Arafat Awajan," Text-based Authorship Identification - A survey",IEEE ISIICT 2018.
2. Siddharth Swain, Gaurav Mishra and C. Sindhu, "Recent Approaches on Authorship Attribution Techniques – An Overview", IEEE International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.
3. Jianbin Ma, Ying Li, GuifaTeng "CWAAP: An Authorship Attribution Forensic Platform for Chinese Web Information" in JOURNAL OF SOFTWARE, VOL. 9, NO. 1, JANUARY 2014.
4. Waheed Anwar, Imran SarwarBajwa "An Empirical Study on Forensic Analysis of Text using LDA based Authorship Attribution", IEEE Access 2018.
5. PelinCanbay, Hayri sever, EbruAkcapninarsezer, "Determining of Discriminative Block Sizes for Authorship Attribution on the Turkish Text", IEEE 2018
6. JurgitaKapociuteDzikiene, AlgimantasVenˇckauskas, RobertasDamaševiˇcius, "A Comparison of Authorship Attribution Approaches Applied on the Lithuanian Language", Proceedings of the Federated Conference on Computer Science and Information Systems 2017.
7. M. Tahmid Hossain, Md. Moshiur Rahman, Sabir Ismail, MdSaiful Islam, "A Stylometric Analysis on Bengali Literature For Authorship Attribution",20th International Conference of Computer and Information Technology (ICCIT),2017.
8. JianPeng , Kim-Kwang Raymond Choo, Helen Ashman, "Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution", IEEE TrustCom/BigDataSE/ISPA
9. 2016.
10. Jianbin Ma, Ying Li, GuifaTeng "CWAAP: An Authorship Attribution Forensic Platform for Chinese Web Information" in JOURNAL OF SOFTWARE, VOL. 9, NO. 1, JANUARY 2014.
11. RituBanga, PulkitMehndiratta,"Authorship Attribution for textual data on Online Social Networks", Tenth International Conference on Contemporary Computing ( IC3) 2017.
12. Li, J., Chen, H., & Huang, Z. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and classification Technique", Journal of the American Society for Information Science, 57(3), 378–393. doi:10.1002/asi,2006.
13. Mike Kestemont et al , "Cross-domain Authorship Attribution and Style Change Detection", Overview of the Author Identification Task at PAN-2018.
14. Yaakov HaCohen-Kerner et al., "Cross-domain Authorship Attribution: Author Identification using Char Sequences, Word Uni-grams, and POS-tags Features", Notebook for PAN at CLEF 2018.
15. Daniel Kopev et al., "Recursive Style Breach Detection
16. with Multifaceted Ensemble Learning", arXiv:1906.06917v1 [cs.CL] 17 Jun 2019
17. Daniel Kara´ et al., "OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection", Notebook for PAN at CLEF 2017
18. R. Ragel, P. Herath and U. Senanayake, "Authorship detection of SMS messages using unigrams," *2013 IEEE 8th International Conference on Industrial and Information Systems*, Peradeniya, 2013, pp. 387-392.
19. V. Vysotska, Y. Burov, V. Lytvyn and A. Demchuk, "Defining Author's Style for Plagiarism Detection in Academic Environment," *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, 2018, pp. 128-133.
20. V. Q. Marinho, G. Hirst and D. R. Amancio, "Authorship Attribution via Network Motifs Identification," *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, Recife, 2016, pp. 355-360.
21. Tempestt Neal, KalaivaniSundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.* 50, 6, Article 86 (November 2017), 36 pages.


## AUTHORS PROFILE

**Monali Pradeep Mohite ,** received her Bachelor of engineering degree in Information Technology from RTMNU Nagpur, India in 2013 and her Master of Technology degree in Computer Science and Engineering from Yeshwantrao Chavan College of Engineering,Nagpur,India in 2015. She is Ph.D. Student in School of Computing Science and Engineering in Vellore Institute of Technology, Chennai, India. Her current research areas include Machine Learning and Social Media Forensics.

**S. Renuka Devi,** received her doctorate degree in Information and Communication Engineering from Anna University, Chennai, India in the year 2015. She is currently working as an Associate Professor in the School of Computing Science and Engineering, VIT University, Chennai, India. Her area of interest includes Network Security and Cryptography.