IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728,p-ISSN: 2319-765X, Volume 8, Issue 4 (Sep. - Oct. 2013), PP 42-50
www.iosrjournals.org

Literary Analysis using CUSUM Technique on Bharathiar Writings

Arbind Kumar Singh[1] and G. Manimannan[2]
[1]*Senior Lecturer, Department of Statistics, S.N.S.R.K.S College Saharsa, B. N. Mandal University, Madhepura, Bihar*
[2]*Assistant Professor, Department of Statistics, Madras Christian College, Chennai.*

Abstract: *Stylometry is an attempt to capture the essence of the style of a particular author by reference to a variety of quantitative criteria, usually lexical in the nature, called discriminators, or more succinctly the statistical analysis of literary style. A written passage of any kind can be analysed by the method called CUSUM analysis. This analysis reveals whether an article is written by one person or more than one person. In this paper, an attempt is made to authorship attribution on the basis of CUSUM technique to certain articles written on Indian freedom movement published in the magazine called* ***India.*** *Seven articles written by renowned Tamil poet Bharathiar and another six articles not attributed to any author, but belonging to the same period, are considered for authorship identity in the present study. The three features of writings used in this analysis are (i) the use of the 2, 3 and 4 letter words, (ii) words starting with a vowel and (iii) the third combination of these two together. Among the six unattributed articles, CUSUM analysis establishes that all of the writings are very close to Bharathiar's style. This result supported the claims made by many scholars that these six articles could have been written by Mahakavi Bharathiar (MB).*
Keywords: *Authorship, Stylometry, CUSUM Analysis and Test of divergence.*

I. Introduction

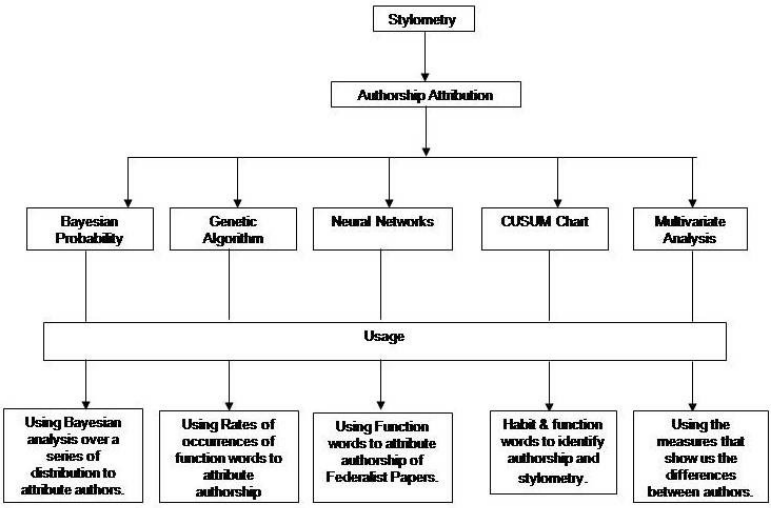
It is possible to attribute a text to a particular author if the general knowledge of writer's style is known. This must be distinguished from what the author writes as, although authors tend to write on similar topics, the general hypothesis must be that each piece of text an author writes is contextually independent, so making the issue of how an author writes is of prime importance. This important concept may be described as an author's style. Therefore, stylometry can be thought of as the study of measuring an author's style.

Stylometry describes style by quantifying aspects of the writing that can be directly measured; sentence length (in words) for example (Tom Ashford, 2001). If stylometric methods are to be useful in measuring an author's style, some assumptions must be made to ensure their reliability. The first main assumption must be that individuals have their own consistently unique writing style. There must always be measurable aspects of an author's style that never change. If an author's writing style is not consistently unique, it will be impossible to distinguish between authors. Another assumption is that authors have both conscious and unconscious aspects to their writing style.

Tweedie and Bayyen (1997) have argued that all stylometrics referring directly to vocabulary usage are subject to much variation for a given author and consequently these stylometrics are less reliable than syntax-based measures. Nowadays many researchers using statistical techniques have been outlined and experimented with Stylometry.

The problem of attributing authorship to text is known long even before the computer was invented and the introduction of computers has increased the power of computation and allows large corpora of text to be analysed in a short period of time.

McEney and Oaks (1996), describe most prevalent techniques of stylometry for authorship attribution and some of these techniques are listed below, along with some of the details of their implementations.



1.1 I Length of Text Analysis

Text analysis method is not considered for analyzing whole works of Literature. Farringdon (1996) suggests that the ideal sample is approximately fifty consecutive sentences and if there are too many sentences it is hard to notice divergences of the lines. One can *bore holes* in various parts of a piece of literature to tests if consistence is maintained within the work and this is recommended if reliable results are to be obtained.

1.2 Sentence Pattern

If the CUSUM method is automated, it is essential to precisely define what is meant by a word or a sentence. The earliest method of defining a word is by using spaces as delimiters that will separate them (Tom Ashford, 2001). The common method is to examine punctuation marks. A sentence can be ended either by a full stop or a question mark or an exclamation mark.

1.3 Semi Sentences

Sentences such as, 'Atu nāṭṭāṭci cuyarājyam′ and other similar ones are known as semi sentences and can cause problems for analysis. The string 'Kāṅkiras tantai āṅkilayār' is not a proper sentence, it cannot be considered for exhibiting styles and it merely adds factual content. These semi sentences can either be omitted, or added to a neighboring sentence. Name of the year, time and date can also cause the same problem. The easy solution to this problem is to remove year, date and time etc (Jill Farington, 1996). In the present research, an attempt is made to apply CUSUM technique to the problem of authorship attribution for articles of ambiguous authorship and to assign them to the contemporary writers of the same period. Two sets of variables such as *sentence length* and *habitual words* are made use of for convergence equation and divergence deduction for CUSUM. Subsequently, results of authorship attribution are discussed.

II. Database

The present study deals with the literary work of the famous Tamil poet Subramaniya Bharathiar popularly known as Mahakavi Bharathiar (1882-1921). He was a well-known poet and freedom fighter of the nineteenth century. He was the editor of *India*, a news magazine published in the year 1906. In this magazine Bharathiar and other writers have written anonymously articles, editorials and short stories. Ilasai Maniyan (1975) has compiled all these articles and has brought out a book entitled *Bharathi Dharisanam*. Six articles from this compiled book are taken up for author attribution in this study. To identify Bharathiar's style, it is also necessary to identify and set aside elements, which have the common stylistic characteristics of the writers of the same period. In this connection we have considered seven articles written by the poet himself on the same topic in other magazines of the same period. All these thirteen articles deal with the common topic, namely, India's freedom movement.

Out of these thirteen articles, seven were written by the poet Bharathiar himself which we designate as *knowns*, six were selected at random from Ilasai Maniyan's edited book, which are referred to as *unknown*. The numbers of sentences selected randomly from each of the thirteen articles are given in *Table 1*.

Table 1. Number of Sample Sentences Selected From Seven Articles

Articles	Samples	Articles	Samples
Known		***Unknown***	
Article 1	20	Article 8	20
Article 2	20	Article 9	20
Article 3	20	Article 10	20
Article 4	20	Article 11	20
Article 5	20	Article 12	20
Article 6	20	Article 13	20
Article 7	20		
Total	140	Total	120

We have considered the three most common features of writings used in the analysis: (i) the use of the 2, 3 and 4 letter words, (ii) words starting with a vowel and (iii) combination of these two. These three variables are identified for each sentence. If we have n sentences and if we identify p variables from each sentence then we have a data matrix of size n×p. Thus each article was converted as a data matrix and these data matrices form basis for the literary data analysis of this study. As there are thirteen articles, there are thirteen different matrices and the problem is to compare the data matrices of the ***unknown*** articles with ***known***. *Table 2 and 3* lists the sentence lengths and counts of two, three, four and vowel beginning words in articles of known and *unknown*. The most noticeable feature in both extracts is the wide variation in sentence length and habitual words.

III. Methodology

The CUSUM method also known as the QSUM technique did not begin as a stylometric analysis tool. This method is normally used in the manufacturing industry as a measure of quality control. The transfer of this technique to stylometry is a natural one. Instead of measuring deviations in product quality, the method is adapted to measure deviation in the stylometric measures found in a text. All kinds of written passages can be analysed with the help of this method.

3.1 Calculation of CUSUM Method

Main six distinct steps involve the calculation of CUSUM techniques are as follows:
Step 1. Counting the occurrences of measures (e.g. habitual words and two and three letter words)
Step 2. Calculating the mean values
Step 3. Calculating the sentence deviations from the mean values
Step 4. Calculating the cumulative sum of the deviations
Step 5. Plotting the graph and
Step 6. Interpretation of the graph

Step by step will describe and applied to an every article. The article comprises the first ten sentences of a Bharathiar, written to the magazine *India* in 1907.

3.2 Counting the Occurrences of the Measures

It‘s the simplest of the stages and involves examining each sentence in the sample text and creating a count for each measure based on their occurrences.

3.3 Calculation of the Mean Values

A further simple stage obtains the mean per sentence value for each discriminator.

Step 1. To calculate the mean sentence length $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} length\ (x_i)$

3.3 Calculating Sentence Deviations from the Mean Value

The sentences individual counts are then compared to the mean values. New values are produced for each sentence representing that sentences deviations from both measure’s mean values. In the passage, the first three sentence deviations from sentence length (**sl**) mean can be calculated.

Step 1. Calculate the differences between the lengths of each individual sentence and $\bar{x}$.

Step 2. To calculate the Sentence Deviation = $length\ \ (x_i) - \bar{x}$

Sentences	Words	Deviation Values	Inference
1	9	-3.9	Less than the mean
2	27	14.1	Greater than the mean
3	15	2.1	Greater than the mean
Mean sentence length: 12.9			

3.4 Cumulative Sum of the Deviations.

The steps in CUSUM process are discussed briefly in the following sections: The term *habit* is interchangeable with the term 'style'. The defining habit of an author is one that most clearly characterizes the style of an author; the occurrences of this habit are relative to how much the author articles. Jill Farington (1996) suggests 9 habits that may be useful for this technique. The three most common features of writings used in the analysis are (i) the use of the 2, 3 and 4 letter words, (ii) words starting with a vowel and (iii) combination of these two.

This stage takes the mean deviations of the sentences and creates a new cumulative sum value for each measure per sentence. This involves summing all the previous deviation values for that measure up to and including the sentence currently being calculated. For example the first three sentences of the passage would yield the following values for the sentence length measure.

Sentences	Words	Deviation Values	CUSUM
1	9	-3.9	-3.1
2	27	14.1	11.0
3	15	2.1	13.1
Mean sentence length: 12.9			

As one should expect, the cumulative sum of deviations sums to zero at the last sentence, as a zero value represents the mean for the sample, which the deviations will all inherently sum to. After that we calculate standardised cumulative sums. This step uses the same formulae in as before. Here the text is represented as the set of the sentences, $H_{w1}, H_{w2}, ..., H_{wn}$. H_{wi} is the number of occurrences of the defining habit found in sentence X_i.

Step 1. Calculate the mean value for the defining habit $\overline{H}_w$.

Step 2. Calculate the difference between the value of the habit for a given sentence and $\overline{H}_w$ diff. Habit word.

Step 3. Calculate a running sum of the differences from the mean CUSUM habit words.

IV. Result and Discussion

A graph is plotted for each sentence versus its cumulative sum deviation value. The graph will have two lines, one representing the standardized sentence length CUSUM values for each measures and another representing the standardized habit words CUSUM values for each measure. They will most likely not share similar deviation values, so the two lines are plotted on different Y-axes for easy comparison between the two lines. The mean value (zero) is plotted center way up the Y-axes as the CUSUM value always falls on both sides of zero. Care must be taken when deciding the scale for the graph. If the scale were too small (little variation in y-axis values) the result would be *flattened so as to produce a faintly waving line and hence lose all detailed information*. If the scale is too big (large variations in Y-axes values) the graph will have *grossly exaggerated peaks and troughs, giving far too much detail to be useful* (both quotes from Farringdon, 1996).

In order to conduct the style test for authorship, the relationship between the two-plotted graph lines was observed. If the two lines track each other closely, it is an indication of homogeneous authorship. This is due to the fact that the styles being measured remain consistent throughout the sample. If the graph line diverges significantly, it suggests that more than one author wrote the sample. *Table 2 and 3* lists the CUSUM sentence lengths and CUSUM habit words calculation in two categories of known and *unknown.*

TABLE 2 WORD COUNT DATA FROM BHARATHIAR FIRST KNOWN ARTICLE

Sample No.	Sentence Length	Habit Words	Standardised CUSUM of sentence length	Standardised CUSUM of habit words	Straight line equation
1	9	6	-.68	-.76	-.31
2	27	25	2.37	1.94	2.76
3	15	10	2.48	2.34	3.21
4	18	17	3.96	3.31	4.32
5	9	8	3.68	2.55	3.45
6	14	14	4.57	2.75	3.68
7	9	7	4.09	1.99	2.82
8	13	9	4.00	2.00	2.83
9	13	5	3.13	2.01	2.84
10	14	8	2.84	2.21	3.07
11	13	11	3.15	2.22	3.08
12	11	5	2.27	1.85	2.65
13	9	2	.81	1.09	1.79
14	8	8	.53	.13	.71
15	8	7	.05	-.82	-.38
16	10	6	-.63	-1.38	-1.02
17	14	10	-.52	-1.18	-.79
18	14	11	-.22	-.98	-.56
19	24	14	.68	1.14	1.86
20	7	6	.00	.00	.55
Total	259	189			
Mean	12.95	9.45			
Standard Deviation	5.20	5.10			

TABLE 3 WORD COUNT DATA FROM BHARATHIAR FIRST UNKNOWN ARTICLE

Sample No.	Sentence Length	Habit Words	Standardized CUSUM of sentence length	Standardized CUSUM of habit words	Straight line equation
1	10	6	-.58	-.77	-.73
2	23	18	1.81	1.30	.94
3	16	11	2.60	1.71	1.27
4	12	8	2.47	1.42	1.04
5	12	11	2.35	1.83	1.37
6	24	20	4.97	4.36	3.42
7	13	8	5.07	4.07	3.18
8	8	9	4.03	4.01	3.14
9	10	13	3.44	4.89	3.85
10	7	7	2.17	4.36	3.42
11	10	5	1.59	3.36	2.61
12	11	7	1.24	2.83	2.18
13	12	10	1.11	3.01	2.32
14	9	6	.30	2.24	1.70
15	12	7	.17	1.71	1.27
16	15	10	.73	1.89	1.42
17	11	6	.38	1.12	.80
18	13	9	.48	1.06	.75
19	14	12	.81	1.71	1.27
20	9	2	.00	.00	-.11
Total	251	185			
Mean	12.55	9.25			
Standard Deviation	4.37	4.24			

V. CUSUM Data Scaling

To plot two cumulative sums (CUSUMS) together, one of the CUSUMS has to be scaled by a scaling factor. An alternative technique of scaling the CUSUMS is used in this project. By this method the error between the two CUSUMS is minimized. For this method, the two separate CUSUMS are indicated as function $S(x)$ and $H(x)$, for a document of sentences $(S_1 \; to \; S_n)$, then, X ranges from $1 \text{ to } n$. Figure 1.0 shows a typical CUSUM graph, under this representation, before any scaling has occurred to either function. It is clear from the CUSUM graph that the two plotted functions are not well scaled together. To identify divergences between the two functions it is important the CUSUM graph is scaled appropriately (Mathew Stephen Hersee, 2001).

To find the optimal association, an *error function* is first defined, which characterizes the degree of divergence between the CUSUMS. Then the *scaling factor* a and a *raising factor* b, are calculated, such that $S(x) = a.H(x) - b$.

5.1 Divergences Detection for CUSUM Data

Fig 1.0 is a typical CUSUM graph; under a manual interpretation the graph has most significant divergences between sentences 9 to 13 (Table 4). This manual process of detecting divergences is automated by calculating the error between the two CUSUMS, at every sentence S_i the error E_i is calculated by the formula:

$$E_i = (CUSUMSent_i - CUSUMHabit_j)^2 .$$

TABLE 4 BASIC ERROR CALCULATIONS FOR SCALED CUSUM GRAPH

Sentences number	$CUSUMSent_i$	$CUSUMHabit_i$	$Error\,(E_i)$
1	-.68	-.76	0.01
2	2.37	1.94	0.18
3	2.48	2.34	0.02
4	3.96	3.31	0.42
5	3.68	2.55	1.28
6	4.57	2.75	3.31
7	4.09	1.99	4.41
8	4.00	2.00	4.00
9	3.13	2.01	1.25
10	2.84	2.21	0.4
11	3.15	2.22	0.86
12	2.27	1.85	0.18
13	.81	1.09	0.07
14	.53	.13	0.15
15	.05	-.82	0.76
16	-.63	-1.38	0.58
17	-.52	-1.18	0.44
18	-.22	-.98	0.59
19	.68	1.14	0.22
20	.00	.00	.00

The crucial stage now is to ensure that a divergence, between two CUSUMS, is large enough to be a *significant divergence*. Essentially there must be a cut off point, a threshold such that if E_i is greater than this threshold, and then there is a significant divergence at S_i, conversely if E_i less than this threshold, then the divergence at S_i is not significant. In essence the threshold classifies the document into a set G, where G_i equals '1' if E_i is above threshold *positive calculation* or *-1* otherwise *negative classification.* To calculate static threshold a *cut off* factor C is required. The threshold functions of the average error, $\overline{E}$ and the constant C. For error E_i, at sentence S_i, average error $\overline{E}$ and cut off C. If $E_i \geq C * \overline{E_i}, S_i$ receives positive classification, S_i receives negative classification. Table 4 a static threshold where cut off C = 2. Table 5 illustrates the sentences 6, 7 and 8 diverges above static threshold.

TABLE 5 ILLUSTRATION OF A STATIC THRESHOLD

Sample Sentence	*Error* (E_i)	$E_{\geq} \geq C^* \overline{E}\,(1.92)$	*Classification* (G_i)
1	0.01	NO	-1
2	0.18	NO	-1
3	0.02	NO	-1
4	0.42	NO	-1
5	1.28	NO	-1
6	3.31	YES	1
7	4.41	YES	1
8	4.00	YES	1
9	1.25	NO	-1
10	0.04	NO	-1
11	0.86	NO	-1
12	0.18	NO	-1
13	0.07	NO	-1
14	0.15	NO	-1
15	0.76	NO	-1
16	0.58	NO	-1
17	0.44	NO	-1
18	0.59	NO	-1
19	0.22	NO	-1
20	0.00	NO	-1

The above results clearly explains the analyses and provides the allocation of authorship of Bharathiar's known and unknown articles.

FIG 1.0 TO 1.6 SAMPLE CUSUM GRAPH FOR THE FIRST TWENTY SENTENCES OF AN ARTICLE OF BAHRATHIAR'S WRITTEN IN 1906 (SCALED AND UNSCALED)

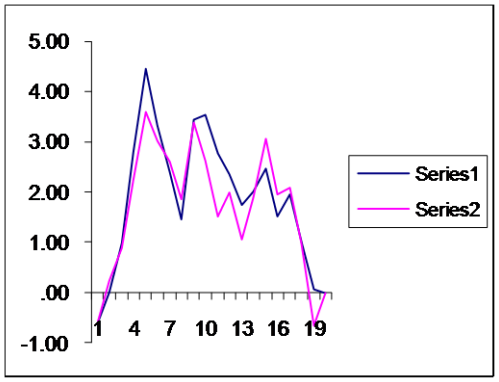


Fig 1.0

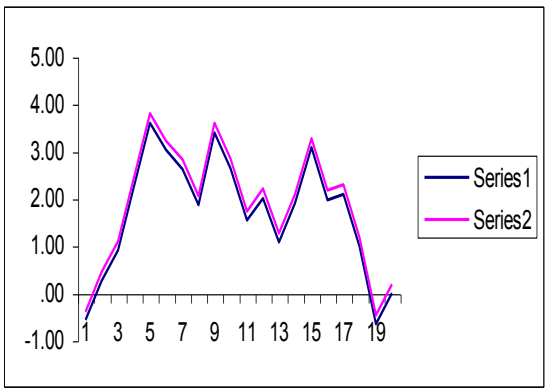


Fig 1.1

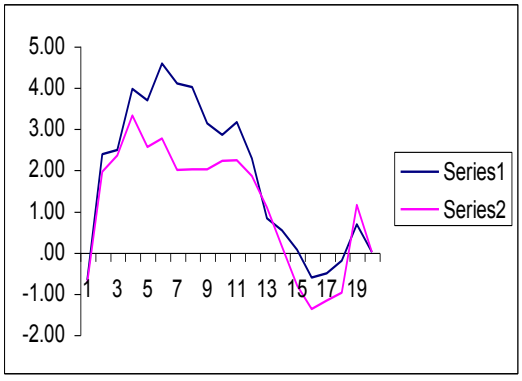


Fig 1.2

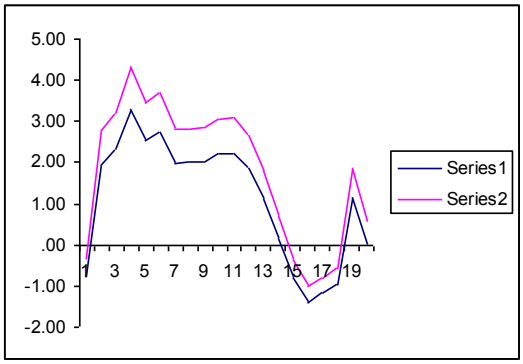


Fig 1.3

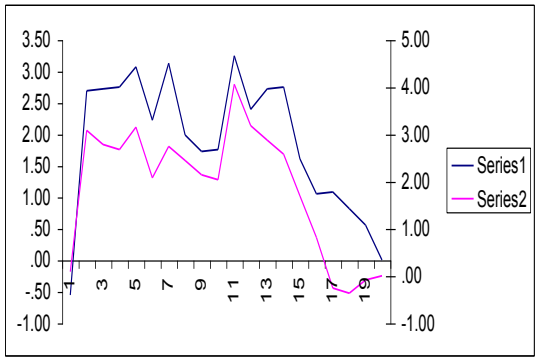


Fig 1.4

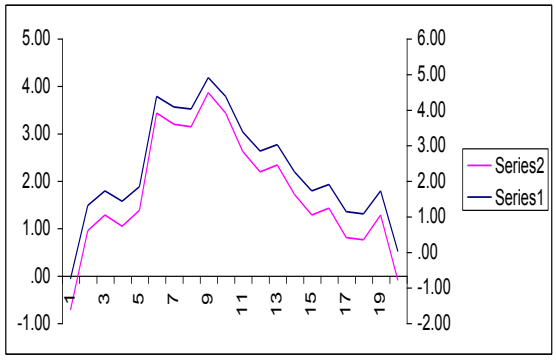


Fig 1.5

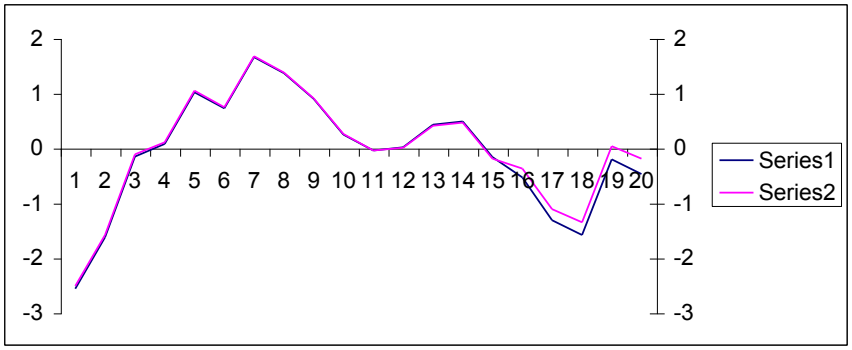


Fig 1.6

Fig. 1.0 indicates that the author is consistent throughout the two pieces, but there is slight disturbance visible at sentence 3, 7, 15, 17 and 19. This can be accounted for the change in context. Fig. 1.1 shows that there is only a small divergence at the joining point to end point of these two samples, significantly small enough to represent homogeneous authorship.

Even if Fig. 1.2 shows the most diverged graph, the lines still track each other very closely, especially at the joining point of the samples. This is also supposed to be a strong evidence of single authorship. From Fig.1.3 it is clear that the scaled sentence length CUSUM remains consistent but slightly divergence above the CUSUM for habit words.

Although Fig. 1.4 and 1.5 show the most diverged graph, the lines still track each other very diverging, especially at the joining point to end point of the samples. This is a strong evidence of single authorship.

Fig. 1.6 shows that there is only a small divergence at the end point of these two samples, significantly small enough to represent homogeneous authorship.

VI. Conclusion

Assignment of articles of ambiguous authorship to the contemporary Tamil scholar, namely Mahakavi Bharathiar (MB), is taken up in the present research. The oppressive attitude of the then British regime compelled all the patriots to write articles on the same theme for anonymous publications without mentioning their names. A written passage of any kind can be analyzed by the method called CUSUM analysis. This analysis will reveal whether a passage is written by one person or more than one person. In this paper, an attempt is made to authorship attribution on the basis of CUSUM technique to certain articles written on Indian freedom movement published in the magazine called **India.** Seven articles written by renowned Tamil poet Bharathiar and another six articles not attributed to any author, but belonging to the same period, are considered for authorship identity in the present study. The three features of writings used in this analysis are (i) the use of the 2, 3 and 4 letter words, (ii) words starting with a vowel and (iii) the third combination of these two together. All the articles are attributed to Mahakavi Bharathiar (MB). This result supported the claims made by many scholars that these six articles could have been written by Mahakavi Bharathiar (MB). Our recent research also supported the claim that these articles with ambiguous authorship were written by Mahakavi Bharathiar (MB).

References

[1] A. F. Bissell (1995), Weighted Cumulative Sums for Text Analysis using Word Counts, *Journal of Royal Statistical Society*, Series A, 158, part 3, pp. 525-545.
[2] J. Tweedie, and R. H. Baayen (1996), Lexical *constant* in stylometry and authorship studies www.cs.queensu.ca/achallc97/papers/s004.html
[3] J. Farrington (1996), How to be literary detective: Authorship Attribution, members.aol.com/qsum/QsumIntroduction.html.
[4] Manimannan G. and Bagavandas. M (2001), The authorship attribution: the case Bharathiar, presented at National conference on Mathematical and Applied Statistics, Nagpur University, Nagpur.
[5] Mathew Stephen Herse (2001), Automatic Detection of Plagiarism: An Approach Using the Qsum Method, University of Sheffield.
[7] T. McEnery and M, Oakes, *Lancaster University*, Authorship Identification and Computational Stylometry**.**
[8] Tom Ashford (2001), Computerised Determination of Disputed Authorship: The CUSUM Method. www.dcs.shef.ac.uk/intranet/teaching/public/projects/archive/ug2001/pdf/u8tja.pdf