Tareef Kamil Mustafa*

Non-word Attributes' Efficiency in Text Mining Authorship Prediction

https://doi.org/10.1515/jisys-2019-0068
Received March 13, 2019; previously published online April 24, 2019.

Abstract: Literature scripts can be compared to paintings, in an artistic way as well as in the perspective of financial value, whereas the value of these scripts rise and fall depending on their author's popularity. Authors' scripts represent a specific style of writing that can be measured and compared using a text mining field called Stylometric. Stylometric analysis depends on some features called authorship attributes, and these attributes or features can be used in special algorithms and methods to reach that aim. Generally, each method selected in the Stylometric field uses a variety of attributes to reach higher prediction accuracy. The aim of this research is to improve the accuracy of authorship prediction in literary works based on the artistic writing style of the authors. To achieve that, a new set of attributes will be used with the Stylometric Authorship Balanced Attribution method, which was chosen in this research among several other machine language methods because of its delicateness in authorship prediction projects. The attributes that have been used by most of the researchers were word frequencies (single word, pair of words, or trio of words), which led to some prediction mistakes. In this research, a new set of attributes is used to decrease these mistakes. These proposed non-word attributes are named sentence length, special characters, and punctuation symbols. The results obtained by using these proposed attributes were excellent.

Keywords: Machine learning, Stylometric, authorship attribution, SABA, non-word attribute.

1 Introduction

Stylometric in general is the criteria for analyzing the writing style and language habits in any text [8]. It can be implemented on many simple or rather complex issues in real life. One of them is the Stylometric authorship, which is the concept of recognizing an unknown text (its writer is not guaranteed or known) to determine to whom it belongs in a set of nominated authors [9]. Usually, this subject refers to the statistical analysis of the literary style of authors based on the characteristics of their writing or expression [14]. To reach that goal, some attributes that describe and summarize the writing style of each author need to be selected. Those attributes generally depend on a set concerning word counts. These attributes are a single frequent word, a pair of words, and a trio of words [7]. The attributes have been very efficient for the majority of researchers in this field; however, there is still much space for accuracy development. Stylometric Authorship Balanced Attribution (SABA) is a machine learning method that uses a set of more effective attributes as compared to the frequent words algorithm [10]. The results of the SABA algorithm have given the most accurate authorship prediction thus far by analyzing the artistic writing style for authorship recognition and prediction. The effective attributes used in SABA include the single word attribute, pair of words attribute, and trio of words attribute, which is a text that contains three consecutive words.

Nevertheless, the proposed dataset in the present research will tackle the prediction efficiency of the SABA method by using a new set of attributes, namely sentence length, punctuation tools, and special characters. It is worth mentioning that even individual experts in the authorship field face a lot of confusion in selecting the suitable attribute set that is required for predicting the correct author. For implementation purposes, the authors are limited to the 19th century literature period.

***Corresponding author: Tareef Kamil Mustafa,** Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq, e-mail: tareef@scbaghdad.edu.iq

The proposed punctuation and special character attributes are obviously any symbol used to clarify the meaning of alphabetical and numerical symbols, such as comma, dot, asterisk, exclamation mark, question mark, etc. Meanwhile, the sentence length proposed here refers to the number of words counted in a single sentence without being interrupted by any punctuation or special symbol [2].

Stylometric supports the detection of plagiarism, which has become a common worldwide crime. In the case of a stolen article from a daily news journal, it is not a big issue; however, when it comes to stealing a creative narrative text, it is a major crime. The importance of Stylometric authorship attribution appears in identifying and detecting the authorship by using a well-defined method to predict the unknown author's text, i.e. by analyzing his writing style. The term "writing style" is a general expression and is used for getting highly reliable results. The style should be converted into a form of attributes. The issue that this method is trying to solve is recognizing an unknown text written by a nominated author in order to remove any ambiguity about that writer and give credibility to that author. To do so, the authorship attributes within a Stylometric method should be used to predict the ambiguous text and thus connect it to a certain author.

The SABA method is involved in text mining implementation, which is considered the most successful method in the Stylometric field thus far. This method proposed the trio attribute as the most efficient feature, supplanting the single frequent word. This method is an expansion of the Burrows-Delta method, which depends on selecting the effective attributes depending on the cumulative variance value for the maximum frequency for each attribute compared with its mean value, rather than using a simple frequency for the above-mentioned word attributes such as the single word frequency or the trio attribute frequency [10].

To improve the efficiency of the SABA method, a new non-word attribute set is proposed in this work, which includes sentence length, punctuation, and special characters, mixing them all together for a consistent and accurate decision making for the purpose of successful authorship prediction. The present research shows a better authorship detection using the suggested non-word attributes rather than the classical word frequency counts, and this was done by comparing the prediction results in the present research with the latest results of other researchers using the same dataset.

2 Literature Review

Several researchers have worked in the field of Stylometric using various types of attributes for detecting or recognizing the author of an unknown text in several methods; however, they all used word features, selecting single words, pair of words, or trio of words.

Klaussner et al. [7] used the Burrows-Delta method for identifying the author among several suspected candidates. The novelty of the work lied in selecting representativeness and distinctiveness features, assuming that the important features were in the use of some well-known linguistic terms that were utilized rarely by the tested author than most other authors, as well as the frequent linguistic terms.

Sapkota et al. [14] concentrated on improving Structural Correspondence Learning. They proposed a median-based classification instead of the standard binary classification. They suggested dropping down using the frequency of the attribute. For each attribute selected in the training dataset, the median was calculated and used to represent the weight of the selected attribute [14].

Evert et al. [3] proposed a method using Burrows-Delta, Euclidean distance, and the classical bag of words attribute to detect the anonymous author of the script by using precisely the single word attribute.

Mustafa et al. [12] used the SABA algorithm with word attributes. They overcame using the single word attribute alone and expanded their features into the pair and trio attributes to collect more evidence for recognizing the authors of several books. The dataset used in this research was in Arabic language to prove that SABA can tackle the authorship investigations in Arabic language.

Jamil and Mustafa [4] proposed a method improving on the Burrows-Delta method, the SARA (Stylometric Authorship Ranking Attribution) method, by neglecting the word attribute frequencies and favoring to use their rank sequence in the Stylometric selected attribute list, which led to more accurate prediction results.

3 Research Scope and Methodology

The research scope covers novels and theater materials (plays) of authors from the 19th century and are deemed to possess similar artistic criteria among each other, hence providing more challenge in predicting authorship. The non-word attributes used in this research include sentence length, punctuation symbols, and special characters for comparing the prediction results between different attribute sets. Implementation-wise, the language of choice is limited to English for a reasonable justified scientific comparison.

3.1 The Dataset

The dataset employed in this work is sourced from the Gutenberg Book Catalog [15]. The dataset source was introduced first by Zhao and Zobel in 2006 [15]. It was then utilized by Jamil and Mustafa in 2018 [4], Kim in 2015 [6], Pokou et al. in 2016 [13], and Mustafa et al. with their project team for Stylometric research in 2010 [12]. The dataset is a subset of the 19th century English literature. Three authors were selected from the top 100 most downloaded authors. Table 1 shows various works (plays and novels) from the top 3 selected authors.

Ten books were collected for each of the three authors, which resulted in a number of 30 books in total. In selecting these books, choices that give the impression of inconsistency were avoided within the aims and scope of the research, such as poetry, dictionaries, or text in languages other than English. Short stories were avoided as well, especially collections of short stories gathered in one book. Nine books for each author were categorized as the training books, and the 10th book was selected as the test book in a machine learning procedure (see the column "Book type" in Table 1).

Table 1: Plays and Novels from Three Selected Authors.

Author name	Book title	Book type	Downloaded file name
Oscar Wilde	*The Canterville Ghost*	Training	14522
	The Picture of Dorian Gray	Training	dgray10
	A House of Pomegranates	Training	hpomg10
	Lord Arthur Savile's Crime	Training	ldasc10
	The Ballad of Reading Goal	Training	rgaol10
	The Soul of Man	Training	slman10
	The Duchess of Padua	Training	dpdua10
	An Ideal Husband	Training	ihsbn10
	Lady Windermere's Fan	Training	lwfan10
	A Woman of No Importance	Test	awoni10
Jack London	*The People of the Abyss*	Training	1688
	The Road	Training	14658
	A Son of the Sun	Training	21971
	The Sea Wolf	Training	cowlf10
	The Valley of the Moon	Training	vlymn11
	The Mutiny of the Elsinore	Training	2415
	The Scarlet Plague	Training	21970
	The Call of the Wild	Training	callw10
	The Strength of the Strong	Training	sstrg10
	Smoke Bellew	Test	smkbl10a
William Shakespeare	*The Life of Henry the Fifth*	Training	ows2310
	The Merchant of Venice	Training	1ws1810
	As You Like It	Training	1w2510
	The Tragedy of Othello	Training	1ws3210
	The Life of Timon of Athens	Training	1ws3710
	A Midsummer Night's Dream	Training	1ws1710
	The Tragedy of Julius Caesar	Training	1ws2410
	The Tragedy of Antony and Cleopatra	Training	1ws3510
	Hamlet, Prince of Denmark	Training	2ws2610
	All's Well That Ends Well	Test	1ws3010

3.2 SABA Method Procedure Using Non-word Attributes

The SABA method is one of the latest, most successful methods in Stylometric authorship detection. It is an unguided, automated, expert-free, and language-independent method. That is why it was adopted in the present research [10].

Implementing SABA starts from uploading the chosen novels in text mode (with .txt extension). The chunking step is performed to select the attribute set type, whether it is a special character, punctuation symbol, or statement length in words (the number of words counted between any punctuation symbol and special character) with pre- and post-traditional cleansing [13]. Steps 1–5 in the proposed method are feature selection steps that are utilized for machine learning data training, whereas step 6 represents feature extraction for data testing.

Steps 7–9 are responsible for post-processing, to extract the prediction results, data mining visualization (histogram), and calculating the quantitative values.

The present algorithm steps (see Figure 1) are as follows:

1. Calculate the frequency of all selected attributes using Structured Query Language statements in each book from the training procedure. The frequency is extracted from the nine books altogether and not for each single book.
2. Replace the attribute frequencies with the percentage values extracted by dividing each attribute frequency over the total sum of the frequencies. Using a percentage for the frequency of attributes for each 10,000 is essential because of the need for a bigger coefficient that can be managed by the correlation coefficient [1].
3. Repeat steps 1 and 2 for all the training books of any author (Jack London as an example). Put all results together in a new database file containing all frequent attributes.
4. Calculate the average and the standard deviation to extract the values of coefficient of variance (CV) given in Eq. (1) for each attribute listed in the new database. This value will show how much fragmentation each attribute has in its frequency. In other words, it shows the level of consistency for an attribute – if the attribute always appears in high frequency, low frequency, or changes its position (hence, unstable).

$$CV = \left(\frac{Standard\ Deviation}{Mean} \times 100\right)\%, \tag{1}$$

where *Standard Deviation* = training books attribute standard deviation and *Mean* = training books attribute frequency mean.

5. Sort the attributes in descending order based on the CV value in order to select the attributes that represent the lowest non-zero CV values to graph the Stylometric map. This is achieved by excluding 5% of the highest CV attributes that give a shaky representation for the Stylometric author's maps. Note that this step may exclude some maximum item-sets because the maximum sets here are the sets that are ranked as the top-nominated high attributes due to their high frequency.
6. Repeat steps 1–3 above for each of the test books from all authors. Next, select from each of the selected books the same attributes that were chosen for the Stylometric map in step 5.
7. Draw the Stylometric histogram to illustrate the classical visualization data mining case showing the frequency contrast for each attribute before going further in extracting the quantitative final results.
8. Extract the Pearson correlation coefficient r for the particular author's Stylometric map from each test book, hence giving the Pearson value for each test as in Eq. (2):

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\left(\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}, \tag{2}$$

where r = Pearson correlation coefficient, X = attribute frequency for the test book, and Y = attribute ratio for the training book.

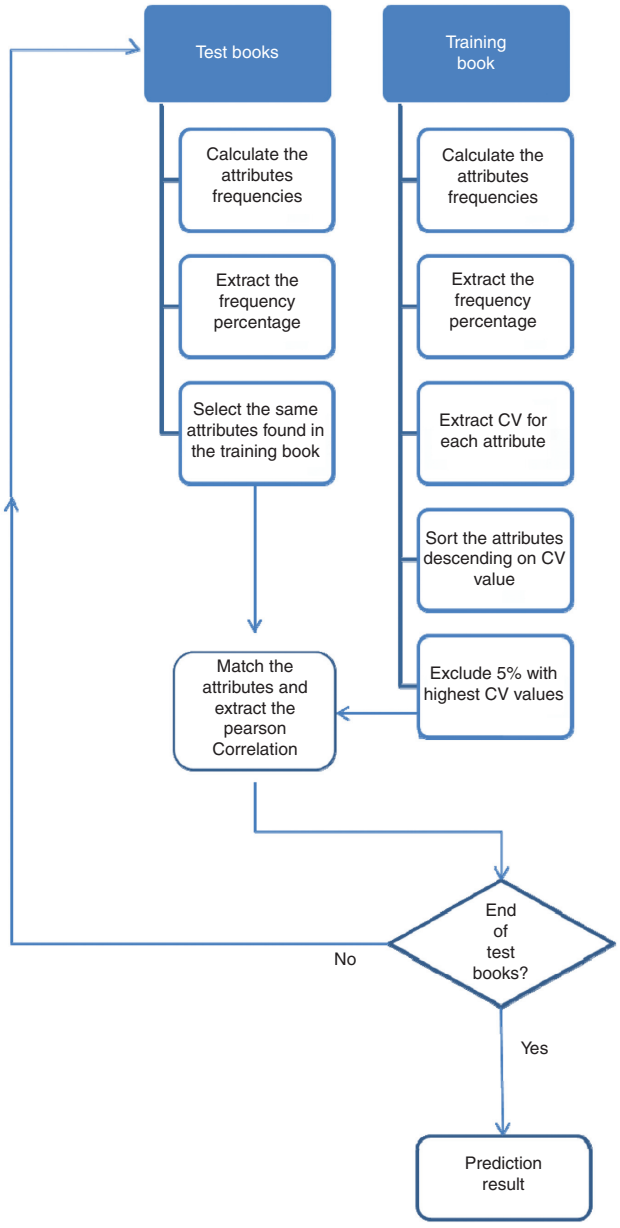

Figure 1: Diagram Summarizing the SABA Algorithm Steps.

9. Prediction is achieved by comparing the Pearson correlation coefficient r for any specific book for any author in the testing set with the Pearson correlation value in the training set for each author's map. The highest Pearson coefficient value of the test book will represent the prediction. Taking Mark Twain as an example, if the Pearson coefficient value of Twain's test value was the highest compared with the other Pearson coefficient in the authors training map, then the prediction is correct and SABA favors the correct author; otherwise, it is a false prediction.

4 Results and Discussion

Using the proposed non-word attribute set that included sentence length, special characters, and punctuation symbols with the SABA algorithm led to more accurate results than using the frequent words attributes. The

bold underlined boxes refer to the goal of the method in which the correlation results of each author are compared with the test book for the same author. It gives the highest value as shown in Table 2, which means that all the prediction tests of all authors of the research scope are successfully detected compared with the previous tests that gave a few false author detection, as shown in Table 3 (see the crosstab cell value between London's training map and Pearson in London test).

In a previous work, Mustafa [11] employed the same dataset; however, the prediction of the Jack London book gave the lowest authorship prediction results when it was implemented using several Stylometric methods, as shown in Table 3. The result of the proposed method showed that it gives better results, as this method positively detected all the cases correctly.

The histogram chart depicted in Figure 2 gives a visual illustration for comparing Shakespeare's map with the three authors' test books. It has been proven quantitatively and with a clear result that the line of Shakespeare's test book is the nearest one to Shakespeare's map line, as shown in Figure 2. The *X*-axis of this histogram is the attribute, while the *Y*-axis represents the frequency for each corresponding attribute [5]. The attributes for Shakespeare starts from the lowest CV values as mentioned previously, which are the comma, dot (full stop), semicolon, 13, 9, 1, 6, 7, . . ., etc. With respect to the number attributes, for example, 13 means that the sentence with a 13-word length is the most frequent with low CV value and represents one of the most dependable attributes for Shakespeare's map, and so on for the rest of the numeric attributes shown in the figure.

Table 2: Prediction Results Showing Authors' Training Maps vs. Authors' Test Books for the Present Research.

Pearson Correlation values	Pearson in Wilde test	Pearson in Shakespeare test	Pearson in London test
Wilde training map	**0.8013**	0.7310	0.7322
Shakespeare training map	0.6956	**0.7447**	0.5570
London training map	0.3707	0.3926	**0.8748**

Table 3: Prediction Results Showing Authors' Training Map vs. Authors' Test Books for the Previous Test [11].

Pearson Correlation values	Pearson in Wilde test	Pearson in Shakespeare test	Pearson in London test
Wilde training map	**0.952347**	0.907094	0.848152
Shakespeare training map	0.848658	**0.975139**	0.873507
London training map	0.984803	**0.827159**	~~0.765925~~

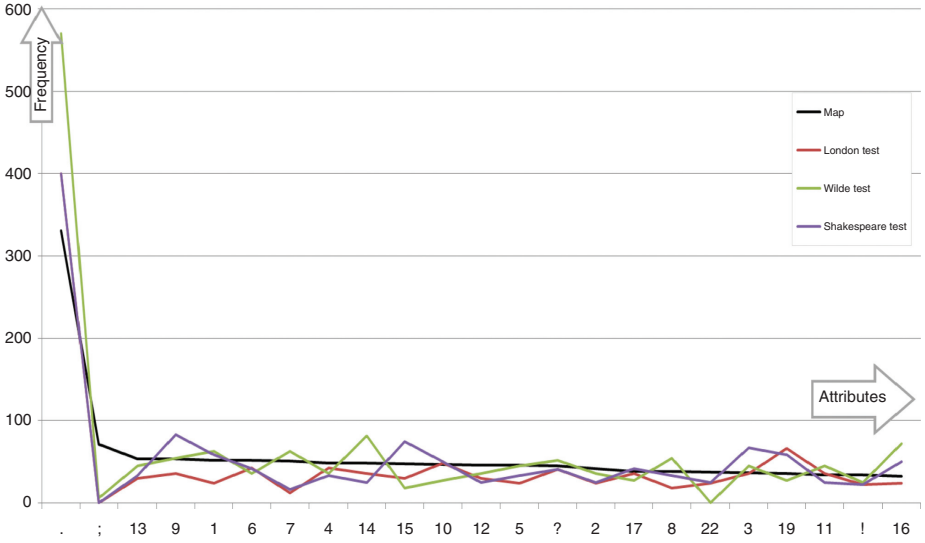

Figure 2: Shakespeare's Map vs. Authors' Test Books.

Figure 3 shows the comparison between the Wilde map and the three authors' test books, which provides an illustrative visual decision that Wilde's curve test book is the nearest curve to the main curve of Wilde's map, which has been proved quantitatively in Figure 3.

Figure 4 depicts the comparison between London's map and the three authors' test books, giving a very clear result that London's curve test book is the nearest curve to the London map curve, as shown previously in Table 2 and proven quantitatively.

Returning to previous results by Mustafa [11], it is clear that all the results achieved by the present research are more accurate, especially for the Jack London case. London's case always failed in the prediction tests using the classic word attributes. See the value 0.8748 in the crosstab cell value between London's training map row and Pearson in London test column in Table 2, which presents the highest value among the three cells in the row presenting the correct author detection.

Notice the first and second rows in Table 2, and compare them with the facing row in Table 3. Both show correct predictions because each author map faces the same author test that holds the highest value; the diagonal line in both tables should hold the highest value in each row, unlike the third row in Table 2. Notice that the London test failed to give a correct prediction (the cross-lined box value in the crosstab cell value between London's training map row and Pearson in Shakespeare test column showed the highest value in

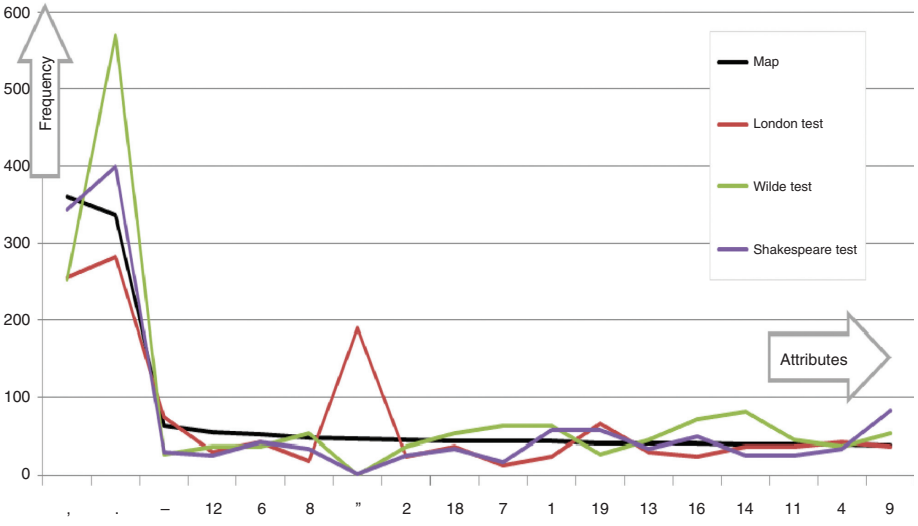

Figure 3: Wilde's Map vs. Authors' Test Books.

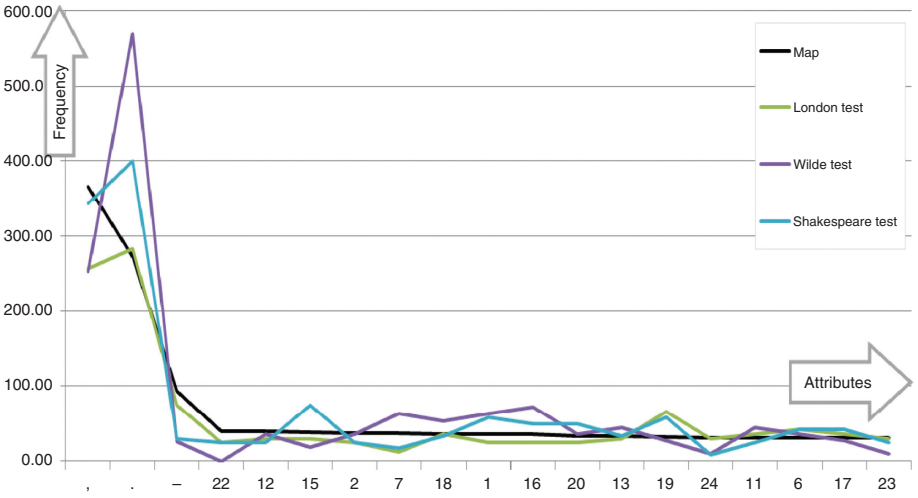

Figure 4: London's Map vs. Authors' Test Books.

the row missing the correct author prediction), unlike the facing results in Table 3 that shows a successful prediction.

5 Conclusions

In this research, a new set of attributes was suggested to replace the SABA method for Stylometric authorship prediction, and was compared with the prediction results implemented by other researches using the traditional frequent words attributes. The following conclusions were made:

- Using the proposed non-word attribute set that includes sentence length, special characters, and punctuation symbols with the SABA algorithm led to more accurate results than using the frequent words attributes, as shown in the implementation done on a standard dataset.
- Decreasing the number of attributes selected by trimming 5% of the attributes depending on the highest CV led to a better attribute selection and process abridgment, and hence more accurate authorship prediction results.
- Depending on the formal attributes of single word, pair of words, and trio of words gave more prediction errors for some authors such as Jack London, while the proposed non-word attribute set enhanced the prediction accuracy for all authors with no prediction errors using the same dataset as previous researches.

Bibliography

[1] A. A. Abdul-Razzaq and T. K. Mustafa, Burrows-Delta method fitness for Arabic text authorship Stylometric detection, *IJCSMC* **3** (2014), 69–78.
[2] A. Brand, L. Allen, M. Altman, M. Hlava and J. Scott, Beyond authorship: attribution, contribution, collaboration, and credit, *Learned Publishing Journal* **28** (2015), 151–155.
[3] S. Evert, T. Proisl, F. Jannidis, I. Reger, S. P. Christof and T. Vitt, Understanding and explaining Delta measures for authorship attribution, *Digit. Scholarsh. Humanit.* **32** (2017), 4–16.
[4] M. T. Jamil and T. K. Mustafa, Ranking attribution: a novel method for Stylometric authorship identification, *IJACST* **9** (2018), 54–61.
[5] P. Juola, Authorship attribution, *Found. Trends. Inf. Ret.* **1** (2008), 233–334.
[6] K. J. Kim, *Information Science and Applications*, Kyonggi University Publications, Suwon, South Korea, 2015.
[7] C. Klaussner, J. Nerbonne and C. Çöltekin, Finding characteristic features in Stylometric analysis, *Digit. Scholarsh. Humanit* **30** (2015), 114–129.
[8] L. Lakshmi and K. P. Pushpendra, A study on author dentification through Stylometry, *IJCSN* **2** (2013), 653–657.
[9] V. Landeiro, D. Reis and A. Culotta, Robust text classification in the presence of confounding bias, in: *Proceedings of the 13 AAAI Conference on Artificial Intelligence*, pp. 186–193, 2016.
[10] T. K. Mustafa, Stylometric authorship balanced attribution prediction method, Thesis, University Putra Malaysia, Malaysia, 2011.
[11] T. K. Mustafa, Text mining authorship detection methods development, in: *Proceedings of the 288th Academicsera International Conference*, pp. 3–8, 2018.
[12] T. K. Mustafa, A. A. Abdul Razzaq and E. A. Al-Zubaidi, Authorship Arabic text detection according to style of writing by using (SABA) method, *AJAST* **5** (2017), 483–490.
[13] Y. J. M. Pokou, P. Fournier-Viger and C. Moghrabi, Authorship attribution using variable length part-of-speech patterns, in:*Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, pp. 354–361, 2016.
[14] U. Sapkota, T. Solorio, M. Montes and S. Bethard, Domain adaptation for authorship attribution: improved structural correspondence learning, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1, pp. 2226–2235, 2016.
[15] Y. Zhao and J. Zobel, Searching with style: authorship attribution in classic literature, in: *Proceedings of the 30th Australasian Computer Science Conference*, pp. 59–68, 2006.