Authorship Arabic Text Detection According to Style of Writing by using (SABA) Method

Tareef Kamil Mustafa[1], Ammar Adil Abdul Razzaq[1], Ehsan Ali Al-Zubaidi[1,*]

[1]College of Science
University of Baghdad, Iraq

[2]Faculty of Physical Planning
University of Kufa, Iraq

[*]*Corresponding author's email: Ihsana.kareem [AT] uokufa.edu.iq*

ABSTRACT--- ***Authorship attribution of a style of writing is a method depend on analyzing texts in text mining, e.g., historical books and novels that famous authors wrote, attempted to measure the author's style, by choosing some attributes that show the author manner of writing. Assuming that these writers have a different way of writing that no other writer have; thus, authorship attribution is the essential of identifying the author of a given text [1].***

Keywords--- Authorship Attribution; Style of writing; text mining

1. INTRODUCTION

In computer science, there is a field called "Text mining" that was taken from Data mining. To be more specific, in the authorship investigation using the style of the author, we use a sub- field of text mining called "Authorship attribution" and "Stylometric Text mining". All these subjects need to be defined to get the picture well clarified.

1.1 Arabic Text Mining

Text Mining is found newly. Formerly it is an unknown information automatically extracting from different written resources. A key element relates the extracted information together to form new facts or new hypotheses to be discovered further by more traditional means of experimentation. In searching, the user is exemplary looking for something that is already recognized and has been written by another person. The problem is pushing aside all the materials that nowadays isn't relevant to your needs to find the related information. At variance what's in text mining, the goal is to invent anonymous information, something that no one yet knows and so could not write so far[2].

Arabic is considered one of the very spoken languages in the world. In fact it a basic language in the Arab nations as well as a secondary language in many other nations. The language alphabet consists of 28 letters plus special character and punctuation symbols, which is. Moreover, the writing direction in Arabic is from right to left[3]. The manner of writing letters in a word changes depending on the location of the letter within the word. So, if the letters come at first, middle or at the end of the word, the letter forms changes. Lastly, there are diacritics in Arabic that are symbols placed above or below the letters to reduplication the letter in the pronunciation or to give short vowels. The most researchers in Arabic text applied learning algorithms only designed for English text without making salient changes[4].

One main problem associated with Arabic text classification the lack of standardized published Arabic and also infrequent. Such works can be used as key data sets for researchers in related fields to compare the results. Actually, most of the related research articles obtain data from online newspapers and websites. Such works ordinarily do not publish their data for other researchers to utilize. therefore the trust in the results derived from such experimental studies is not high enough [5]

However, researchers concluded that Arabic text classification is a very challenging task due to language complexity.

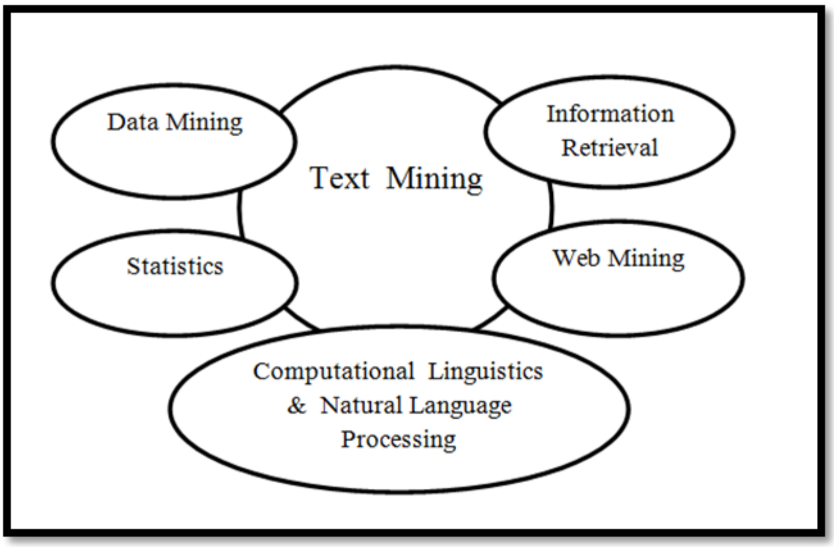


Figure 1: Relations of Text mining with other fields

1.2 Authorship attribution (AA)

: is the process of trying to identify the likely authorship of a given document, given a collection of documents whose authorship is known. Most of the approaches described in the research literature consist of two components, a comparison mechanism, and an indexing mechanism. The indexer converts each document to a set of tokens whose properties are assumed to be characteristic in some way of a certain author. The comparator uses these markers to assign an author to un-attributed documents [6].

2. PROPOSED WORK

In this paper we will measure the accuracy of Stylometric features, so it can be reduced nearly as well as fingerprints of different persons using authorship attributes.

The main aim is test algorithms supports a system of decision making enables users to predict in Arabic text and choose the right author for a specific unknown author's novel under consideration, by using a learning procedure to train the system the Stylometric map of the author and behave as an expert opinion[7]. Test the optimal threshold for authors of Arabic. Compare the effectiveness of many attributes in Arabic as the frequent, pair, trio sentence.

Still the word of frequent is a head of other attributes that give good results in the researches and experiments and still the best parameter and technique that's been used until now is the counting of the bag-of-word with the maximum item set[8] .

Here we will focus on literature written in Arabic language and work on analysis of Arabic text based on the words redundancy as a feature in Arabic books as a frequent, pair and trio-of-words and test results obtained using text mining by computer-assisted authorship attribution is to define a certain characterization of documents that captures the writing manner of authors [9].

We proposed here if Stylometric Authorship Balanced Attribution (SABA) works with Arabic language as well as the attributes "frequent, pair, trio" considered as a constant in Arabic language, if we assumed that we need to threshold of "300" of frequent situations, and we check also the punctuation and symbols in Arabic if it were a good attributes type.

In addition to the difficulty of Arabic language, it is hard to get resources of Arabic books in text formatted.

3. METHODOLOGY

The proposed methodology is to test the algorithms (SABA)that was used in many of the research in different languages using statistical analysis and text mining, but here we will test this algorithm in the Arabic language, which have not been tested previously.

3.1 Stylometric authorship attribution methodology

The methodology of stylometric authorship attribution main steps can be shown in figure (2) which describes the process of converting readable text into the semi-structured data.

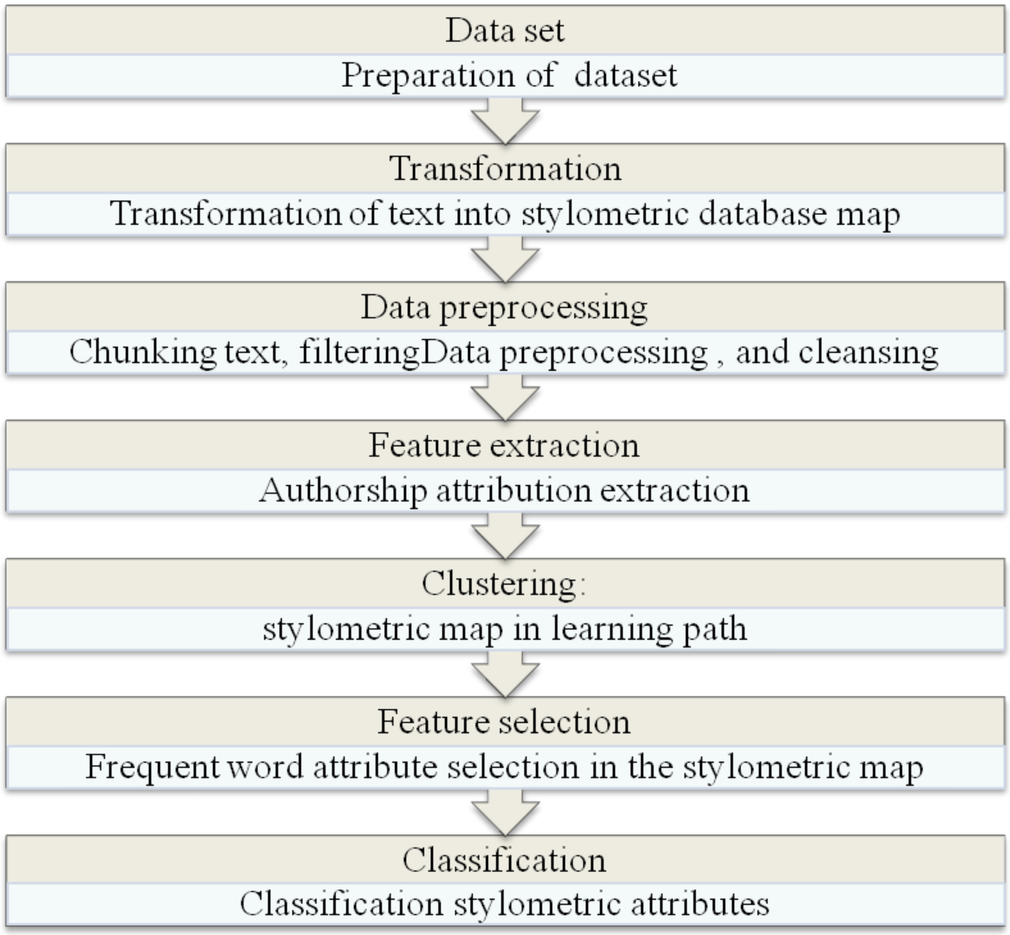

Data set
Preparation of dataset
Transformation
Transformation of text into stylometric database map
Data preprocessing
Chunking text, filteringData preprocessing , and cleansing
Feature extraction
Authorship attribution extraction
Clustering:
stylometric map in learning path
Feature selection
Frequent word attribute selection in the stylometric map
Classification
Classification stylometric attributes

Figure (2): main steps for stylometric authorship attribution (SAA)

3.2 Dataset

This include dataset , Details of the dataset , and dataset plan.

Table (1): Dataset

No.	Name of Author	Number of Books
1	Ibnjuzia - ابن الجوزية	6 (5 for Learn, 1 for test)
2	Sakhawy – السخاوي	6 (5 for Learn, 1 for test)
3	Tusi – الطوسي	6 (5 for Learn, 1 for test)

Table (2): Details of the dataset

No.	Authors' Names	Books Titles	Books Types
1.	Ibnjuzia -ابن الجوزية	طريق الهجرتين وباب السعادتين	Training
		حادي الأرواح إلى بلاد الأفراح	Training
		الصواعق المرسلة في الرد على الجهمية والمعطلة	Training
		شفاء العليل في مسائل القضاء والقدر والحكمة والتعليل	Training
		إغاثة اللهفان من مصايد الشيطان	Training
		أحكام أهل الذمة	Test
2.	Sakhawy – السخاوي	التحفة اللطيفة في تاريخ المدينة الشريفة	Training
		القول البديع في الصلاة على الحبيب الشفيع	Training
		السر المكتوم في الفرق بين المالين المحمود والمذموم	Training
		الغاية فى شرح الهداية فى علم الرواية	Training
		البلدانيات للسخاوي	Training
		فتح المغيث بشرح ألفية الحديث	Test
3.	Tusi – الطوسي	فضائح الباطنية	Training
		المنخول	Training
		الوسيط في المذهب	Training
		العلم في فن المنطق	Training
		تهافت الفلاسفة	Training
		المستصفى	Test

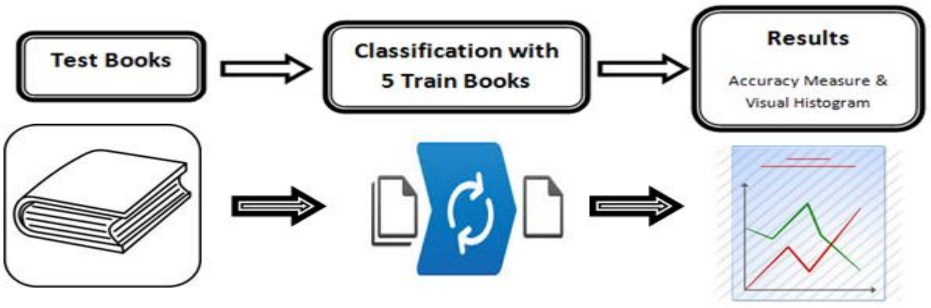
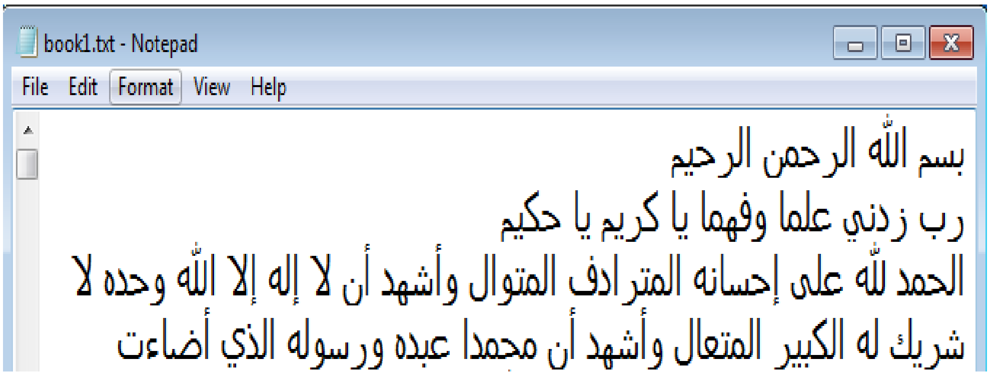


Figure (3): dataset plan

3.3 Transforming text into stylometric database map

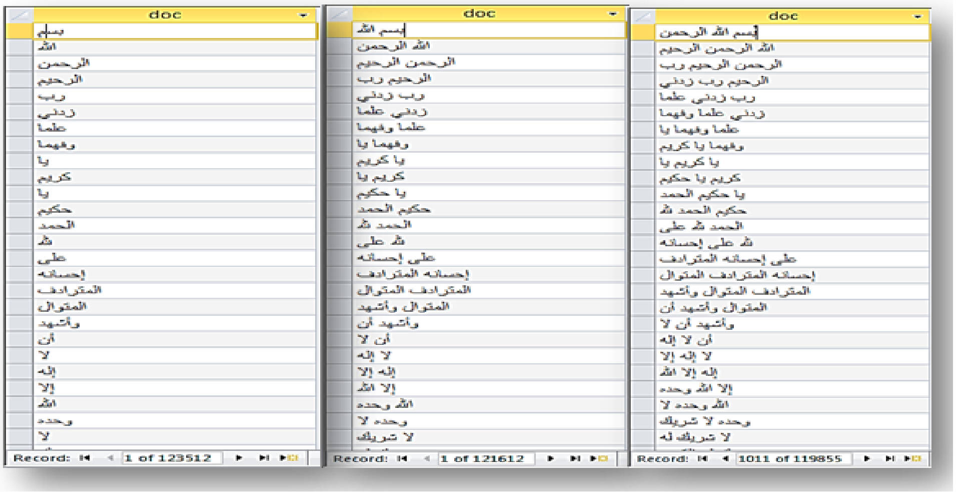
The text transformation process contains chunking, filtering and cleansing text before finally transforming the text into database tuples, in this experiment, changing the data in separated tables which is represented by 3 tables for 5 books for each author respectively, repeating this transformed data process with single word, pair and trio words in the designed database, figure (4) represent a sample of the text and table (3) is representing the transformed text into separate database.



بسم الله الرحمن الرحيم
رب زدني علما وفهما يا كريم يا حكيم
الحمد لله على إحسانه المترادف المتوال وأشهد أن لا إله إلا الله وحده لا
شريك له الكبير المتعال وأشهد أن محمدا عبده ورسوله الذي أضاءت

Figure (4): Sample of the textbook

Table (3): single word and pair and trio words



doc	doc	doc
بسم	بسم الله	بسم الله الرحمن
الله	الله الرحمن	الله الرحمن الرحيم
الرحمن	الرحمن الرحيم	الرحمن الرحيم رب
الرحيم	الرحيم رب	الرحيم رب زدني
رب	رب زدني	رب زدني علما
زدني	زدني علما	زدني علما وفهما
علما	علما وفهما	علما وفهما يا
وفهما	وفهما يا	وفهما يا كريم
يا	يا كريم	يا كريم يا
كريم	كريم يا	كريم يا حكيم
يا	يا حكيم	يا حكيم الحمد
حكيم	حكيم الحمد	حكيم الحمد لله
الحمد	الحمد لله	الحمد لله على
لله	لله على	لله على إحسانه
على	على إحسانه	على إحسانه المترادف
إحسانه	إحسانه المترادف	إحسانه المترادف المتوال
المترادف	المترادف المتوال	المترادف المتوال وأشهد
المتوال	المتوال وأشهد	المتوال وأشهد أن
وأشهد	وأشهد أن	وأشهد أن لا
أن	أن لا	أن لا إله
لا	لا إله	لا إله إلا
إله	إله إلا	إله إلا الله
إلا	إلا الله	إلا الله وحده
الله	الله وحده	الله وحده لا
وحده	وحده لا	وحده لا شريك
لا	لا شريك	لا شريك له
Record: 1 of 123512	Record: 1 of 121612	Record: 1011 of 119855

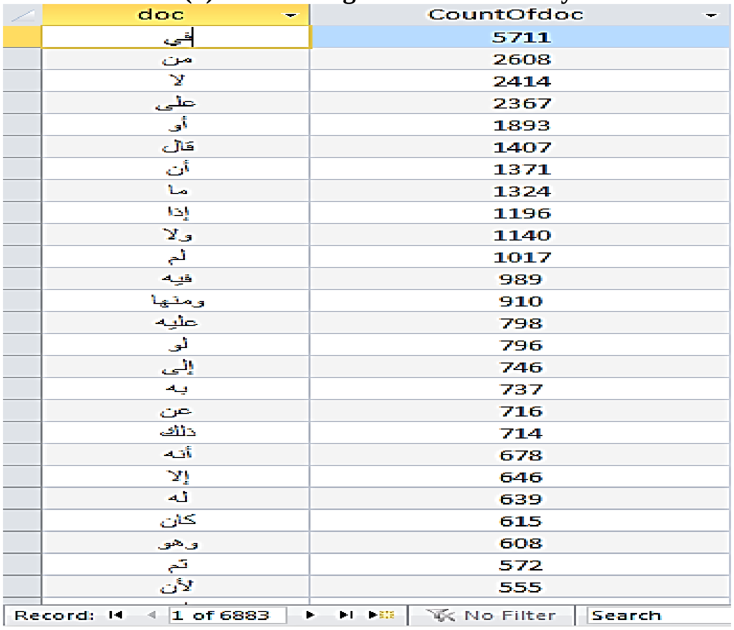
3.4 Data pre-processing

The most common procedures for preparing the data are cleansing and filtering, thus the data can be clearly analyzed without any distortion or noise. Cleaning and filtering operations includes all multi spaces found between the words (the fact that sentences consisting of several words that sometimes are separated by several spaces), multi punctuations, similar signs and titles of sections. Lastly, cleansing operations includes also the removal of the diacritics in Arabic that are symbols placed above or below, duplication symbol letters "الشدّة" to double the letter in the pronunciation or to give short vowels.

3.5 Stylometric authorship attribution features extraction

After a pre-process operation, which includes cleansing and filtering on data, the chunking operation starts on data that depends on the attribute type, whether it is single word, word pair or trio statement. The feature extraction step will be performed by collecting the redundancies of the features in the learning path and store the frequencies in the stylometric database map (SAA map), the SAA Map sorted in separated table, as shown in table (4).

Table (4): descending attributes for a stylometric database map



doc	CountOfdoc
في	5711
من	2608
لا	2414
على	2367
أو	1893
قال	1407
أن	1371
ما	1324
إذا	1196
ولا	1140
لم	1017
فيه	989
ومنها	910
عليه	798
لو	796
إلى	746
به	737
عن	716
ذلك	714
أنه	678
إلا	646
له	639
كان	615
وهو	608
ثم	572
لأن	555

3.6 Clustering Stylometric map

After the completion of the cleansing. Chunking and analyzing in both learning and testing data, the data could be classified to represent the authorship attribution.

Clustering the SAA map is the most important step to differentiate the semantic authorship from the stylometric authorship because the semantics must be supported with some language rules and supported by a database annotated with frequent words or collocations. Whereas the stylometric method follows no prior rules, and it is language independent, so it will not have regards nor intentions to fit the grammatical rules of the language used under consideration.

The clustering in this the methodology is to measure the detection ability of the algorithm. The learning process during clustering uses 5 out of 6 books for each author, leaving one book for testing with the remaining books from 5 other authors.

After the completion of the clustering data for each set of training and testing books, the 300 attributes are selected as a result to compare the stylometric author attribution maps with testing books for each author, the 300 attributes in each author stylometric map usually chosen from thousands of high frequencies attributes through sorting, as a result, the stylometric map for each of the 3 suspected authors against 5 books under investigation will produce one author prediction for each investigated book.

3.7 Attribute selection

Attribute selection is based primarily on the number of frequencies that results from clustering the attributes, These features result from a group of five books for the purpose of making the stylometric map for each author, The sixth book is used for testing purposes.

Due to the different size of the books in this experiment, the percentage measure represents the number of frequencies for each attribute divided by the sum of frequencies for all attributes, and there is a weighted frequency for all used books was obtained.

By comparing the results between the learning data for a specific author and the testing data for all three authors, a comparison measure is needed to compare the accuracy. The statistical measure of Pearson correlation (r) will give weight to each attribute within the range [-1, 1].

The negative results are not common since there is always some relation between the author's style that results from the grammar rules. The comparison for choosing the favorite author will be resulted by the highest positive score that is achieved in the proposed algorithm.

The features were selected to give the nearest estimation between the Stylometric map of the author and other test maps.

3.8 Stylometric attributes classification

After extracting the values of Pearson correlation for all authors, the values are grouped according to the equation called Winnow algorithm, as shown in equation (1).

$y = w1x1 + w2x2 + w3x3 + \ldots + wnxn$ …….(1)

Where x here is the classifier that used to assign parameter, hence wrong authors have the value of negative and the correct author have the value positive, n is a number of test books or a number of predictions in the test = 5, y is the accuracy measure that describes the fitness between the authors map, and the test book under investigation. Lastly, we can be represented by using the computational stylometric measure of Pearson correlation, which is used to find the weight for each classifier in order to produce the final automated result.

Finally, to get a ratio result, the accuracy measure is separated between real positive (correct prediction) and negative (wrong prediction), the amount of negative books' weight should equal the weights of the positive authors.

4. RESULT

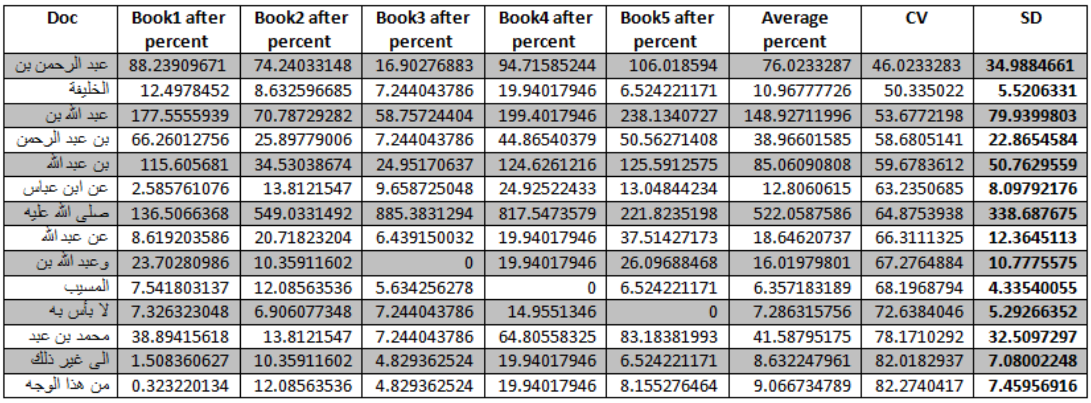
4.1 Testing SABA method

(SABA) method is considered an expansion of Burrow-Delta method, SABA method depends on the coefficient of variance (CV), represented as a measurement of statistical that is not affected by the observation of mean, SABA method formerly tested in English language with high prediction, this research will examine and test this algorithm in Arabic language in the single, pair and trio words.

In SABA method, the test of single, pair and trio words is comparable to the Burrow Delta method in application, but there is an essential difference between them, precisely when selecting the top of 300 attributes, these selections depends on the values of (C.V). The following example in trio words can explain the major steps of extracting the (C.V) and the method of selecting the required attributes.

To apply SABA method, all steps used in the Burrow Delta method are recurred then transform the final stylometric database map to Microsoft Excel, by using the ready functions to elicit the values of the average, the standard deviation (σ) and the (C.V) for each attributes in the data learning step, the (C.V) can be found by dividing the (σ) by the (μ) itself, Finally, we sort the data in ascending order based on the values of the (C.V) and select the top 300 attributes, as shown in table (5).

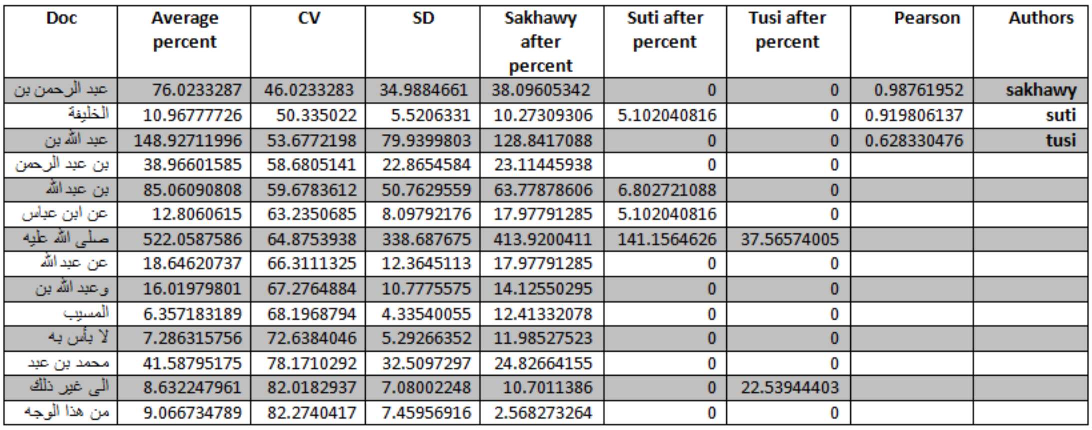
Table (5): SABA stylometric map



Doc	Book1 after percent	Book2 after percent	Book3 after percent	Book4 after percent	Book5 after percent	Average percent	CV	SD
عبد الرحمن بن	88.23909671	74.24033148	16.90276883	94.71585244	106.018594	76.0233287	46.0233283	**34.9884661**
الخليفة	12.4978452	8.632596685	7.244043786	19.94017946	6.524221171	10.96777726	50.335022	**5.5206331**
عبد الله بن	177.5555939	70.78729282	58.75724404	199.4017946	238.1340727	148.92711996	53.6772198	**79.9399803**
بن عبد الرحمن	66.26012756	25.89779006	7.244043786	44.86540379	50.56271408	38.96601585	58.6805141	**22.8654584**
بن عبد الله	115.605681	34.53038674	24.95170637	124.6261216	125.5912575	85.06090808	59.6783612	**50.7629559**
عن ابن عباس	2.585761076	13.8121547	9.658725048	24.92522433	13.04844234	12.8060615	63.2350685	**8.09792176**
صلى الله عليه	136.5066368	549.0331492	885.3831294	817.5473579	221.8235198	522.0587586	64.8753938	**338.687675**
عن عبد الله	8.619203586	20.71823204	6.439150032	19.94017946	37.51427173	18.64620737	66.3111325	**12.3645113**
وعبد الله بن	23.70280986	10.35911602	0	19.94017946	26.09688468	16.01979801	67.2764884	**10.7775575**
المسيب	7.541803137	12.08563536	5.634256278	0	6.524221171	6.357183189	68.1968794	**4.33540055**
لا بأس به	7.326323048	6.906077348	7.244043786	14.9551346	0	7.286315756	72.6384046	**5.29266352**
محمد بن عبد	38.89415618	13.8121547	7.244043786	64.80558325	83.18381993	41.58795175	78.1710292	**32.5097297**
الى غير ذلك	1.508360627	10.35911602	4.829362524	19.94017946	6.524221171	8.632247961	82.0182937	**7.08002248**
من هذا الوجه	0.323220134	12.08563536	4.829362524	19.94017946	8.155276464	9.066734789	82.2740417	**7.45956916**

After building relationships between the final stylometric map and the five test books for all authors, we obtain on the final trio test in SABA method, as shown in table (6).

Table (6): Final trio test in SABA method



Doc	Average percent	CV	SD	Sakhawy after percent	Suti after percent	Tusi after percent	Pearson	Authors
عبد الرحمن بن	76.0233287	46.0233283	34.9884661	38.09605342	0	0	0.98761952	sakhawy
الخليفة	10.96777726	50.335022	5.5206331	10.27309306	5.102040816	0	0.919806137	suti
عبد الله بن	148.92711996	53.6772198	79.9399803	128.8417088	0	0	0.628330476	tusi
بن عبد الرحمن	38.96601585	58.6805141	22.8654584	23.11445938	0	0		
بن عبد الله	85.06090808	59.6783612	50.7629559	63.77878606	6.802721088	0		
عن ابن عباس	12.8060615	63.2350685	8.09792176	17.97791285	5.102040816	0		
صلى الله عليه	522.0587586	64.8753938	338.687675	413.9200411	141.1564626	37.56574005		
عن عبد الله	18.64620737	66.3111325	12.3645113	17.97791285	0	0		
وعبد الله بن	16.01979801	67.2764884	10.7775575	14.12550295	0	0		
المسيب	6.357183189	68.1968794	4.33540055	12.41332078	0	0		
لا بأس به	7.286315756	72.6384046	5.29266352	11.98527523	0	0		
محمد بن عبد	41.58795175	78.1710292	32.5097297	24.82664155	0	0		
الى غير ذلك	8.632247961	82.0182937	7.08002248	10.7011386	0	22.53944403		
من هذا الوجه	9.066734789	82.2740417	7.45956916	2.568273264	0	0		

By getting on the weights for each parameter, multiply each Pearson value by -1 if it is the wrong author for the formerly known result or by +1 if it is the right author, as shown in table (7).

Table (7): Final SABA results

Author name	Pearson (Sakhawy)	weight	sign	Result
Sakhawy	0.976621	4	1	3.90648
Suti	0.896984	1	-1	-0.897
Tusi	0.626008	1	-1	-0.626
SABA RESULT				**2.38348**

5. CONCLUSION

Stylometric authorship balanced attribution (SABA) that is able to forecast with higher accuracy and independent from human judgments, which means that the approach does not rely on the domain experts. This method implemented by merging three methods, which are called the computational method ,the Burrows-delta method, and the algorithm of Winnow. Stylometric authorship balanced attribution (SABA) method also uses a set of more effective attributes in comparison with frequent words method. This leads in higher Stylometric prediction thus far, having more accurate for author artistic writing style for authorship recognition and prediction. The effective attributes are represented by the frequent word, pair and the trio, while both are multiple words attributes.

The SABA method is compared against three other approaches using the computational method, the Burrows-delta method ,and the Winnow algorithm method. The results showed that the SABA method produces superior prediction accuracy and even provides a completely correct result during the final phase of the experiment.

6. REFERENCES

[1] Argamon, S. and S. Levitan, 2005. Measuring the usefulness of function words for authorship attribution. Proceeding of the Joint Conference on Association for Literary and Linguistic Computing Computer Humanities, June 16-19, University of Victoria, Canada, pp: 2-4. http://lingcog.iit.edu/doc/paper_162_argamon.pdf.

[2] Mofleh Al-diabat (2012), " Arabic Text Categorization Using Classification text Mining", Department of Computer science /Al Albayt University. Applied Mathematical Sciences, Vol. 6, 2012, no. 81, 4034 – 4047.

[3] Zaho Y., Zobel J. (2008), "Search with manner: Authorship Attribution in Classic Literature", the Thirtieth Australasian Computer Science Conference, Ballarat, Jan. 2007.

[4] J. F. Burrows, "Delta: a gauge of stylistic difference and an index to likely authorship," Literary and Linguistic Computing 17, pp. 268–288, 2002.

[5] D. Hoover, "Delta prime" Literary and Linguistic Computing 19.4, pp. 478–496, 2004.

[6] D. Hoover, "Testing burrows' delta," Literary and Linguistic Computing 19.4, pp. 454–476, 2004.

[7] J. Burrows, "The English of Juvenal: Computational stylistics and translated texts," Style 36, pp. 678–700, 2002.

[8] Sterling Stein and Shlomo Argamon (2007), "A Mathematical Explanation of Burrows's Delta", Linguistic Cognition Laboratory Department of Computer Science Illinois Institute of Technology Chicago.

[9] S. K.-M. J. Higgins, Concepts in Probability and Stochastic Modelling. Duxbury Press,1 ed., 1995.