Ranking Attribution: A Novel Method for Stylometric Authorship Identification

Marwa Taha Jamil, Dr. Tareef kamil Mustafa
College of Science, University of Baghdad
Baghdad, Iraq

***Abstract*—Stylometric Authorship attribution is one of the essential approaches in the text mining. The present research endorses a Stylometric method called Stylometric Authorship Ranking Attribution (SARA) overcomes the usual problems which are processing time and accurate prediction results, without any human opinion that relays on the domain expert. This new method also uses the most effective attributes used in the Stylometric authorship prediction frequent word bag counts, whether it was frequent single, pair or trio words attributes, which are the most successful attributes in Stylometric prediction, having more alibi for author artistic writing style for our authorship recognition and prediction proposed technique. The experiments show that the proposed method produces superior prediction accuracy and even provides a completely correct result at the final stage of our experimental tests regarding the dataset scope.**

Keywords—Data mining; text mining; Stylometric Authorship Attribution; SARA

I. Introduction

Data mining is the evaluation of observational data units to find authorized relationships and the evaluation of statistics in novel methods that are each obvious and beneficial to the statistics owner [1]. Text mining (TM) [2], additionally recognized as understanding discovery in textual database(KDT) [3] or textual content data mining [2], of which new fascinating expertise is created, many defined it also as the process of extracting previously unknown, understandable, achievable and practical patterns or understanding from the series of large and unstructured textual content information or corpus. Text mining uses the same evaluation approach and techniques as statistics mining. However, information mining requires structured data, whilst textual content mining aims to discover patterns in unstructured statistics [4]. The problem of text mining has gained growing attention in current years because of the big quantities of textual content data, which created a variety of social network, web, and other information-centric applications. Unstructured statistics is the most natural form of information which can be produced in any application scenario. As a result, there has been an extraordinary need for graph techniques and algorithms which can successfully manner a broad range of textual content purposes [1]. Another foremost issue is a multilingual text refinement dependency that creates problems. Only a few tools are available that aid multiple languages [5]. Text mining is generally composed of three steps: text preprocessing, text mining operations, post-processing. Text preprocessing tasks inclusive of information selection, classification and characteristic extraction normally convert the documents into intermediate forms, which have to be appropriate for distinct mining purpose. Text mining operations are the central phase of a text mining system and encompass clustering, association rule discovery, trend analysis, sample discovery and different know-how discovery algorithms. Post-processing tasks manipulate facts or understanding coming from text mining operations, such as comparison and resolution of knowledge, interpretation, and data visualization representation [6]. The upcoming sections in this research will illustrate the latest methods and approaches of a certain subfield in the text mining area that is concerned about the text corpus in literature and the writing style of its authors before stepping into the proposed method details.

II. Literature Review

An essential trouble in authorship attribution is the choice of stylometric aspects that are linguistic expressions of unique authors. Sets of proposed facets may vary, depending on accessible data, the supposed generality of their extraction approach and applicability to precise languages.

The easiest elements describe statistical residences of documents: word length, sentence length, and vocabulary richness. Function phrases are points primarily based on word frequencies. In contrast to text categorization problems, where the most established words are considered useless or even unsafe for classification, in authorship attribution problems they are frequently used as non-public fashion markers. However, not all the most universal phrases are exact candidates to be blanketed to that set of features: an important characteristic is an instability [7], i.e. the possibility to be replaced with the aid of every other word from the dictionary.

Other word-based elements are phrase sequences (n-grams). An instance of this approach can be observed in [8], the place classification using word sequences used to be examined on 350 poems in Spanish through five authors giving about 83% accuracy.

Features, which normally supply very excessive accuracy measures are personality n-grams, i.e. sequences of n characters extracted from phrases performing in documents. They are considered language independent, i.e. they can be extracted from texts in a variety of languages regardless of persona units used. See, for example, [9] for reviews on authorship attribution of English, Greek, and Chinese texts. In our opinion very accurate effects of their utility need to be handled with caution: there is an apparent useful dependence between report content and personality n-grams, so they may additionally represent and alternative representation of feature phrases (what is probable good) or they may also simply render document content material (what appears to be worse).

Tareef proposed a new Stylometric approach recognised as the Stylometric Authorship Balanced Attribution (SABA) which in a position to analyze texts in text mining, e.g., novels and performs by means of famous authors, attempting to measure the author’s style, by way of deciding on some attributes that exhibit author's style of writing, assuming that these writers have a one of a kind way of writing that no different creator has, with greater accuracy prediction and impartial from human judgments, which ability that the technique does not count on the domain experts. This method is implemented by using merging three methods, which are called the computational approach, the Winnow algorithm, and the Burrows-delta method. The algorithm regarded an unguided mannequin and it tested in the English language correctly with noticeable prediction [10].

III. Stylometric Authorship Attribution

Stylometry is the study of writing style based totally on linguistic elements and is typically applied to authorship attribution troubles [11].

SAA was once begun as a “Content analysis” and was described as “understanding data now not as a series of bodily activities but as symbolic phenomena and to strategy their evaluation unobtrusively. Methods in the natural sciences do now not want to be worried about meanings, references, consequences, and intentions. Methods in social research that derive from these tough disciplines manipulate to omit these phenomena for convenience”. The time period content material evaluation is about 50 years old. Webster’s English Dictionary has listed it solely considering 1961 [12].

IV. Stylometric Authorship Balanced Attribution (SABA)

The SABA method is compared towards three different strategies the use of the computational approach, the Winnow algorithm method, and the Burrows-delta method. The results showed that the SABA method produces most useful prediction accuracy and even presents a completely right end result during the closing stage of the test [10].

The SABA method way is by neglecting the maximum values for the attribute frequencies and replacing it with “balanced” frequency. The idea that the right attributes are the “stabilized” or “balanced” attributes rather than attribute with the maximum frequencies. This means that in a written paragraph from a novel with assuming 10000 words, if a specific writer had used a specific word between 200-250 times in all of his books, then consider the attribute “word” has a “stabled” frequency percentage, hence is not a maximum frequency count[10].

V. Burrows Delta Method

While many methods have been utilized to the hassle of computerized authorship attribution, John F. Burrows’s “Delta Method” [13] is an especially simple, yet effective. The purpose is to robotically determine, based on a set of known education archives labeled by using their authors, who the most probably creator is for an unlabeled check document. The Delta technique makes use of the most usual words in the education corpus as the facets that it makes use of to make these judgments. The Delta measure is described as: The suggestion of the absolute differences between the z-scores for a set of phrase variables in a given text-group and the rankings for the same set of word-variables in a target text [14].

VI. Methodology

Data is taken from the web site www.Gutenberg.org. The dataset is an incredible cross segment of nineteenth century English writing as appropriately as various work. Utilizing this accumulation; we assembled books from 5 of the best 100 most downloaded writers; collected 10 books from every one of the 5 writers and they are Charles Dickens, Jack London, William Shakespeare, Mark twain and Oscar Wilde.

Both algorithms (Burrow-Delta and SABA methods) sharing same first steps, starting by uploading the chosen novels in text mode (with .txt extension), steps of cleaning and chunking are performed (removing double spaces, punctuation marks, special characters, symbol and others) before the implementation of the process of transforming text into Microsoft Access 2010 database files; taking into account that every single record contains frequent or a pair or trio words.

All tests implemented in this experiment by using Microsoft Access 2010 database and Visual C#, and choose ten books for the famous author(Charles Dickens, Jack London, William Shakespeare, Mark twain and Oscar Wilde) (nine for Learn, one for test).

A. Burrow Delta Method

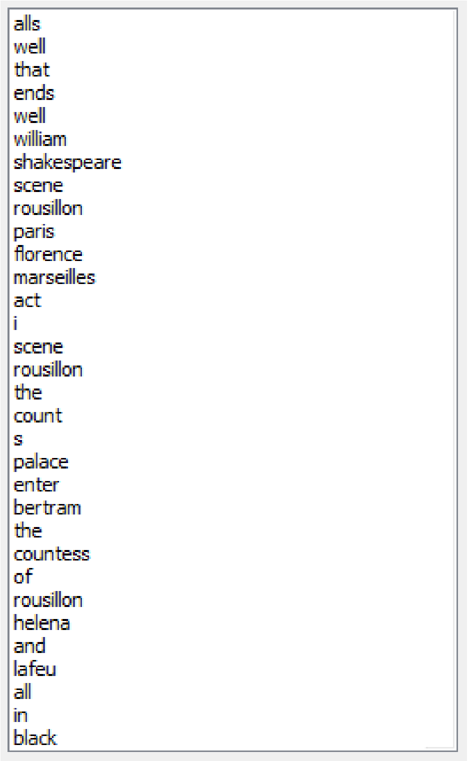
Burrow Delta represents the mean of the outright contrasts between the z-scores for an arrangement of word factors in a given text-gathering and the z-scores for a similar arrangement of word-factors in an objective text. The working steps will be implemented in detail in the case of frequent, pair, trio words.

The first step is to transform the book to be tested in text mode (.txt) into a separated list of book words. The final result of this is shown in Fig. 1. This operation will be executed for all learning and testing books.

Next, group the similar records, and calculate to the redundancy of these records, finally store the result in a separate table, the final result of this is shown in Fig. 2.

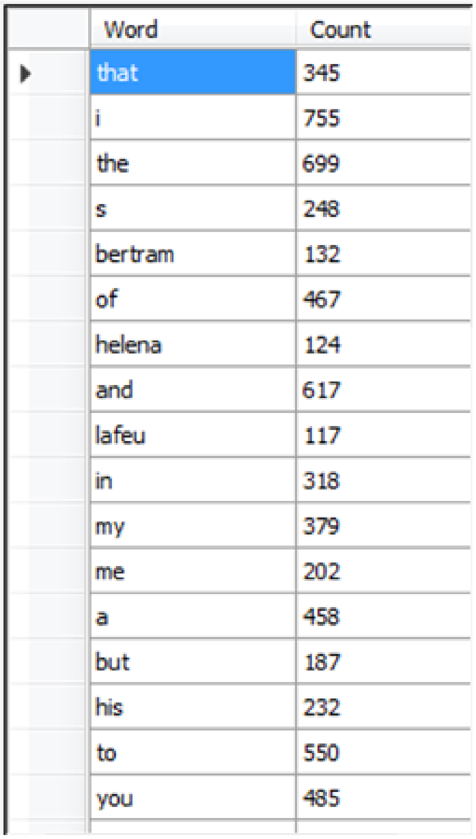
The next step is to cancel the differences between the size of books, by taking the percentage that speaks to the number of frequencies for each property separated by the entirety of frequencies for every one of the qualities multiplied by1000 in order to get a frequency that equal in weight for all used books and give true indication about the style of the author, the final result of this is shown in Fig. 3.

The following are making a stylometric map, by Merging and assembling all of the nine books (learning data) of the author which is being tested in a single table and make a relationship between their fields, calculate the arithmetic average of the redundancies.



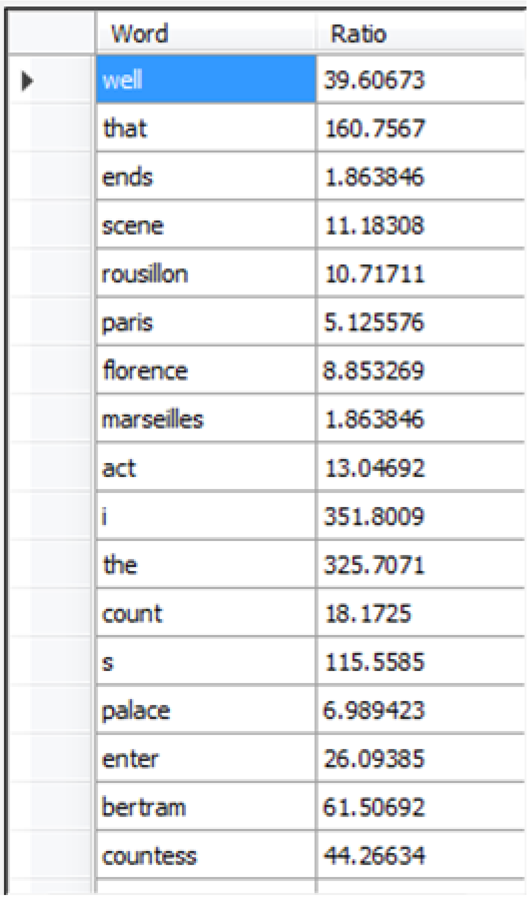
alls
well
that
ends
well
william
shakespeare
scene
rousillon
paris
florence
marseilles
act
i
scene
rousillon
the
count
s
palace
enter
bertram
the
countess
of
rousillon
helena
and
lafeu
all
in
black

Fig. 1. List of book words.



Word	Count
that	345
i	755
the	699
s	248
bertram	132
of	467
helena	124
and	617
lafeu	117
in	318
my	379
me	202
a	458
but	187
his	232
to	550
you	485

Fig. 2. Records redundancy.



Word	Ratio
well	39.60673
that	160.7567
ends	1.863846
scene	11.18308
rousillon	10.71711
paris	5.125576
florence	8.853269
marseilles	1.863846
act	13.04692
i	351.8009
the	325.7071
count	18.1725
s	115.5585
palace	6.989423
enter	26.09385
bertram	61.50692
countess	44.26634

Fig. 3. Book word ratio.

Index the total arithmetic average descending as shown in the following steps:

1) Merge all the learning data and save the result in a single table.

2) Assemble the result of merging data from the previous table and save the result in a new single table.

3) Make a relationship between their fields, and calculate the arithmetic average of the redundancies.

By calculating the average for all fields of the learning data and sorting it in descending, the stylometric map is ready now for the purpose of testing with other authors' books.

The Stylometric map is prepared for the purpose of examination and testing it, by building connections between the stylometric outline the five test books for all writers to get a new distribution of attributes based on the stylometric map that has been extracted.

In addition, this operation isolates the features that do not participate in any redundancy, that means if there are no common attributes between the learning books and testing books the main attributes will be isolated it by this operation, this step is important in order to make the stylometric map more stronger and reflecting a true style of the author.

After sorting the stylometric database map in the descending order based on the average percentage value for each attribute member in attribution set.

For Pearson, during the last step, select top 300 attributes that have the highest average percentage value in the stylometric map. Extract the Pearson correlation for the particular author's stylometric map from each of the five test books, hence giving five Pearson values. By having the weights for every parameter, increase each Pearson esteem by - 1 on the off chance that it is the wrong creator for the already known outcome or by +1 on the off chance that it is the correct writer.

For Spearman, a new table is configured that consist of 5 maps and 1 test. Each word corresponds to the ratio and the Rank (this rank is based on rank). Then works on it a word search function of the test, search on each map if found, take the rank for that word (only in this map), if not, they are compensated by zero. The result of this procedure is a table consisting of the test words only correspond to the word rank value and the rank of the word that was found at the specific map. The next step is applying spearman equation which also has a range between 1 to -1.

The Spearman connection between two factors is equivalent to the Pearson relationship between the rank estimations of those two factors; while Pearson's connection surveys straight connections, and Spearman's relationship evaluates the monotonic relationship.

B. SABA Method

The stylometric authorship balanced attribution (SABA) technique thought about an advancement of the calculation of Burrow-Delta strategy, this strategy relies upon the coefficient of difference (CV), which is spoken to as a factual estimation that isn't influenced by the perception of mean. Then will be analyzed and tried this calculation in English dialect in the regular, match and trio words.

In SABA technique, the trial of successive, match and trio words is like the Burrow Delta strategy in application, however there is basic contrast between them, precisely while choosing the highest point of 300 characteristics, these determinations rely upon the estimations of coefficient of variety (C.V), the accompanying case visit words can outline the real strides of removing the (C.V) And the strategy for choosing the required properties.

To apply SABA technique, rehash all the past strides as their request in the Burrow Delta strategy, at that point change the last stylometric guide to remove the estimations of the normal, the standard deviation (S.D) and the coefficient of variety (C.V) for every trait in the learning of the data, the (C.V) can be found by isolating the standard deviation by the mean itself, Finally, record the data in rising request in light of the estimations of the coefficient of variety (C.V) and select the main 300 qualities. In the wake of building connections between the last stylometric delineate the test books for all writers as we did on the Burrow Delta test, get the last successive test in SABA technique.

For Pearson, by having the weights for every parameter, duplicate each Pearson esteem by - 1 on the off chance that it is the wrong creator for the beforehand known outcome or by +1 in the event that it is the correct creator.

For Spearman, if there are no rehashed data esteems, a flawless Spearman relationship of +1 or −1 happens when every one of the factors is an ideal monotone capacity of the other. It merits saying that the utilization of Spearman is it requires less investment to contrast and Pearson and utilize basic numbers and less unpredictable in light of the utilization of the Rank rather than copies.

VII. Results

A. Burrow Delta Method and Pearson

The first step in this test is done on three authors only was the expectations of true and 0% error rate whether for frequent or pair or trio.

After applying it to five authors, it was found that there was an error of 20%.

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table I.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table II.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table III.

- Summary

The results of the prediction for the frequent word and word pair were better than the trio. Although the results of trio words are less accurate than pair and frequent word, because the frequent word results and word pair don't contain any percentage of error prediction.

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$pair\ words\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$trio\ words\ prediction\ error = \frac{1}{5} \times 100\% = 20\%$$

TABLE. I. PEARSON CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare test**	**Pearson in Wilde test**	**Pearson in London test**	**Pearson in Twain test**
Dickens Stylometric Map	0.852674	0.586294	0.639235	0.655396	0.835108
Shakespeare Stylometric Map	0.66921	0.768426	0.615545	0.592736	0.718333
Wilde Stylometric Map	0.782839	0.622714	0.701601	0.638623	0.802786
London Stylometric Map	0.775219	0.554586	0.575797	0.738936	0.827077
Twain Stylometric Map	0.760399	0.597347	0.587423	0.70423	0.890519

TABLE. II. PEARSON CORRELATION COEFFICIENT RESULTS IN A FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare test**	**Pearson in Wilde test**	**Pearson in London test**	**Pearson in Twain test**
Dickens Stylometric Map	0.657954	0.351077	0.304583	0.449358	0.610181
Shakespeare Stylometric Map	0.385408	0.607154	0.372539	0.340183	0.411975
Wilde Stylometric Map	0.484741	0.383454	0.515655	0.428261	0.560584
London Stylometric Map	0.492758	0.321433	0.251817	0.539384	0.491636
Twain Stylometric Map	0.532386	0.409412	0.326204	0.482538	0.684761

TABLE. III. PEARSON CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare test**	**Pearson in Wilde test**	**Pearson in London test**	**Pearson in Twain test**
Dickens Stylometric Map	0.299347	0.073595	0.262487	0.293538	0.243407
Shakespeare Stylometric Map	-0.07354	0.220364	0.187214	0.092378	0.066574
Wilde Stylometric Map	0.215399	0.102512	0.339212	0.259741	0.264372
London Stylometric Map	0.259402	-0.02263	0.226403	0.349504	0.388979
Twain Stylometric Map	0.269146	0.108761	0.237624	0.371188	0.509085

However the experiment showed that the frequent word and word pair is the higher predicted values, and represents the best attribute according to the true prediction values for all results. This test use complex equations and numbers and take more time compared with the use of Spearman and Rank algorithm.

B. Burrow Delta Method and Spearman

The first step in this test is done on three authors only was the expectations of true and 0% error rate whether for frequent or pair or trio.

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table IV.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table V.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table VI.

- Summary

The results of the prediction for the frequent word, pair and trio were best possible, because of all results don’t contain any percentage of error prediction.

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$pair\ words\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$trio\ words\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

TABLE. IV. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare test**	**Spearman in Wilde test**	**Spearman in London test**	**Spearman in Twain test**
Dickens Stylometric Map	0.819973	0.330999	0.464541	0.480149	0.758905
Shakespeare Stylometric Map	0.406229	0.666872	0.305217	0.217681	0.490131
Wilde Stylometric Map	0.663092	0.389021	0.544767	0.411393	0.688602
London Stylometric Map	0.67824	0.234515	0.302373	0.648795	0.74929
Twain Stylometric Map	0.673973	0.329172	0.383627	0.549095	0.847885

TABLE. V. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare test**	**Spearman in Wilde test**	**Spearman in London test**	**Spearman in Twain test**
Dickens Stylometric Map	0.514772	-0.18968	-0.23314	0.062879	0.47578
Shakespeare Stylometric Map	-0.09823	0.404939	-0.14157	-0.33312	0.035592
Wilde Stylometric Map	0.158636	-0.0908	0.158825	-0.06006	0.386241
London Stylometric Map	0.218483	-0.30366	-0.41439	0.257186	0.336782
Twain Stylometric Map	0.227489	-0.15582	-0.24058	0.063646	0.577259

TABLE. VI. SPEARMAN CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare test**	**Spearman in Wilde test**	**Spearman in London test**	**Spearman in Twain test**
Dickens Stylometric Map	-0.38487	-0.88527	-0.65054	-0.559	-0.52368
Shakespeare Stylometric Map	-0.99799	-0.67918	-0.80241	-0.93396	-0.89839
Wilde Stylometric Map	-0.62677	-0.86367	-0.445	-0.70271	-0.64397
London Stylometric Map	-0.51367	-0.9769	-0.74875	-0.30116	-0.26849
Twain Stylometric Map	-0.47065	-0.8657	-0.63585	-0.37992	-0.02225

However, the experiment showed that all test have perfect predicted values and represents the best attribute according to the true prediction values for all results. In this experiment the Speed and accuracy at a high rate, using the Spearman equation, which is less complex than Pearson's equation, it takes less time to compare with Pearson, work faster because taking from the test only 300 attributes means we did not adopt all the attributes values. Cancellation of CV and adoption of Ratio, use simple and less complex numbers because of the use of the Rank algorithm instead of the frequencies. Change the experience from 5 test 1 map To 5 map 1 test. It is worth mentioning that in this experiment was obtained perfect results.

C. *SABA method and Pearson*

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table VII.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table VIII.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table IX.

- **S**ummary

The results of the prediction for the frequent word and word pair were worse than the trio. Although the results of trio words are better accurate than pair and frequent word, because the trio word results don't contain any percentage of error prediction.

TABLE. VII. PEARSON CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare**	**Pearson in Wilde**
Dickens Stylometric Map	0.538038	0.428293	0.478812
Shakespeare Stylometric Map	0.555541	0.546308	0.500479
Wilde Stylometric Map	0.566413	0.451176	0.422471

TABLE. VIII. PEARSON CORRELATION COEFFICIENT RESULTS IN THE FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare**	**Pearson in Wilde**
Dickens Stylometric Map	0.490773	0.32738	0.300934
Shakespeare Stylometric Map	0.382706	0.44047	0.372926
Wilde Stylometric Map	0.405736	0.257335	0.294741

TABLE. IX. PEARSON CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare**	**Pearson in Wilde**
Dickens Stylometric Map	0.232979	0.06033	0.203679
Shakespeare Stylometric Map	-0.09015	0.18197	0.180051
Wilde Stylometric Map	0.146199	0.049749	0.336961

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{2}{3} \times 100\% = 66\%$$

$$pair\ words\ prediction\ error = \frac{1}{3} \times 100\% = 33\%$$

$$trio\ words\ prediction\ error = \frac{0}{3} \times 100\% = 00\%$$

However the experiment showed that the frequent word and word pair is the less predicted values, and represents the worse attribute according to the true prediction values for all results. Use CV, this cause the consumption time to be longer than the ratio used. It also has long equations and complex numbers. Because there is a false expectation in the frequent (Table VII) and Pair (Table VIII), this test was not applied to all authors because the error rate will increase.

D. SABA Method and Spearman

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table X.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table XI.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table XII.

- Summary

Results of the prediction for the trio word was best possible, because of other results contain percentage of error prediction.

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{1}{3} \times 100\% = 33\%$$

$$pair\ words\ prediction\ error = \frac{0}{3} \times 100\% = 0\%$$

$$trio\ words\ prediction\ error = \frac{1}{3} \times 100\% = 33\%$$

However the experiment showed that the pair word is the higher predicted values, and represents the best attribute according to the true prediction values for all results.

Use CV, this cause the consumption time to be longer than the ratio used. It also has long equations and complex numbers. Because there is a false expectation in the frequent (Table X) and Pair (Table XI), this test was not applied to all authors because the error rate will increase.

TABLE. X. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare**	**Spearman in Wilde**
Dickens Stylometric Map	0.47352	0.094352	0.223405
Shakespeare Stylometric Map	0.32946	0.359226	0.17037
Wilde Stylometric Map	0.402448	0.102134	0.153217

TABLE. XI. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare**	**Spearman in Wilde**
Dickens Stylometric Map	0.28902	-0.19774	-0.19657
Shakespeare Stylometric Map	-0.09847	0.171194	-0.14929
Wilde Stylometric Map	0.073531	-0.25251	-0.10344

TABLE. XII. SPEARMAN CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare**	**Spearman in Wilde**
Dickens Stylometric Map	-0.37887	-0.86592	-0.61241
Shakespeare Stylometric Map	-1.01489	-0.71803	-0.80471
Wilde Stylometric Map	-0.62162	-0.90985	-0.36788

VIII. Conclusions

The first contribution is gain, a better prediction accuracy by involving the statistical Pearson correlation and Spearman correlation as a main weighting factor in the SABA and burrows method. And do not overlook that using the Spearman algorithm which is less complex compared to Pearson with the burrows algorithm led to optimal prediction results. The next contribution is improving the feature extraction process by introducing a new set of more dependable attributes, such as the word pair and the trio, in addition to the use of classical frequent words. The results showed that using Spearman correlation coefficients measure leads to, zero error prediction, Speed, and accuracy at a high rate, the Spearman Equation which is less complex than the Pearson Equation and it takes less time to compare with Pearson. The main consideration in this treatise is that the results are best when used ratio rather than CV, use simple numbers and less complicated because of the use of the Rank algorithm instead of frequencies matches. Conducting optimal predictors result in SARA compared with SABA and burrows. Replace ratio value with attribute ranks make the calculations more easy and speedy.

References

[1] Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

[2] Stańczyk, U. (2016). The class imbalance problem in construction of training datasets for authorship attribution. In *Man-Machine Interactions 4* (pp. 535-547). Springer, Cham.

[3] Korasidi Andriana Maria (2016) “Authorship Attribution Forensics: Feature selection methods in authorship identification using a small e-mail dataset.” Master thesis, University of Athens.

[4] Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

[5] White, D. R., & Joy, M. S. (2004). Sentence-based natural language plagiarism detection. *Journal on Educational Resources in Computing (JERIC)*, *4*(4), 2.

[6] Zhang, Y., Chen, M., & Liu, L. (2015, September). A review on text mining. In *Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on* (pp. 681-685). IEEE.

[7] Cohen, J. (1988). Statistical power analysis for the behavioral sciences 2nd edn.

[8] Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66.

[9] Sullivan M. “Fundamentals of Statistics”. Pearson Education. Canada, 2010.

[10] Mustafa, Tareef Kamil (2011), “Stylometric authorship balanced attribution prediction method”. PhD thesis, Universiti Putra Malaysia.

[11] Dauber, E., Overdorf, R., & Greenstadt, R. (2017, June). Stylometric Authorship Attribution of Collaborative Documents. In *International Conference on Cyber Security Cryptography and Machine Learning* (pp. 115-135). Springer, Cham.

[12] Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

[13] Burrows, J. (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, *17*(3), 267-287.

[14] Burrows, J. (2002). The Englishing of Juvenal: computational stylistics and translated texts. *Style*, *36*(4), 677-698.