Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights

Viktor Shynkarenko and Inna Demidovich

*Dnipro National University of Railway Transport named after academician V. Lazaryan 2,
academician Lazaryan str., Dnipro, 49010, Ukraine*

Abstract

In this work we try to improve the results of texts and their fragments attribution using the classification method of the least distance in Euclidean space of images, by selecting weights for each of the image measures. For weights determination the genetic algorithm was used. Images are formed using statistical and modified recurrent analysis and the text complexity indicators. We will try to identify the effectiveness for each of them.

It was found that this method usage improves the efficiency of the text attribution and the reliability of authorship determination of the texts from the control sample reaches 80-91%.

Keywords 1

natural language texts, recurrence analysis, frequency analysis, text complexity, text authorship, classification, genetic algorithm

1. Introduction

The authorship determination of natural language texts is quite a relevant topic [1]. A technique that allows determining the authorship of a particular text with sufficient reliability can be widely used in many areas such as education, jurisprudence, literary criticism, etc. Despite a large number of researches [1], there is no way to determine the authorship of even literary texts with a 100% guarantee.

In this work, it is planned to increase the authorship determination reliability for natural language literary texts. We plan to select the text attributes that have the greatest information content among the selected indicators, that according to other studies adequately reflect the author's syllable and style. In addition, the effectiveness of modified for working with texts recurrent analysis to determine the author's style, and, as a consequence, the authorship of the text will be investigated.

To determine the ponderability for each of the indicators presented, weights will be used that are tuned using a genetic algorithm.

2. Related works

One of the problems authorship determination methodologies of the literary texts is the complexity of choosing text parameters that will determine the author's style [2].

Various methods of text attribution have been used, but results with the greatest accuracy are obtained by using the text's character frequency [3], N-gram [4,5] and their variations, as well as the frequency of words (all of them or any of their separate categories [6]) and word's parts [7]. The previously cited studies show that the usage of N-grams allows reflecting the personal author style with the best accuracy due to the fixation of the sequences of lexical structures [6, 8].

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: shinkarenko_vi@ua.fm (V. Shynkarenko); 2019demidovichinn@gmail.com (I. Demidovich)
ORCID: 0000-0001-8738-7225 (V. Shynkarenko); 0000-0002-3644-184X (I. Demidovich)

These techniques were widely used to determine the authorship of texts that were written in various languages and various topics [9, 10, 12] and have shown their effectiveness. Similar studies also were done for texts in the Ukrainian language [11].

In this work, the most informative indicators from various classes, reflecting the author's style, are distinguished to determine the authorship of literary texts.

3. Methods

3.1. The classes of indicators in the text's image

The text's image is a vector that contains the indicators measurements of the following classes: the frequency analysis, the text perception complexity and modified recurrence analysis for natural language texts and it is used for authorship determination. Let's consider each class of these indicators.

3.1.1. Frequency analysis of texts

Frequency analysis is one of the most common text analysis methods. For many languages and a large number of authors, linguists compiled a frequency dictionary of the author's language or his single texts [12-15]. The frequency of a single character occurrence calculation for a specific text is the basis of such text processing [16, 17]. Based on the obtained data, we can conclude that each text will have its own individual frequency structure. However, during the analysis of the sufficiently large text, it was found that the occurrence frequency of a specific letter in the alphabet will be very close to the frequency of its occurrence in any text of various authors and to the frequency of its use in the language. It will be not so much a characteristic of the author's style or text as the language that the author use [18, 19].

To solve this problem and obtain more reliable information that could characterize the text, it was decided to use larger structures than the letters of the alphabet –N-grams [20].

This method arose relatively recently and is often used to detect plagiarism [21]. An N-gram is a sequence of characters in a text with length N. Depending on the value of N the occurrence frequency of individual words or phrases will be analyzed in the text. Using this method of the text frequency analysis, it becomes possible to distinguish texts by the author's style, since it is formed through not least the service parts of speech, inserted words and structures that are inherent in the particular author's style [22].

In previous studies [32], we get the best result in the authorship determination of the text using 4-grams. In this work, to compare the effectiveness of the methods, 1- and 4-gram frequency analysis was performed.

We formed the text's image with the frequency of each character in the text in the case of 1-gram (letters) and 100 of the most frequently encountered 4-gram.

Let's give an example of the text's image formation using the work of T. Shevchenko "Saul", the work contains 2148 analyzed symbols. The diagram presents the text analysis results (Figure 1).

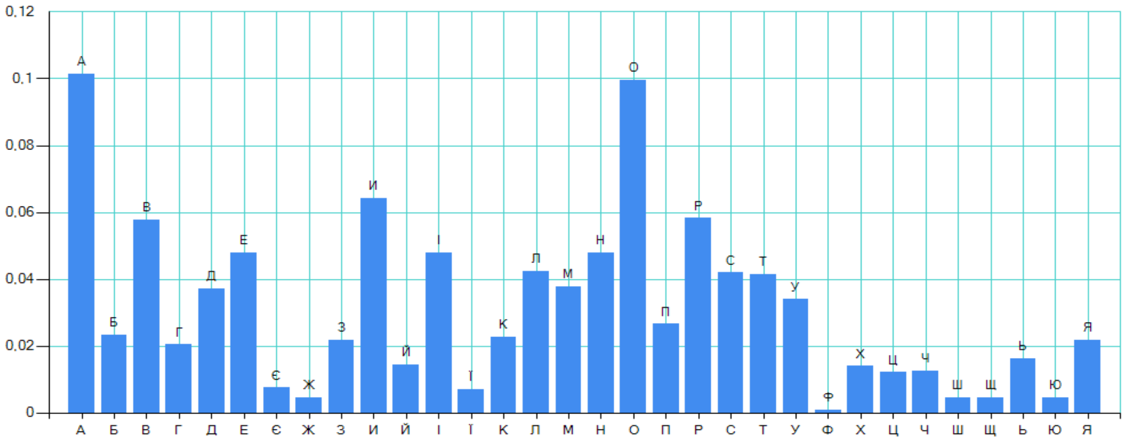

Figure 1: Text character frequency chart

The data are arranged in alphabetical order and allow you to visually evaluate the occurrence frequency of each letter, typical for a given text.

To assess the uniqueness of the obtained values and their applicability for attribution, we should compare the obtained data with information on the average occurrence frequency of the letters in the Ukrainian language. According to a study [31], the most frequently encountered letters of the Ukrainian language are О, А, Н. Letters И, Т, В, Е, Р, І, С, К, М will have the secondary frequency of propagation.

Letters А, О, И, Р, В, Н, І, Е, Л, С have the highest frequency of occurrences in the studied text "Saul" by T. Shevchenko. As you can see, there is a certain discrepancy with the average frequencies of letters in the language and in the studied work, which allows us to take them as one of the author's style characteristics.

The initial values in the vector of the text's image of the "Saul" by T. Shevchenko, include the frequency indicators (Figure 1: Text character frequency chart): $X' =$[0.1014, 0.0233, 0.0577, 0.0205, 0.0372, 0.0480, 0.0074, 0.00466, 0.0219, 0.0642, 0.0144, 0.0480, 0.0070, 0.0228, 0.0424, 0.0377, 0.0480, 0.0996, 0.0265, 0.0582, 0.0419, 0.0414, 0.0340, 0.0009, 0.0140, 0.0121, 0.0126, 0.0046, 0.0046, 0.0163, 0,0047, 0.0219, …]

3.1.2. The text perception complexity indicators

The next characteristic of the author's text is its structural complexity and complexity of perception. There are a number of metrics that help determine the level of text difficultness. Among them are the number of sentences, words, syllables and letters in the text, as well as the average number of words, syllables, letters in sentences and words.

This analysis was carried out to assess the complexity of the perception of the English text [23], but it is also applicable to any other languages.

Using this method of text analysis, headings, subheadings and formulas are most often ignored, since they are not complete sentences.

This data also contains certain information about the author's writing style. However, such indicators consider the text complexity, but don't reflect its content and word order, based on this, these indicators don't have sufficient efficiency for analyzing the author's style on their own, but can be used in conjunction with other indicators.

Let's give an example of these indicators values according to the poem "Saul" by T. Shevchenko (Table 1).

Table 1
Indicators of the text perception complexity in the work of T. Shevchenko "Saul"

Indicator	Value
Number of words	458
Number of syllables	937
Number of sentences	36
Number of characters	2148
Average number of words in sentences	12.72
Average number of syllables in sentences	26.03
Average number of letters in sentences	59.67
Average number of syllables in words	2.25
Average number of letters in words	4.69

There is the text's image vector of "Saul" by T. Shevchenko, with the values of the text perception complexity indicators (Table 1): $X' =$[… 12.72, 26.03, 59.67, 2.25, 4.69, …]

The text complexity indicators also include data of the words number with different lengths in the text. For the studied poem, these indicators will be as follows (Table 2).

Table 2
Indicators of the word frequency with different lengths in the work of T. Shevchenko "Saul"

Word length	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	16	12	7	15	0.2	13	9.6	6.1	3	3	0.8	4	0.4	0	0.2

In the studied text, the largest word consists of 15 letters; there are no words of 14 letters in the text (Table 2). Based on the analysis of all available texts, it was decided to consider words with a length of 22 characters.

There is the text's image vector of "Saul" by T. Shevchenko, supplemented with the presented meanings Table 2: X' =[… 16, 12, 7, 15, 0.2, 13, 9.6, 6.1, 3, 3, 0.8, 4, 0.4, 0, 0.2, 0, 0, 0, 0, 0, 0, 0,…]

3.1.3. Modified recurrence analysis

The recurrence analysis is used to study time series and processes. We modified this type of analysis for its application in the processing of natural language texts. It is based on the quantitative analysis of RQA [25, 26, 27] recurrence plot that was used by Zbilut J. P. and Webber Jr. C. [24].

The text is transforming into a time series to apply the recurrence analysis for text processing. The value of each point in the series is the occurrence frequency of the N-gram, and the advancement to the next N-gram is considered as the time unit. The usage of recurrence analysis allows to some extent count the microstructure of the text, its individual language [33] and the author's style.

The resulting time series is a characteristic exclusively for this text and allows further research based on these data, including the construction of the phase space and the recurrence plot. There is the time series of the text of T. Shevchenko "Saul" (Figure 2).

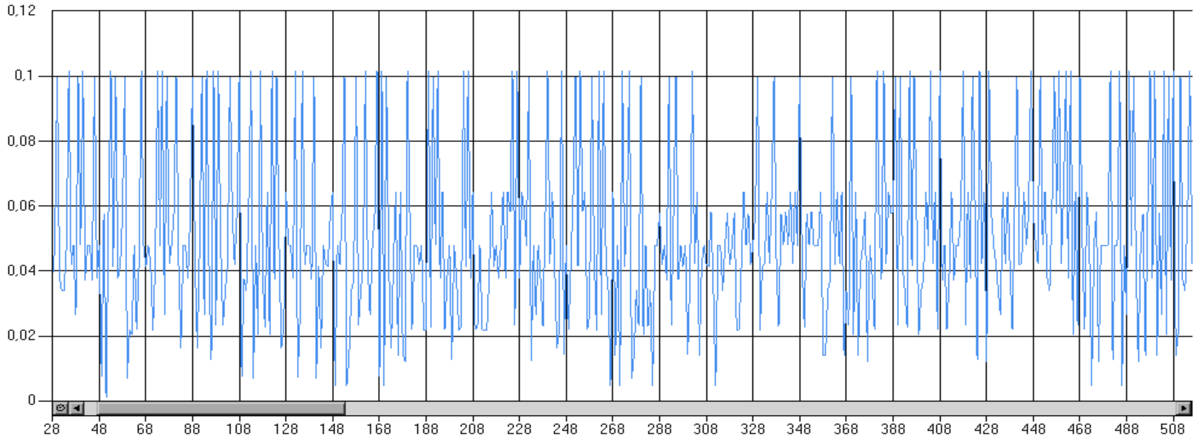

Figure 2: Time series of the text of T. Shevchenko "Saul"

According to the rules of constructing a recurrence plot, its size corresponds to the size of the text in characters and displays the number of repeated elements within the text. Based on the value of ε, the symbols in the processed text that have the same frequency or differ from each other less than 0.002 will be perceived as the same symbol and will be displayed as a filled point in the recurrence plot.

This plot displays repeating states at different moments in time (respectively, places in the text). The numbers of indicators are calculated using a recurrence chart:

- the recurrence rate in the analysis of the text displays the total number of repetitions of each statistically close sequence of characters.

$$RR = \frac{1}{K^2} \sum_{i,j=1}^{K} R_{i,j}^{n,\varepsilon}, \tag{1}$$

where K – number of the considered states, $R_{i,j}$ – *i*, *j*-th point of the recurrence plot, ε – the recurrence threshold in *i* moment, *n* – phase space dimension;

- determinism expresses the lengths frequency distribution of *l* diagonal lines in the in the plot $P^{\varepsilon}(l)$, K – the absolute number of such lines. This indicator displays the number of all sequences repetitions of statistically close N-grams with any length

$$DET = \frac{\sum_{l=l_{min}}^{K} lP^{\varepsilon}(l)}{\sum_{i,j}^{K} R_{i,j}^{n,\varepsilon}}; \tag{2}$$

- the mean line length L, which for the text analyzing process, will display the average number of repeated N-gram sequences

$$L = \frac{\sum_{l=l_{min}}^{K} lP^{\varepsilon}(l)}{\sum_{l=l_{min}}^{K} P^{\varepsilon}(l)}; \tag{3}$$

- divergence - the reciprocal of the diagonal structures maximum length

$$DIV = \frac{1}{max(\{l_i; i=1..K_l\})}; \tag{4}$$

- entropy is an indicator of the diagonal structures frequency distribution, which for the text analyzing process, will display the frequency distribution repetitions of consecutive statistically close N-grams sequences

$$ENTR = -\sum_{l=l_{min}}^{K} p(l)\ln(p), \tag{5}$$

where

$$p(l) = \frac{P^{\varepsilon}(l)}{\sum_{l=l_{min}}^{K} P^{\varepsilon}(l)}; \tag{6}$$

- laminarity displays the frequency distribution of the horizontal structures lengths on a recurrence plot and, for text analysis, expresses the frequency distribution repetitions of statistically close N-gram sequences

$$LAM = \frac{\sum_{v=v_{min}}^{K} vP^{\varepsilon}(v)}{\sum_{i,j}^{K} R_{i,j}^{n,\varepsilon}}, \tag{7}$$

where v – the length of the horizontal line on the plot $P^{\varepsilon}(v)$;

- time trapping – the average length of horizontal structures. As part of text analysis - the average length of statistically close N-gram sequences repetitions

$$TT = \frac{\sum_{v=v_{min}}^{K} vP^{\varepsilon}(v)}{\sum_{v=v_{min}}^{K} P^{\varepsilon}(v)}. \tag{8}$$

For T. Shevchenko's text "Saul" this indicators have next values Table 3.

Table 3

Recurrence analysis indicators for the text "Saul" by T. Shevchenko

Indicator	Value
RR	0.019
DET	0.002
DIV	0.111
L	2.236
ENTR	0.6
LAM	8.2
TT	2.275

There is the text's image vector of "Saul" by T. Shevchenko, supplemented with the indicators of the recurrence analysis Table 3: $X' =$[… 0,019, 0.002, 0.111, 2.236, 0.6, 8.2, 2.275].

And the text "Saul" by T. Shevchenko is put in correspondence with the text's image – vector $X' =$[0.1014, 0.0233, 0.0577, 0.0205, 0.0372, 0.0480, 0.0074, 0.00466, 0.0219, 0.0642, 0.0144, 0.0480, 0.0070, 0.0228, 0.0424, 0.0377, 0.0480, 0.0996, 0.0265, 0.0582, 0.0419, 0.0414, 0.0340, 0.0009, 0.0140, 0.0121, 0.0126, 0.0046, 0.0046, 0.0163, 0,0047, 0.0219, 12.72, 26.03, 59.67, 2.25, 4.69, 16, 12, 7, 15, 0.2, 13, 9.6, 6.1, 3, 3, 0.8, 4, 0.4, 0, 0.2, 0, 0, 0, 0, 0, 0, 0, 0,019, 0.002, 0.111, 2.236, 0.6, 8.2, 2.275].

3.2. Method of text authorship determination

3.2.1. Minimum distance classification

To determine the authorship of the text, the theory of pattern recognition is used, or rather, its recognition method based on the minimum distance to the standard [33].

The essence of the well-known method [28] is as follows.

There are M classes of images $\omega_1, \omega_2, \ldots, \omega_M$, each is associated with a specific author and an image X_l of the text, which authorship must be established. This text is known to belong to one of these authors.

The standards of images for each class are determined $Z_1, Z_2, \ldots, Z_M$. The standard of the class is a vector containing the average values of each indicator according to the author's texts in the training sample.

In the previous paper [33], it was assumed that the image X_l belongs to the class ω_i (the text belongs to the i-th author) if $\rho(X_l, Z_i) < \rho(X_l, Z_j)$ for $\forall j \neq i$, where $\rho(X, Z)$ – the distance between the images of X and Z in Euclidean space.

Different indicators can have different units and scales. To solve this problem, the minimax normalization of each indicator in the vectors X and Z is used.

The image of the text includes 66 indicators. The information content of each in the pattern recognition problem is different. In this regard, in order to increase the recognition efficiency, it was decided to use the weight of indicators. In this case, the fitness function has the form:

$$d_{lm} = \sum_{i=1}^{N} w_i (x_{il} - z_{im})^2, \tag{9}$$

where i – indicator's number in the vector;

l – investigated text number, l = 1..L;

m – author's standard number;

w_i – the weight of the i-th indicator;

N – the number of indicators in the vector (the text's image);

x_{il}, z_{ik} – elements of vectors X_l and Z_k.

We will assume that the image X_l belongs to the class ω_i (the text belongs to the i-th author) if $d(X_l, Z_i) < d(X_l, Z_j)$ for $\forall j \neq i$.

The task is to find such weights of the indicators $\mathrm{w_i}$, so that the recognition accuracy is the maximum value. A genetic algorithm was used to solve this problem.

3.2.2. A genetic algorithm for determining the text indicators weights

Genetic algorithms are used to solve the optimizing problem of the multi parametric functions' value. All presented tasks are formed as functions that depend on a number of parameters, the global maximum or minimum of which will correspond to the solution of the problem.

The genetic algorithm idea is the organization of the evolutionary process to obtain the final optimal solution [29, 30, 34]. It retains biological terminology. Thus, a chromosome is a vector, each position of which is called a gene. Each such vector (individual) is characterized by a certain health function (fitness function). This function determines the quality of the presented solution. The optimization problem can be considered as the problem of finding an individual with the best health function. The search is based on the heredity mechanisms, variability, and the selection and is implemented using various genetic operations. Crossover is an operation in which two chromosomes exchange their parts. Mutation is a random change in one or more positions in a chromosome.

Working with a genetic algorithm, the initial population is usually generated randomly. The only criterion is a sufficient variety of individuals to avoid populations falling into the local extremum.

After the generation of the first generation, the genetic algorithm imitates the evolutionary process as a repetitive process of reproduction and mutation; the probability of an individual's participation in reproduction is directly proportional to its health. The result is a new population, and the old one dies, thus, the health function of all individuals from generation to generation, in average, improves. Subsequently, the process is repeated until the health function stops improving. As a result, individuals with the best health function indicator from the last generation are selected.

In this work, the genetic algorithm has the following characteristics:

- fixed population size;
- fixed length of genes;
- proportional selection;
- individuals for reproduction are selected among the best representatives of the population;
- single point crossing;
- descendants take the place of the previous population;
- a fixed number of randomly generated individuals is added to each population to avoid population degeneration.

The health (fitness) function determines the amount of correctly authorship determination for the texts from the training sample. The initial population is randomly generated. During the simulation of the evolutionary process, the selection of individuals for the next generation was carried out according to the following proportions: 34% of the parent individuals with the best indicators of the health function interbred with each other, 60% of the remaining parent individuals mutated randomly, 6% of the offspring individuals were generated randomly to eliminate the population degeneration. In the experiment, the sample size is fixed at 100 individuals in each generation.

4. Results

4.1. Training and control samples

In the experiment the symbol-by-symbol analysis and 4-gram analysis were done.

The literary texts were selected for the training sample. This is based on a clear representation of the author's style and his personality, as well as reliable information about the authorship.

The training sample consists of 20 works by 11 different Ukrainian authors, and the control sample has 3 works written by the same authors in it. We took only prose works, up to 10 thousand characters or their parts of this size.

We took the following authors for the sample: 1 – I. Bahrianyi, 2 – A. Vyshnia, 3 – M. Vovchok, 4 – A. Dovzhenko, 5 – M. Kotsiubynskyi, 6 – H. Kvitka-Osnovianenko, 7 – P. Myrnyi, 8 – V. Nestaiko, 9 – V. Pidmohylnyi, 10 – I. Franko, 11 – M. Khvylovyi.

4.2. Chromosome formation

In the text's image in experiments with 1-grams, positions from 1 to 32 are occupied by the values of the letters frequency, arranged in alphabetical order (Figure 1), places from 33 to 37 are taken by data on the complexity of the text (Table 1), from 38 to 59 the frequency of words with different lengths in the text (Table 2), from 60 to 66 - indicators of recurrence analysis (Table 3).

For example, the image of the work of T. Shevchenko "Saul":

X' =[0.1014, 0.0233, 0.0577, 0.0205, 0.0372, 0.0480, 0.0074, 0.00466, 0.0219, 0.0642, 0.0144, 0.0480, 0.0070, 0.0228, 0.0424, 0.0377, 0.0480, 0.0996, 0.0265, 0.0582, 0.0419, 0.0414, 0.0340, 0.0009, 0.0140, 0.0121, 0.0126, 0.0046, 0.0046, 0.0163, 0.0047, 0.0219, 12.72, 26.03, 59.67, 2.25, 4.69, 71, 55, 30, 68, 61, 60, 44, 28, 15, 14, 4, 4, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,019, 0.002, 0.111, 2.236, 0.6, 8.2, 2.275].

The formed chromosomes contain 67 elements each element weights in the vector-image of the text and the recurrence threshold weight ε. The initial chromosome values are set randomly.

In experiments with 4 grams, the 100 most common combinations were selected for analysis. In this case, the number of genes in the chromosome was increased to 135 and the first 100 places in the chromosome are taken by weights to the most common 4-grams, weights for the text complexity parameters are on 101-105 places, 106-127 contain weights for data on the number of words with different lengths in the text, the weights for recurrence indicators has 128-134 places and the last place is held by the weight for ε, respectively.

4.2.1. Experiment results with 1-gram

The results of authorship determination using 1-gram are shown in Table 4, shaded cells - cases in which the author of a work from the control sample was identified correctly.

Working with 1-gram, 15 generations of chromosomes were formed. The calculated weights made it possible to improve the result of the text authorship determination four times more from 6 to 24 out of 33 works in total, which was an improvement from 18% to 80%.

Table 4

The results of authorship determination with using weights and without them character by character

The real author of the text		0	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6
Experiment-identified author	without weights	1	9	9	2	2	3	9	10	11	3	11	5	7	3	3	10	9	11
	with weights	1	1	1	2	9	2	3	10	3	4	4	4	4	5	5	6	5	5
The real author of the text		7	7	7	8	8	8	9	9	9	10	10	10	11	11	11			
Experiment-identified author	without weights	7	5	7	6	3	10	7	3	11	8	9	10	10	9	11			
	with weights	7	11	7	8	11	8	9	9	9	10	10	10	11	5	11			

During the experiment, 64 chromosomes were obtained, which gave the largest number of correctly identified authors. To obtain a more complete picture of the indicators significance, there are the intervals of the weights for each of the indicators and their average values (Table 5-Table 8).

Table 5

Calculated weights for the frequencies of the letters in Ukrainian alphabet

Letter	Max	Min	Mean	Letter	Max	Min	Mean	Letter	Max	Min	Mean
А	1,726	1,399	1,596	І	1,331	0,392	0,894	Т	0,984	0,000	0,692
Б	2,550	2,299	2,426	Ї	5,171	4,812	4,988	У	1,221	1,088	1,160
В	1,140	0,921	1,060	Й	3,881	3,545	3,703	Ф	5,509	5,165	5,385
Г	3,287	3,002	3,135	К	1,148	1,114	1,128	Х	2,032	1,517	1,781
Ґ	1,851	1,737	1,800	Л	0,215	0,203	0,208	Ц	3,313	3,070	3,180
Д	0,759	0,699	0,729	М	1,072	0,988	1,030	Ч	4,879	4,497	4,652
Е	0,000	0,000	0,000	Н	2,442	2,175	2,332	Ш	5,544	5,110	5,327
Є	5,049	4,599	4,819	О	3,842	2,309	3,261	Щ	3,850	3,549	3,705
Ж	2,643	2,493	2,564	П	3,590	3,270	3,432	Ь	3,850	3,567	3,738
З	4,842	2,830	3,728	Р	4,039	3,679	3,848	Ю	1,233	0,866	1,009
И	4,076	3,882	3,962	С	2,675	2,518	2,592	Я	5,088	0,805	1,587

According to the data presented in Table 5 among the weights of all chromosomes in the symbol frequency class, the letters Ш, Ф, Ї, Я, Є received the biggest weight.

It should be noted that these letters are not the most frequently encountered letters in the Ukrainian language, which allows us to consider their frequency as an informative characteristic of a particular text and the author's style in general. These indicators weights, except for the letter Я, the frequency also do not show significant fluctuations in values, which only confirm the earlier conclusion.

The frequencies of the letters Е, Л, Д, Т, М are the letters whose frequencies have the secondary weight among all. And the letters Я, З, О, Т, І have the greatest spread in the weights values for the entire time of the experiment.

Table 6

Calculated weights for words with different lengths

Word length	Max	Min	Mean	Word length	Max	Min	Mean
1	5,156	4,696	4,929	12	2,835	1,350	2,336
2	1,010	0,900	0,953	13	4,920	2,218	3,871
3	5,204	2,882	4,479	14	4,599	1,893	3,522
4	4,920	4,679	4,781	15	4,574	4,342	4,458
5	5,529	5,096	5,299	16	0,958	0,912	0,930
6	1,322	1,230	1,278	17	3,136	2,856	2,976
7	1,807	0,406	0,795	18	0,750	0,685	0,717
8	2,378	2,270	2,338	19	4,323	4,055	4,201
9	3,702	3,260	3,467	20	4,171	3,864	4,037
10	3,287	2,928	3,139	21	4,012	3,654	3,817
11	1,552	1,431	1,489	22	3,802	3,567	3,702

According to the data in Table 7 the TT a time trapping indicator has the greatest weight. This indicator represents the average length of statistically close N-gram sequences repetitions in the text. However, its weight also shows significant fluctuations.

Table 7

Calculated weights for recurrence analysis indicators

Indicator	Max	Min	Mean
DET	3,458	1,053	2,124
DIV	4,394	3,914	4,167
ENTR	4,166	3,728	3,990
L	3,874	3,602	3,714
LAM	4,082	1,749	3,361
RR	1,684	1,566	1,628
TT	4,651	2,140	4,086

The second in weight is the DIV (divergence) indicator witch is the inverse of the maximum length of diagonal structures, which in the text reflects the number of characters repeated sequences. The weight of this indicator is quite stable, which allows us to consider it as an important parameter of the text.

The ENTR (entropy) indicator and L (average length of diagonal lines) have less weight, but their fluctuations are also insignificant.

The laminarity indicator LAM, according to the obtained data, also has a significant weight, but in order to its strong fluctuation, this indicator cannot serve as a reliable characteristic of the text.

Table 8
Calculated weights for text complexity indicators

Indicator	Max	Min	Mean
letters in sentences	4,721	4,390	4,536
letters in words	4,309	3,041	3,446
words in sentences	0,319	0,297	0,306
syllables in sentences	0,862	0,793	0,829
syllables in words	4,848	4,509	4,670

Analyzing the obtained data (Table 6 and Table 8) we can assert that the largest among all weight in the text complexity class have indicators of the words frequency with length 5, 3, 1 and 4, as well as the number of syllables in words and letters in sentences.

In addition, these weights do not have significant fluctuations throughout the experiment. But there is a significant weight fluctuation for the words with length 3.

The last indicator in the chromosome – the weight for ε – varied from 0.053 to 2.842 with an average value of 1.095. According to these data, it can be concluded that a significant fluctuation in ε will lead to the fact that symbols sequences that are close in frequency will become indistinguishable. As a result, the recurrence plot will be distorted and will not be able to reflect the author's style fully.

4.2.2. Experiment results with 4 grams

An experiment was also conducted with 4-grams as the most effective option [32]. As a result, we get the following data (Table 9). Shaded cells - cases in which the author of the work from the control sample was identified correctly.

Table 9
The results of authorship determination using 4-grams with calculated weights and without them

The real author of the text		1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6
Experiment-identified author	without weights	0	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	5
	with weights	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	5

The real author of the text		7	7	7	8	8	8	9	9	9	10	10	10	11	11	11
Experiment-identified author	without weights	7	7	4	8	9	8	9	9	9	10	1	1	11	11	9
	with weights	7	7	7	8	8	8	9	9	9	10	1	1	11	11	11

Working with text analysis using 4-grams, 28 generations were formed during the experiment. The obtained weights made it possible to improve the result of text authorship determination only slightly, from 27 to 30 out of 33 works in total.

5. Discussion

According to our previous study [32], that contains authorship text determination without differentiation indicators by information content, the reliability result was 18% of the texts correct attribution for analysis character by character and 82% for analysis using 4-grams. Using the methods described in this work, the results for the character by character analysis improved significantly (up to 80%). The results of the analysis using 4-grams also improved, but to a lesser extent (up to 91%).

In recent years, there have been a lot of researches in the sphere of text attribution using different methods and different lengths and styles texts.

Vadim Moshkina, Ilya Andreeva, Nadezhda Yarushkina in their work [35] conducted a comparative analysis of various attribution methods. The architectures of a convolutional neural network, a multilayer perceptron, and LSTM neural network were proposed to solve this problem. It should be noted that the study was conducted on English poetry and took into account the peculiarities of this particular language. The credibility of determining the authorship for each of the methods studied in the article fell within the range from 74% to 83%.

Rahul Radhakrishnan Iyer, Carolyn Penstein Rose´ in their text attribution work [36] were using stylometric functions and various algorithms. This work was also carried out using English-language texts. Authors were able to achieve 82% confidence in text authorship attribution.

The following studies were carried out for Ukrainian texts and gave the following results: in the work [37] on determining the authorship of journalistic articles, the authors achieved 92% reliability in determining authorship using neural networks. And 79% when determining the authorship of scientific articles, using the Quantitative Method for Automated Text Authorship Attribution Based on the Statistical Analysis of N-grams Distribution [38].

Despite the difference in the methods, languages and styles of the studied texts, the results obtained in our study have sufficient reliability, along with other similar works. It can be concluded that this method can be used in various fields to determine the authorship of Ukrainian-language texts.

6. Conclusions

The usage of the genetic algorithm to find weights for various indicator classes helps to improve the results of natural language texts authorship determination. The calculated weights made it possible to improve the result four times (from 6 to 24 matches by the author). Working with 4 grams we got the result which also improved, but to a lesser extent - from 27 to 30 matches. The results obtained were 80% and 91%, respectively.

The application of this new technique has improved the result of text authorship determination in both cases, which allows us to declare its effectiveness. It helps us to identify most valuable indicators among all that have been analyzed. The effectiveness in authorship determination of different modified recurrence analysis indicators also were defined.

If we consider the presented classes of indicators and their weight in the text authorship determination, then the most important, according to their obtained weights, will be the frequencies of the letters Ф, Ш, Ї, Є, Ч. The text complexity indicators, namely the number of words in the text with a length of 5, 1 and 4 letters and the number of syllables in words and letters in sentences are the most important indicators in their class. The indicators of divergence, time trapping and entropy are the most informative for recurrence analysis.

In the future, to improve the result, it is planned to expand the number of analyzed indicators and conduct research using the stems of words. It is also planned to highlight a small list of various nature indicators; which combination will give the best result in determining the authorship of a natural language text.

7. References

[1] P. Juola, Authorship attribution. Foundations and Trends in Information Retrieval 1(3) (2006) 233–334. doi:10.1561/ 1500000005.

[2] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. 60(1) (2009) 9–26. doi: 10.1002/asi.20961.

[3] I. I. Drozdova, A. D. Obuhova, Opredelenie avtorstva teksta po chastotnyim harakteristikam (determining the authorship of the text by frequency characteristics), in: Tehnicheskie nauki v Rossii i za rubezhom: materialyi VII Mezhdunarodnoy nauchnoy konferentsii, Buki-Vedi, Moskva, 2017, pp. 18-21.

[4] Yunita Sari, Andreas Vlachos, Mark Stevenson, Continuous N-gram Representations for Authorship Attribution, in: Proceedings of the 15th Conference of the European Chapter of the

Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp.267-273. doi: 10.18653/v1/E17-2043.

[5] I. Markov, J. Baptista, O. Pichardo-Lagunas, Authorship Attribution in Portuguese Using Character N-grams. Acta Polytechnica Hungarica 14(3) (2017) 59–78. doi: 10.12700/APH.14.3.2017.3.4.

[6] D. L. Hoover, Frequent word sequences and statistical stylistics. Literary and Linguistic Computing 17 (2) (2002) 157-180. doi: 10.1093/llc/17.2.157.

[7] G. O. Sidorov, Automatic Authorship Attribution Using Syllables as Classification Features, Rhema journal 1 (2018) 62-81.

[8] J. Kruczek, P. Kruczek, M. Kuta, Are n-gram Categories Helpful in Text Classification, in: Computational Science ICCS 2020, Springer, Amsterdam, 2020, pp. 524-537. doi: 10.1007/978-3-030-50417-5_39.

[9] E. Stamatatos, On the robustness of authorship attribution based on character n-gram features. Journal of Law & Policy 21 (2) (2013) 427–439.

[10] H. Gómez-Adorno, JP. Posadas-Durán, G. Sidorov, Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing 100 (2018) 741–756. doi: 10.1007/s00607-018-0587-8

[11] O. Marchenko, A. Anisimov, A. Nykonenko, T. Rossada, E. Melnikov, Authorship attribution system. Artificial Intelligence 2 (2016) 77-85.

[12] G. Wimmer, G. Altmann, L. Hřebíček, S. Ondrejovič, S. Wimmerová, Úvod do analýzy textov, Univerzita Komenského v Bratislave, Bratislava, 2003.

[13] I.I. Popesku, G. Altmann, Some aspects of word frequencies. Glottometrics 13 (2006) 23-46.

[14] R. Köhler, G. Altmann, Aims and Methods of Quantitative Linguistics, in: G. Altmann, V. Levickij, V. Perebyinis (Eds.), Problems of Quantitative Linguistics, Ruta, Chernivtsi, 2005, pp. 12–41.

[15] V.I. Perebyinis, Statystychni metody dlia linhvistiv (statistical methods for linguists), 2nd. ed., Nova Knyha, Vinnytsia, 2013.

[16] P.M. Alekseev, Frequency dictionaries, in: R. Kohler, G. Altmann, R. G. Piotrowski (Eds.), Quantitative linguistics: an international handbook, Mouton de Gruyter, Berlin, 2005, pp. 312-324.

[17] I. Popescu, Word frequency studies, Mouton de Gruyter, Berlin–New York, 2009. doi: 10.1515/9783110218534

[18] S.M. Sukhorolska, O. I. Fedorenko, Metody linhvistychnykh doslidzhen: navch. posibnyk dlia studentiv, aspirantiv i naukovtsiv (methods of linguistic research: textbook guide for students, graduate students and researchers), Lvivskyi natsionalnyi universytet im. I.Franka, Lviv, 2009.

[19] M.B. Chatuev, A.M. Chepovskiy, Chastotnyie metodyi v kompyuternoy lingvistike (frequency methods in computational linguistics). Moskovskiy gosudarstvennyiy universitet pechati, Moskva, M, 2011.

[20] V. YU. Gudkov, E. F. Gudkova, N-grammy v lingvistike (N-grams in linguistics). Vestnik Chelyabinskogo gosudarstvennogo universiteta 24 (239) (2011) 69-71.

[21] V.Yu. Taranukha, Ispolzovanie kombinirovannykh kriteriev dlia avtomatizirovannoho opredelenyia zaimstvovaniy (using Combined Criteria for Automated Determination of Borrowings), in: «Innovatsyy v nauke»: sbornik statei po materialam XXXII mezhdunarodnoi nauchno-prakticheskoi konferentsii, SibAK, Novosibirsk, 2014, pp.15-18.

[22] M.N. Kozhyna, L.R. Duskaeva, V.A. Salimovskiy, Stilistika russkoho yazyka (stylistics of the Russian language), Nauka, Moskva, M, 2008.

[23] Yu. V. Rohushyna, Ispolzovanie kriteriev otsenki udobochitaemosti teksta dlia poiska informatii, sootvetstvuiushchei realnym potrebnostiam polzovatelia (the usage of criteria for evaluating the readability of the text to find information that meets the real needs of the user.). Problemy prohramiuvannia 3 (2007) 76-88.

[24] J. P. Zbilut, C. L. Webber Jr., Embeddings and delays as derived from quantification of recurrence plots. Physics Letters A 171 (3-4) (1992) 199–203. doi: 10.1016/0375-9601(92)90426-M

[25] N. Marwan, M. C. Romano, M. Thiel, J. Kurths, Recurrence Plots for the Analysis of Complex Systems. Physics Reports 438 (5–6) (2007) 237–329. doi: 10.1016/j.physrep.2006.11.001

[26] N. Marwan, How to avoid potential pitfalls in recurrence plot based data analysis. International Journal of Bifurcation and Chaos 21 (4) (2011) 1003–1017. doi: 10.1142/S0218127411029008

[27] V.B. Kiselev, Rekurrentnyi analiz – teoriya i praktika (recurrent analysis - theory and practice). Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnolohiy, mekhaniki i optiki 29 (2006) 118-127.

[28] P. S. Sisodia, V. Tiwari, A. Kumar, A Comparative Analysis of Remote Sensing Image Classification Techniques, in: International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, Delhi, (2014). 1418-1421. doi: 10.1109/ICACCI.2014.6968245.

[29] D. Whitley, A Genetic Algorithm Tutorial. Statistics and Computing 4 (1998) 65-85. doi: 10.1007/BF00175354.

[30] I. V. Kalinina, O. I. Lisovychenko Vykorystannia henetychnykh alhorytmiv v zadachakh optymizatsii (genetic algorithms usage in optimization problems). Adaptyvni systemy avtomatychnoho upravlinnia: mizhvidomchyi naukovo-tekhnichnyi zbirnyk. № 1(26) (2015) 48–61.

[31] S.O. Sushko, L.Ya. Fomichova, Ye.S. Barsukov, Chastoty povtoriaemosti bukv i bihramm v otkrytykh tekstakh na ukrainskom yazyke (requencies of letters recurrence and bigrams in plain texts in Ukrainian). Zakhyst informatsii 12 (3), (2010) 91-98.

[32] V.I. Shynkarenko, I.M. Demidovich Determination of the attributes of authorship of natural texts. Artificial Intelligence 3 (2018) 27-35.

[33] V. Shynkarenko, O. Kuropiatnyk, Constructive Model of the Natural Language. Acta Cybernetica 23 (4) (2018) 995–1015. doi: 10.14232/actacyb.23.4.2018.2.

[34] V. Shynkarenko, P. Ilchenko, H. Zabula, Tools of investigation of time and functional efficiency of bionic algorithms for function optimization problems. CEUR Workshop Proceedings 2139 (2018) 270 – 280.

[35] V. Moshkina, I. Andreeva, N. Yarushkinaa, Solving the problem of determining the author of text data using a combined assessment. CEUR Workshop Proceedings 2782 (2020). 112-118

[36] R. Iyer, C. Rosé, A Machine Learning Framework for Authorship Identification FromTexts. ArXiv abs/1912.10204 (2019).

[37] M. Lupei, A. Mitsa, V. Repariuk, V. Sharkan, Identification of authorship of Ukrainian-language texts of journalistic style using neural networks. Eastern-European Journal of Enterprise Technologies, 1(2), (2020) 30–36. https://doi.org/10.15587/1729-4061.2020.195041

[38] V. Lytvyn et al. Development of the Quantitative Method for Automated Text Content Authorship Attribution Based on the Statistical Analysis of N-grams Distribution. Eastern-European Journal of Enterprise Technologies, 6 (2), (2019) 28-51. doi:10.15587/1729-4061.2019.186834.