The chi-square test and the Student's t-test used for authorial style characterization

Vasyl Teslyuk[1,*,†], Iryna Khomytska[1,†], Iryna Bazylevych[2,†], Valentyna Holtvian [1,†] and Olena Durytska [1,†].

[1] *Lviv Polytechnic National University, Lviv, 79013, Ukraine*
[2] *Ivan Franko National University of Lviv, Lviv, 79000, Ukraine*

Abstract

In this research, we combine two classical statistical tests for author identification – the chi-square test and the Student's t-test. Application of these statistical tests for analysis of distribution of parts of speech is the novelty of the research. The research was conducted on the material of the belles-lettres and scientific styles. The research has proved that the chosen statistical tests give good results for determining the specificity of parts of speech distribution and phoneme distribution. The results of our research allow us to identify the style differentiating capability of each part of speech. Authors and styles are differentiated by the parts of speech which ensure statistically significant results. The calculations were carried out in Java. The structure of the developed software is based on the modular principle. The test validity of the obtained results is 95%. The results can be applied in authorship attribution.

Keywords

Chi-square test, Student's t-test, Distribution of parts of speech, Phoneme distribution, Belles-lettres style, Scientific style, Authorship attribution

1. Introduction

It is topical to offer an approach to identifying the authorial style patterns. Each authorial style consists of certain patterns typical of a certain author. The individual psychological approach to reality and the way of thinking reflect the reality differently. A variety of individual associations in each life situation creates a peculiar image of the surrounding world. To disclose the specificity of the authorial reflection of reality is the task of the researcher dealing with authorship attribution. Increasing attention to author identification can be explained by a wide practical application of the results of this research. The author is identified to establish justice in cases of slander in anonymous letters, forged documents, copyright infringement and other violations of human rights. The researchers try to reveal

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine
* Corresponding author.
† These authors contributed equally.
vasyl.m.teslyuk@lpnu.ua (V. Teslyuk); Iryna.khomytska@ukr.net (I. Khomytska); i_bazylevych@yahoo.com (I. Bazylevych); valentyna.i.holtvian@lpnu.ua (V. Holtvian); olena.durytska@lnu.edu.ua (O. Durytska)
00000-0002-5974-9310 (V. Teslyuk); 0000-0003-3470-7191 (I. Khomytska)

the patterns the author follows in the manner of writing. In most cases, researchers analyze the author's word stock, the distribution of the most frequently and the least frequently used words. However, here, we deal with the syntactic and the phonological levels. It is expedient to analyze the distribution of parts of speech and phonemes in the researched text. The difference between the authorial styles is the difference between the individual patterns used by the authors. The difference is established by various methods and techniques. The most efficient are those that ensure high level of test validity (95% – 99%). However, 95% test validity is considered classical and is applied in most cases. Powerful classical statistical tests (the Student's t-test, the chi-square test, the Lehmann-Rosenblatt test, the Wilcoxon test), allow us to obtain the results with high accuracy. The data clustering and the discriminant analysis give also good results. The statistical tests can be checked for efficiency on the phonological, lexical and syntactic levels. The reliability of the results can be enhanced by the use of several tests. The purpose of this research is to prove that the chi-square test and the Student's t-test are efficient statistical tests to differentiate texts by parts of speech distribution and phoneme distribution. The text differentiation by parts of speech distribution is a novel approach of this research.

2. Related works

The analysis of recent research has shown that the machine learning and classical methods are often applied for authorship attribution. In most cases, the content of the researched texts is emotionally colored [1]. Thus, an attempt was made to detect aggression in social media using the deep learning models. The models were tested on the Cyber-Troll dataset and gave the result – F1 score of 97% [2]. Convolutional neural networks (CNN) gave good results for author identification. The applied algorithm of this research was classical [3]. For fake news detection, the use of feature stacking gave the results of 93.39%. In the research, random forest and extra tree models were used for bagging [4, 5]. The textual semantic analysis of the Reddit statements was conducted with the help of the software toolbox LIWC-22 (Linguistic Inquiry and Word Count). On the basis of the analysis, two cognitive sub-models with linguistic psychological and social apprehension were developed [6]. The individual authorial conceptualization was characterised by the quantitative markers [7]. An intellectual analysis system aimed at determining the text authorship attribution probability for Ukrainian-language artistic works was developed [8 – 10]. For Ukrainian tweets analysis, algorithms using Levenstein distance, that is fuzz sort and fuzz set ensured good results. The best result is fingerprint similarity reaching 70% [11]. The research presented in this paper, has proved that the chi-square test and the Student's t-test are powerful statistical tests for texts differentiation by parts of speech distribution and phoneme distribution. Statistically significant results have been obtained with a high level of test validity – 95%. Consequently, the results are reliable and may be used for further research or practically applied in author identification.

3. Methods and software

3.1. The proposed combination of methods

In this research, we combine the chi-square test and the Student's t-test. The two tests were used in our previous research in different combinations: with the Lehmann-Rosenblatt test, the Wilcoxon test, the data clustering and the discriminant analysis [12, 13]. The tests were efficient in each combination. The algorithm of text differentiation in this research is the given below.

1. Choose the texts from J. K. Rowling's creation.
2. Choose the texts from K. Ashley's creation.
3. Determine the most frequently used parts of speech for each author.
4. Let the sample size be equal for the texts compared.
5. Calculate the absolute, mean and relative frequency of occurrence of parts of speech and phonemes for the two samples.
6. Use the Pearson's normality test for two samples:

$$\hat{\chi}_n^2 = \sum_{i-1}^{N} \frac{(v_i - np_i)^2}{np_i}, \quad (1)$$

where N is a number of intervals [14 – 16].

Use the Student's t-test:

$$t = (\bar{\xi} - \bar{\eta})/s\sqrt{\frac{n+m}{nm}} \geq t_{\alpha;(n+m-2)}, \quad (2)$$

where $\bar{\xi}$ and $\bar{\eta}$ are the values of mean frequencies of occurrence of parts of speech and phoneme groups for the two samples n and m [17 – 19].

Use the chj-square test:

$$X_n^2(p) = n\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{(v_{ij} - n_j v_{i.})^2}{n_j v_{i.}} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{v_{ij}^2}{n_j v_{i.}} - 1\right). \quad (3)$$

3.2. The developed software

The text differentiation program is developed on the Java programming language [20]. The structure of the program is based on the modular principle and consists of the following modules:

1. Module of data input.
2. Module of forming samples of parts of speech.
3. Module of determining the most frequently used parts of speech.
4. Module of calculating the relative frequencies of occurrence of parts of speech.
5. Module of forming samples of English phonemes.
6. Module of calculating the mean frequencies of occurrence of phonemes.
7. Module of carrying out the Pearson's test
8. Module of carrying out the Student's t-test.

9. Module of carrying out the chi-square test.
10. Module of data output.

The software has the following structure of classes: Main, SampleProcessor, PartsOfSpeechProcessor, PhonemeProcessor, PartsOfSpeechUtils, PhonemeUtils, StatisticProcessor.

The researched text files are downloaded in the class Main.

The texts are transcribed in the class SampleProcessor.

The samples of parts of speech are formed in the class PartsOfSpeechProcessor.

The samples of phonemes are formed in the class PhonemeProcessor.

The relative frequencies of occurrence of word combinations are calculated in the class PartsOfSpeechUtils.

The mean frequencies of occurrence of phonemes are calculated in the class PhonemeUtils.

The Pearson's test, the Student's t-test and the chi-square test are carried out in the class StatisticProcessor.

4. Results of the study

The applied chi-square test and Student's t-test have proved to be efficient for text differentiation. "Harry Potter and the Philosopher's Stone" by J. K. Rowling and "Sebring" by K. Ashley were differentiated with the help of the chi-square test. For text differentiation, the two texts were tagged by parts of speech (POS) in natural language processing (NLP). The tagging was done in the following way:

"CC", # Coordinating conjunction
"CD", # Cardinal number
"DT", # Determiner
"EX", # Existential there
"FW", # Foreign word
"IN", # Preposition or subordinating conjunction
"JJ", # Adjective
"JJR", # Adjective, comparative
"JJS", # Adjective, superlative
"LS", # List item marker
"MD", # Modal
"NN", # Noun, singular or mass
"NNS", # Noun, plural
"NNP", # Proper noun, singular
"NNPS",# Proper noun, plural
"PDT", # Predeterminer
"POS", # Possessive ending
"PRP", # Personal pronoun
"PRP$",# Possessive pronoun
"RB", # Adverb

"RBR", # Adverb, comparative
"RBS", # Adverb, superlative
"RP", # Particle
"SYM", # Symbol
"TO", # to
"UH", # Interjection
"VB", # Verb, base form
"VBD", # Verb, past tense
"VBG", # Verb, gerund or present participle
"VBN", # Verb, past participle
"VBP", # Verb, non-3rd person singular present
"VBZ", # Verb, 3rd person singular present
"WDT", # Wh-determiner
"WP", # Wh-pronoun
"WP$", # Possessive wh-pronoun
"WRB" # Wh-adverb

In Figure 1, we present a fragment of the tagged text "Harry Potter and the Philosopher's Stone" by J. K. Rowling

```
Model POS NLP accuracy: 0.901142771595389


Tagged text by POS NLP:
[[('J.', 'NNP'), ('K.', 'NNP'), ('Rowling', 'NNP'
), ('did', 'VBD'), ("n't", 'RB'), ('hold', 'VB'),
('useful', 'JJ'), ('as', 'IN'), ('she', 'PRP'), (
), ('.', '.')], [('They', 'PRP'), ('did', 'VBD'),
), ('be', 'VB'), ('.', '.')], [('The', 'DT'), ('D
e', 'IN'), ('that', 'DT'), ('.', '.')], [('When',
'RB'), ('happily', 'JJR'), ('as', 'IN'), ('she',
'), ('throwing', 'VB'), ('his', 'PRP$'), ('cereal
```

Figure 1: A fragment of the tagged text Harry Potter and the Philosopher's Stone" by J. K. Rowling

For calculations, the two samples were used.

For Harry Potter and the Philosopher's Stone" by J. K. Rowling:

111, 1182, 22, 0, 1350, 599, 34, 15, 9, 272, 1302, 577, 355, 5, 15, 78, 1282, 302, 849, 4, 1, 120, 0, 260, 5, 532, 942, 157, 279, 233, 113, 53, 44, 0, 82.

For "Sebring" by K. Ashley:

145, 979, 15, 0, 1159, 422, 35, 8, 0, 351, 1202, 548, 407, 1, 11, 6, 1710, 448, 759, 4, 2, 99, 0, 517, 0, 847, 971, 132, 249, 268, 146, 63, 61, 0, 95.

The application of the chi-square test has proved that the homogeneity hypothesis is rejected and the differences between the compared texts are statistically significant: por_zn=qchisq(0.95,34)

> por_zn

[1] 48.60237

The style differentiation has been carried out by the Student's t-test on the material of Show's drama and the scientific style (classical mechanics). Three cases of style differentiation were considered: 1 – any position in the word; 2 – the beginning of the word; 3 – the end of the word. Statistically significant differences were obtained in position 1 for all except for two groups of phonemes and in positions 2, 3 – for all except one group of phonemes. The results prove the Student's t-test efficiency. The data are given in Tables 1 – 3.

In Tables 1 – 6, we use such designations: GP – the group of phonemes; SD – Show's drama; SC – the scientific style (classical mechanics); L – labials; D – dorsals; C – coronals; V – velars; N – nasals; S – sonorous; F – fricatives; T – stops; S is the value of dispersion; t is the Student's statistic; $2Q$ is the level of significance; $\bar{x}$ is the mean value of frequencies of phoneme groups; $\Sigma(x_i - \bar{x})^2$ is a sum of squares of difference of the value of middle of the interval and the mean value of frequencies of phoneme groups, $\bar{x}_1 - \bar{x}_2$ is the value of difference between the researched samples.

Table 1

The results of the calculations for the comparison between Show's drama and the scientific style in an unidentified position

GP	SD $\bar{x}$	SD $\Sigma(x_i - \bar{x})^2$	SC $\bar{x}$	SC $\Sigma(x_i - \bar{x})^2$
L	121,1	2992,71	123,5	5465,75
D	390,2	9504,84	425,1	11442,71
C	18,6	1557,56	8,4	419,36
V	68,5	2707,75	60,9	2282,71
N	92,2	6098,84	87,9	3125,91
S	162,6	5351,56	186,8	2872,24
F	202,6	5055,36	211,1	4202,71
T	234,2	5611,44	220,4	6461,16

In Table 1 (continuation), we see the style differentiating capability of groups of phonemes. In the groups of dorsals, coronals, velars, sonorous and fricatives, the differences between the researched texts are statistically significant.

Table 1 (continuation)

The essential differences between Show's drama and the scientific style in an unidentified position

GP	S	t	$2Q$	$\bar{x}_1 - \bar{x}_2$
L	11,87	0,80	> 20%	Unessential
D	18,68	7,36	< 0,1%	Essential
C	5,74	7,00	< 0,1%	Essential
V	9,12	3,28	< 0,2%	Essential
N	12,40	1,37	> 10%	Unessential
S	11,71	8,14	< 0,1%	Essential
F	12,42	2,69	< 1%	Essential

In Table 2, you can see the data of a sum of squares of difference of the value of middle of the interval and the mean value of frequencies of phoneme groups for Show's drama and the scientific style in the position at the beginning of a word the end of a word.

In Table 2 (continuation), we can see the essential differences revealed in the position at the beginning of a word for the groups of labials, dorsals, coronals, velars, nasals, sonorous and stops.

In Table 3, we give the data of a sum of squares of difference of the value of middle of the interval and the mean value of frequencies of phoneme groups for Show's drama and the scientific style in the position at the end of a word.

In Table 3 (continuation), we see the style differentiating capability of the groups of labials, dorsals, velars, sonorous, fricatives and stops for the comparison of Show's drama and the scientific style in the position at the end of a word.

Table 2

The results of the calculations for the comparison between Show's drama and the scientific style at the beginning of a word

GP	SD $\bar{x}$	SD $\Sigma(x_i - \bar{x})^2$	SC $\bar{x}$	SC $\Sigma(x_i - \bar{x})^2$
L	59,1	3807,51	52,0	2711,00
D	95,8	1735,44	80,4	3005,36
C	16,4	1743,56	1,5	57,75
V	33,6	1619,16	21,0	1297,00
N	11,5	441,75	2,9	170,71
S	71,4	3009,16	30,4	687,16
F	71,6	3747,56	70,0	2138,00
T	62,2	1316,24	54,9	3164,71

Table 2 (continuation)

The essential differences between Show's drama and the scientific style at the beginning of a word

GP	S	t	$2Q$	$\bar{x}_1 - \bar{x}_2$
L	10,42	2,68	< 1%	Essential
D	8,89	6,82	< 0,1%	Essential
C	5,48	10,71	< 0,1%	Essential
V	6,97	7,12	< 0,1%	Essential
N	3,19	10,60	< 0,1%	Essential
S	7,85	20,57	< 0,1%	Essential
F	9,90	0,64	> 50%	Unessential
T	8,64	3,33	< 0,2%	Essential

Table 3

The results of the calculations for the comparison between Show's drama and the scientific style at the end of a word

GP	SD $\bar{x}$	SD $\Sigma(x_i - \bar{x})^2$	SC $\bar{x}$	SC $\Sigma(x_i - \bar{x})^2$
L	26,2	742,84	18,5	583,75
D	142,5	1247,75	125,8	2447,44
C	–	–	–	–
V	15,4	633,36	10,1	432,71
N	35,2	811,44	37,0	1889,00
S	60,7	1740,39	54,2	3338,24
F	49,3	1919,99	56,5	1423,75
T	74,4	2069,16	43,3	1220,79

Table 3 (continuation)

The essential differences between Show's drama and the scientific style at the end of a word

GP	S	t	$2Q$	$\bar{x}_1 - \bar{x}_2$
L	4,70	6,45	< 0,1%	Essential
D	7,85	8,38	< 0,1%	Essential
C				
V	4,22	4,95	< 0,1%	Essential
N	6,71	1,06	> 20%	Unessential
S	9,20	2,78	< 1%	Essential
F	7,47	3,43	< 0,2%	Essential
T	7,40	16,54	< 0,1%	Essential

The results obtained for the comparison of Show's drama and the scientific style have shown that in three cases of phoneme's position in a word the differences between the compared texts are statistically significant for almost all groups of phonemes. Consequently, the Student's t-test is efficient for solving a text differentiation task. In another comparison, we have obtained statistically significant differences between Byron's emotive prose and the scientific style. In Tables 4 – 6, we see the data for three cases of phoneme's position in a word.

Byron's emotive prose differs essentially from the scientific style in an unidentified position for the groups of labials, dorsals, nasals, sonorous and fricatives (Table 4 (continuation)).

Table 4

The results of the calculations for the comparison between Byron's emotive prose and the scientific style in an unidentified position

GP	BE $\bar{x}$	BE $\Sigma(x_i - \bar{x})^2$	SC $\bar{x}$	SC $\Sigma(x_i - \bar{x})^2$
L	139,7	3253,22	123,5	5465,75
D	402,7	7542,39	425,1	11442,71
C	6,7	313,99	8,4	419,36
V	60,9	2243,51	60,9	2282,71
N	80,0	3005,00	87,9	3125,91

S	210,4	5463,56	220,4	6461,16
F	204,2	10770,24	186,8	2872,24
T	194,1	12960,31	211,1	4202,71

Table 4 (continuation)
The essential differences between Byron's emotive prose and the scientific style in an unidentified position

GP	S	t	$2Q$	$\bar{x}_1 - \bar{x}_2$
L	12,5	5,29	< 0,1%	Essential
D	17,79	4,96	< 0,1%	Essential
C	3,50	1,91	5%	Unessential
V	8,69	0,00	100%	Unessential
N	10,11	3,08	< 0,5%	Essential
S	14,10	2,79	< 1%	Essential
F	15,05	4,54	< 0,1%	Essential

In Table 5, you can see the data of a sum of squares of difference of the value of middle of the interval and the mean value of frequencies of phoneme groups for Byron's emotive prose and the scientific style in the position at the beginning of a word.

Table 5
The results of the calculations for the comparison between Byron's emotive prose and the scientific style at the beginning of a word

GP	BE $\bar{x}$	BE $\Sigma(x_i - \bar{x})^2$	SC $\bar{x}$	SC $\Sigma(x_i - \bar{x})^2$
L	64,0	1058,00	52,0	2711,00
D	93,2	13225,44	80,4	3005,36
C	1,3	49,99	1,5	57,75
V	30,6	2169,56	21,0	1297,00
N	6,7	382,39	2,9	170,71
S	48,4	2865,56	30,4	687,16
F	33,9	10265,51	70,0	2138,00
T	56,8	1875,44	54,9	3164,71

At the beginning of a word, statistically significant differences have been obtained for the groups of labials, dorsals, velars, nasals, sonorous and fricatives (Table 5 (continuation)).

Table 5 (continuation)
The essential differences between Byron's emotive prose and the scientific style at the beginning of a word

GP	S	t	$2Q$	$\bar{x}_1 - \bar{x}_2$
L	7,93	5,96	< 0,1%	Essential

D	16,45	3,06	< 0,5%	Essential
C	1,34	0,59	50%	Unessential
V	7,60	4,97	< 0,1%	Essential
N	3,04	4,93	< 0,1%	Essential
S	7,69	9,21	< 0,1%	Essential
F	14,38	3,81	< 0,1%	Essential
T	9,17	0,82	> 20%	Unessential

In Table 6, we present the data of a sum of squares of difference of the value of middle of the interval and the mean value of frequencies of phoneme groups for Byron's emotive prose and the scientific style in the position at the end of a word.

Table 6

The results of the calculations for the comparison between Byron's emotive prose and the scientific style at the end of a word

GP	BE $\bar{x}$	BE $\Sigma(x_i - \bar{x})^2$	SC $\bar{x}$	SC $\Sigma(x_i - \bar{x})^2$
L	29,2	1353,42	18,5	583,75
D	135,2	4720,84	125,8	2447,44
C	–	–	–	–
V	8,6	407,56	10,1	432,71
N	30,7	1022,79	37,0	1889,00
S	48,0	1879,00	54,2	3338,24
F	65,0	6553,56	56,5	1423,75
T	58,1	2194,71	43,3	1220,79

Byron's emotive prose differs essentially from the scientific style in the case of the end of a word for the groups of labials, dorsals, nasals, sonorous, fricatives and stops (Table 6 (continuation)).

Table 6 (continuation)

The essential differences between Byron's emotive prose and the scientific style at the end of a word

GP	S	t	$2Q$	$\bar{x}_1 - \bar{x}_2$
L	5,68	7,41	< 0,1%	Essential
D	10,93	3,39	< 0,2%	Essential
C				
V	3,74	1,58	> 10%	Unessential
N	6,97	3,56	< 0,1%	Essential
S	9,32	2,62	< 2%	Essential
F	11,53	2,90	< 1%	Essential
T	7,54	7,72	< 0,1%	Essential

In this research, the Student's t-test is efficient for style differentiation. Statistically significant differences have been revealed in comparisons of the belles-lettres style (Show's drama; Byron's emotive prose) and the scientific style (classical mechanics) for the three cases of phoneme's position in a word.

The analysis of the results obtained by the chi-square test in this research, has shown that this test is efficient for authorship attribution on the syntactic level. The Student's t-test has given good results on the phonological level for style differentiation. The results have been obtained with the test validity of 95%.

5. Discussions

The chi-square test in this research has been used on the syntactic level for author identification. In our previous research, we used the test on the phonological and lexical-semantic levels [12, 13]. The test was efficient on these levels. In this paper, we have proved efficiency of the chi-square test on the syntactic level. Consequently, the chi-square test ensures reliable data (the level of test validity – 95%) on the phonological, lexical-semantic and syntactic levels.

The Student's t-test in this research has been used for style differentiation. The results of testing have shown statistically significant differences between the belles-lettres style (Shaw's drama, Byron's emotive prose) and the scientific style (classical mechanics). The level of test validity is 95%.

According to the analysis of similar research, the authorial style was identified by deep learning models in an attempt to detect aggression in social media. The models were tested on the Cyber-Troll dataset and ensured the result – F1 score of 97% [2]. In another research, the random forest and extra tree models were used for fake news detection. The use of feature stacking gave the results of 93.39%. [4, 5]. The algorithms, using Levenstein distance for Ukrainian tweets analysis, ensured reliable results. The best result is fingerprint similarity – 70% [9].

Having analyzed the results obtained in our research with the help of the chi-square test and the Student's t-test, we can state that this combination of tests is efficient for style differentiation and author identification on three language levels: the phonological, lexical-semantic and syntactic. As the test validity of the results is high – 95%, it is recommended to apply this combination of tests for solving the tasks of authorship attribution.

Conclusions

It is topical in modern research to propose a new approach to the authorial style identification. The novelty of the research is an application of the chi-square test and for analysis of distribution of parts of speech on the material of American emotive prose.

The chi-square test was performed on the material of the belles-lettres style ("Harry Potter and the Philosopher's Stone" by J. K. Rowling and "Sebring" by K. Ashley). For text differentiation, the two texts were tagged by parts of speech (POS) in natural language processing (NLP). The task of an authorial style differentiation has been solved with a level of test validity of 95%.

The Student's t-test was performed on the material of the belles-lettres style (Show's drama, Byron's emotive prose) and the scientific style (classical mechanics). Statistically significant differences were obtained in three cases of style differentiation: 1 – any position in the word; 2 – the beginning of the word; 3 – the end of the word. The style differentiating capability of phoneme groups (labials, dorsals, coronals, velars, nasals, sonorous, fricatives and stops) was revealed in position 1 for all except for two groups of phonemes and in positions 2, 3 for all except one group of phonemes. The results prove the Student's t-test efficiency. The calculations were carried out in Java. The structure of the developed software is based on the modular principle. The test validity of the obtained results is 95%.

The goal of this research has been attained. The research has proved that the chi-square test and the Student's t-test are efficient statistical tests to differentiate texts by parts of speech distribution and phoneme distribution.

The practical application of this research involves the author identification and style differentiation. In our future research, we will choose some other syntactic features for authorial styles differentiation.

References

[1] P. Hajibabaee, M. Malekzadeh, M. Ahmadi, M. Heidari, A. Esmaeilzadeh, R. Abdolazimi, J. H. J. Jones, Offensive language detection on social media based on text classification, in: Proceedings of the IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0092-0098, doi: 10.1109/CCWC54503.2022.9720804.

[2] U. Khan, S. Khan, A. Rizwan, G. Atteai, M. M. Jamjoom, N. A. Samee, Aggression detection in social media from textual data using deep learning models, Applied Sciences 12(10) 5083 (2022). doi:10.3390/app12105083.

[3] F. Mohades Delami, H. Sadr, M. Nazari, Using machine learnjng-based models for personality recognition, Big Data and Computing Visions 1(3) (2022) 128-139. doi:10.22105/bdcv.2021.142588.

[4] N. Lina, S. Fua, S. Jianga, Fake news detection in the Urdu language using CharCNN-RoBERTa, CEUR Workshop Proceedings, vol. 2826/T3-2, 2020.

[5] M. Shoaib Farooq, A. Naseem, F. Rustam, I. Ashraf, Fake news detection in the Urdu language using machine learning, PeerJ Computer Science 9:e1353 (2023). doi: 10.7717/peerj-cs.1353.

[6] S. Albota, Creating a model of war and pandemic apprehension: textual semantic analysis, in proceedings of the 7th International conference on computational linguistics and intelligent systems. Vol. II: Computational linguistics workshop. Kharkiv, Ukraine, April 20-21, 2023, pp. 228–243.

[7] O. Levchenko, M. Dilai, Qualitative and Quantitative Markers of Individual Authorial Conceptualization, in proceedings of the 7th International conference on computational linguistics and intelligent systems. Vol. II: Computational linguistics workshop. Kharkiv, Ukraine, April 20-21, 3396, 2023, pp. 1-19.

[8] R. Romanchuk, V. Vysotska, V. Andrunyk, L. Chyrun, S. Chyrun, O. Brodyak, Intellectual Analysis System Project for Ukrainian-language Artistic Works to Determine the Text

Authorship Attribution Probability, in: Proceedings of the 18th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2023, Lviv, Ukraine, 19-21 October, 2023, Doi:10.1109/CSIT61576.2023.10324012.

[9] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, vol. 2125, 2018, pp. 1–25.

[10] Hou, R., & Huang, C.-R. (2020). Robust stylometric analysis and author attribution based on tones and rimes. Natural Language Engineering 26(1) 2020 49–71. doi:10.1017/S135132491900010X.

[11] O. Prokipchuk, V. Vysotska, Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning. Radio Electronics, Computer Science, Control 2 (2023) 103. doi: 10.15588/1607-3274-2023-2-11.

[12] I. Khomytska, V. Teslyuk, K. Prysyazhnyk, N. Hrytsiv, The Lehmann-Rosenblatt test applied for determination of statistical parameters of Charles Dickens's authorial style, in Proceedings of IEEE XVIth Scientific and Technical Conference on Computer Science and Information Technologies. CSIT 2021, Lviv, Ukraine, 22–25 September, vol. 2, 2021, pp. 64–67. doi:10.1109/CSIT52700.2021.9648789.

[13] I. Khomytska, V. Teslyuk, I. Bazylevych, Yu. Kordiiaka, Machine learning and classical methods combined for text differentiation, in Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems. Vol. I: Main Conference, Gliwice, Poland, May 12-13, CEUR Workshop Proceedings, vol. 3171, 2022, pp 1107-1116.

[14] Th. S. Gries, Statistics for Linguistics with R: A Practical Introduction (Trends in Linguistics: Studies & Monographs), Mouton de Gruyter, 2009, p. 348.

[15] R. Bhattacharya, E. C Waymire, A Basic Course in Probability Theory (2nd ed.), Springer, 2016 edition, February 16, 2017.

[16] V. S. Perebyjnis, Statystychni metody dlia lingvistiv, Nova Knyha, Vinnytsia, Ukraine, 2013.

[17] P. C. Gomez, Statistical Methods in Language and Linguistic Research. University of Murcia, Spain, 2013.

[18] A. Kornai, Mathematical Linguistics, Springer, 2008.

[19] V. M. Turchyn, Matematychna statystyka, Navch. Posib., Vydavnychyj tsentr “Akademia”, Kyiv, Ukraine, 1999.

[20] A. Batyuk, V. Voityshyn, V. Verhun, Software Architecture Design of the Real-Time Processes Monitoring Platform, in: Proceedings of the IEEE Second International Conference on Data Stream Mining & Processing, DSMP 2018, Lviv, Ukraine, 2018, pp. 98-101. doi: 10.1109/DSMP.2018.8478589.