Automated Identification of Authorial Styles

Iryna Khomytska [1], Vasyl Teslyuk [1], Iryna Bazylevych [2] and Iryna Karamysheva[1]

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

Abstract

The problem of improvement of the software for authorial style identification is topical today and requires new approaches. The proposed approach consists in the use of the efficient classical and machine learning methods which ensure reliable data with a test validity of 95%. These are the following methods: the chi-square test and the discriminant analysis. The methods have been applied on the level of letters of Cyrillic alphabet which proved to be appropriate for an author identification task. Typical statistical characteristics have been established for Ukrainian authors' styles. With the help of these characteristics, the author can be identified. The proposed structure of the software system is the novelty of the research. The developed software system is based on a modular principle. The algorithm of the author's identification is realized on the Python programming language. The level of automation is high.

Keywords:

Chi-square test, Discriminant analysis, Ukrainian authors' styles, Author identification, Modular principle of software system.

1. Introduction

The information technologies for text differentiation and author identification have been widely used recently. These technologies are aimed at establishing the authorial characteristics typical of a certain author. However, the authorial features never occur alone, separated from the other text features. Different text features related to a functional style, genre and topic are combined and cause the complexity of authorship attribution. The problem of separation of the authorial features lies at the crux of the author identification. For every particular text, a certain part of vocabulary is typical of a certain topic and can occur in a text of any author. This vocabulary cannot identify a particular author. Therefore, some specific layer of vocabulary should be identified. If the author's specificity is clearely expressed, that is an easy case of characterizing the author. If the author's distinctive features are minimal, it is hard to draw a demarcation line between the general text specificity (functional style, genre) and the authorial specificity. In any case, the vocabulary characteristics of a certain style and topic should be identified as a preparatory stage of the author identification. For the purpose of characterizing the specificity of vocabulary of a certain topic, frequency dictionaries can be compiled. Such dictionaries list the most commonly used words for a particular sphere of communication. The authorial specific vocabulary can be separated from the layer of commonly used words. However in documents and formal papers, the standards and formalities prevail over a free expression of a thought. Consequently, the authorial features can hardly be noticed. In this case, the text features should be thoroughly studied and viewed from all possible sides. In our research, to avoid the ambiguity caused by the mentioned difficulties, we compare the texts with clearly expressed authorial specificity. These are the texts from emotive prose which is rich in specific expressive means. Expressive and emotional specificity of authorial styles is reflected in frequency of occurrence of language units. The texts from emotive prose by Ukrainian writers are researched in this paper. The developed program system uses the statistical tests (the chi-square test and the discriminant analysis) which are the most appropriate for the task of authorship attribution on the chosen language level (letters of Cyrillic alphabet, stop words,

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine

EMAIL: Iryna.khomytska@ukr.net (I. Khomytska); vasyl.m.teslyuk@lpnu.ua (V. Teslyuk); i_bazylevych@yahoo.com (I. Bazylevych); iryna.d.karamysheva@lpnu.ua (I. Karamysheva)
ORCID: 0000-0003-3470-7191 (I. Khomytska); 0000-0002-5974-9310 (V. Teslyuk).

punctuation marks, spaces). The purpose of the research is to prove that the chi-square test and the discriminant analysis combined are efficient for the task of author identification. The novelty of the research is the proposed structure of the program system based on a modular principle and a combination of the chi-square test and the discriminant analysis applied in the Ukrainian language.

2. Related Works

The difficulties related to separation the authorial specific features from the other language features make the author identification a hard problem to solve. Different approaches have been tried and a lot of methods have been applied over the last decades. The problem has been studied on nearly all language levels. Nevertheless, a perfect solution has not been found yet and the problem is still topical.

In recent research [1 – 4], the machine learning methods were applied to recognize the author of a given text. The use of the classification algorithms ensured obtaining the acknowledgement text, for some classifiers, with an accuracy of 92%. The authors were deduced in the Portuguese language. The extracted stylometric features (text relevant attributes) suggested that the applied technique was effective to distinguish the author or the ghost writer of a given text [1]. In our research, a classical approach is used. The significance level is 0.05 and all the results have been obtained with the test validity of 95%.

The approaches to authorship attribution comprise the attempts to find the best solutions of separating the distinctive authorial style features from the rest of the characteristics of the researched text. Among the most efficient approaches are the following: text distortion for identifying the distinctive features of the authorial style [5]; leveraging the discourse information [6]; the use of orthogonal similarity relations [7]; the use of topic models [8]. Stylistic features of poetry and other styles are often determined on the basis of stylometric analysis [9 – 11].

Authentication of misinformation generated by some dubious sources is a task of great importance. This task was approached with the following machine learning methods: logistic regression and naive Bayes algorithms [12]. In the pre-processing phase, stop words and punctuation marks were removed. The texts were tokenized and stemmed. This way, certain specific to Twitter features were extracted. The highest precision of 91.1% was obtained using the method of logistic regression [12]. The method applied in this paper involves the use of the chi-square test and the discriminant analysis. These two tests have been applied for characterizing the authorial styles of Ukrainian writers. The two tests ensure higher precision than the method of logistic regression.

A similarity metric was used to compare the pieces of a text with the most relevant words [13]. According to this approach, the words, corresponding to the nodes, were to be taken into account in order to enhance representation of a text with complex networks. The applied method involved constructing co-occurrence network for a text, obtaining dissimilarity matrices, joining them and analyzing the obtained data with a standard supervised learning algorithm. In most cases, the precision rates were above 90% and the maximum value was 98.75% [13]. Our research ensures a classical level of accuracy (95%). Another attempt to apply the neural network method was made for a sentiment analysis in English newspapers. The method was used as a public opinion influences identification tool [14]. This approach may also be used in an emotion recognition system project of English newspapers.

A quantitative approach was used in a textual semantic analysis to highlight some important issues of semantics [15]. Quantitative parameters of linking words in political speeches of Bill Clinton were analyzed using Python [16].

For solving the task of authorship attribution in Arabic tweets, the support vector machine as a supervised learning algorithm was used for classification of relevant text features. The Bag of Words (BOW) approach proved to be efficient. The performance of different classifiers was tested. Different feature sets were created and combined. The combination of feature sets improved the results [17]. As it has been proved in our research, it is recommended to combine a machine learning method with a classical one, as the latter gives more reliable results.

The random forest approach, using WEKA 3.8 tool, was tested authenticating Arabic poems. This approach was chosen because of a higher accuracy for decision trees. The dataset was tested on the basis of twelve features. The overall precision was 76.4%. The research included four stages: data collecting, data cleansing, feature extracting and classifying. The method was applied on the level of letters, words and word length [18]. The twelve linguistic features chosen for the research is quite a sufficient number and the results would be higher if some powerful classical statistical methods were combined with the NLP methods.

The analysis of the related works has shown that, in most cases, for authorship attribution, the machine learning methods give results with an accuracy of 70% – 90%. A similarity metric, that involves constructing a co-occurrence network for a text, ensures a higher accuracy – up to 98.75% [13]. In our research we apply a combination of classical (the chi-square test) and machine learning (the discriminant analysis) methods. The chosen classical level of significance of 5% makes it possible to obtain the results with a precision of 95%.

3. Methods and Software

3.1. The Proposed Combination of Methods

The mathematical support of this research is based on two statistical tests: one – a powerful classical method – the chi-square test and a machine learning method – the discriminant analysis. These two methods combined ensure reliability of the results with the first mentioned method, and simplicity in use with the second method. Both methods have proved to be efficient on the level of letters of Cyrillic alphabet. The methods have been tested on the material of texts from Ukrainian emotive prose. The algorithm for the chi-square test is the following:

1. Prepare samples of 51000 letters for the comparison.
2. Form portions of 1000 letters for the samples that are to be compared.
3. Obtain the values of frequencies of occurrence of letters in each portion.
4. Obtain the values of frequencies of occurrence of letters in each sample.
5. Obtain the values of relative frequencies of occurrence of letters in each portion.
6. Obtain the values of relative frequencies of occurrence of letters in each sample.
7. Use the chi-square test for two compared samples [19, 20].
8. Analyze the obtained results.

Using the chi-square test, we verify the H_0 hypothesis: the observations are done with the same variable. We use the statistics:

$$X_n^2(p)=\sum_{i=1}^{s}\sum_{j=1}^{k}(\nu_{ij}-n_j p_i)^2\big/(n_j p_i)\,. \tag{1}$$

To estimate the unknown parameters $p_1,\dots,p_s$, we use the maximum likelihood method:

$$L(p)=c\prod_{i,j}p_i^{\nu_{ij}}=c\prod_{i}p_i^{\nu_{i.}},\ \nu_{i.}=\sum_{j=1}^{k}\nu_{ij}\,. \tag{2}$$

The method of indefinite Lagrange factor is employed to obtain estimates $\hat{p}$ of parameters p_i : $\hat{p}_i=\frac{\nu_{i.}}{n},\ i=1,\dots,s$, where $n=n_1+\dots+n_k=\sum_{i,j}\nu_{ij}$ is the total of the observations. As a result, the formula for the criterion statistics is the following [21, 22]:

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

The H_0 hypothesis is rejected if the value t of the statistics (3) satisfies the inequality $t \geq \chi^2_{1-\alpha,(s-1)(k-1)}$.

The next step will be the algorithm for the discriminant analysis:

9. Obtain the mean values of frequencies of occurrence of letters, stop words, punctuation marks and spaces for each sample.
10. Construct the vectors.
11. Write the regression equations for the obtained data.
12. Obtain the coefficients for the regression equations.
13. Employ the formula of Mahalanobis distance [23]:

$$D^2(x/G_k) = (n-g)(x-\bar{x}_k)^T W^{-1}(x-\bar{x}_k), k = 1,...,g, \qquad (4)$$

where G_k stands for a set of authors, x stands for an object having p variables, n is a number of the researched literary works, g is a number of the chosen authors, W^{-1} stands for an inverse covariance matrix, $\bar{x}_k$ stands for the vector of the values of the mean for the variables from *k-th* group of the objects.

3.2. The Developed Software

A topical issue of computer linguistics is the development of information technologies for an automated identification of the authorial style. Every automated information system characterized by a certain technology is aimed at transforming the input data into some expected information. Therefore, the information technology structuring involves the development of a classification and a coding system, an organization of collecting and transferring information and different methods to access the data [24]. The developed information technology processes the researched texts by different authors using chosen statistical methods. The obtained data are statistical characteristics typical of a certain authorial style. These data form the author's statistical parameters. Python programming language has been used for automated identification of the authorial style. The developed structure of information technology for automated authorship attribution is shown in Figure 1.

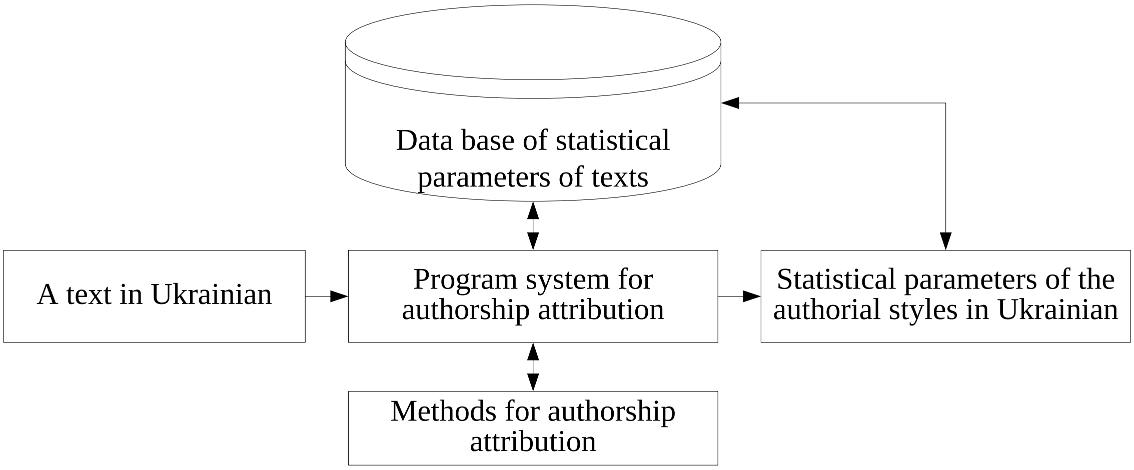

Data base of statistical parameters of texts
A text in Ukrainian
Program system for authorship attribution
Statistical parameters of the authorial styles in Ukrainian
Methods for authorship attribution

Figure 1: The structure of information technology for automated authorship attribution

The automated author identification has been done using Python. The algorithm of performing the chosen mathematical tests involves standard functions and libraries of Python. The tools of Python were used for work with different protocols. The process of automated identification of the authorial style consists of two main stages: the first stage is preparatory before the statistical calculations, and the second stage is the stage of statistical calculations. On the first stage, we make the following changes: all the letters in uppercase are changed into the letters in lowercase, only one space is left between the words, a space is put at the beginning of a text. Then, we sort the linguistic units. For calculations, we have chosen letters, stop words, punctuation marks and spaces. Differentiation of authorial styles and author identification is done using the chi-square test and the discriminant analysis.

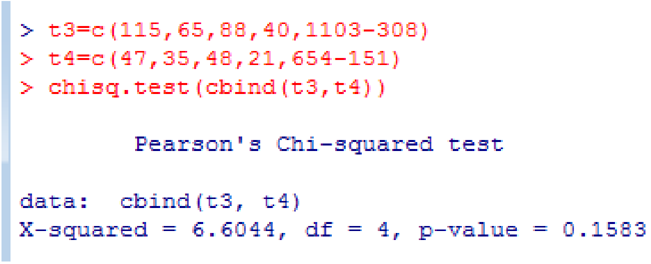
The algorithm of the software system functioning includes the following steps: text files uploading, sample formation, sample division into portions, calculations of frequencies of occurrence of linguistic units in each portion and sample, application of the chi-square test and the discriminant analysis and analysis of the data obtained. The algorithm is presented in Figure 2.

The developed structure of the software system for author identification is shown in Figure 3. The program is based on a modular principle. The main modules are: a module of file opening, a module of sample setting, a module of text analysis, applying the chi-square test and the discriminant analysis, a module of results visualization, a module of data storing.

The module “data storing” gives an access to data base. The module “interface” ensures a connection between the user and the software system. The interface is written with the help of library PyQt5 which has Qt Designer. One of the biggest classes of PyQt5 is Widgets having tables, lists and other means of results visualization. Quick visualization is ensured by a considerable level of NumPy and Qt Qraphics View Framework. For more efficient work, a module MainWindowUI has been developed. It shows the program main window, imports all next modules and a file containing a code of the program interface. Modules StatisticLetters, StatisticStopWords, StatisticPunctuationMarks, StatisticSpaces are involved in the statistical analysis of letters, stop words, punctuation marks and spaces. Module StatisticTests is responsible for texts differentiation and author identification by the chi-square test and the discriminant analysis. Module re is responsible for editing a text before processing. Module GraphCreate has functions of graphical presentation of the obtained data. In every tab of the interface, there are options for building the graphs that show the results of the statistical analysis [24].

4. Results of the Study

The authorship attribution has been done on the material of Ukrainian emotive prose. The statistical parameters of the literary works by I. Franko, O. Honchar and L. Hlibov have been obtained by the chi-square test and the discriminant analysis. The chi-square test has been performed on the level of letters of Cyrillic alphabet. The relative frequencies of occurrence of letters have been calculated as a stage of the algorithm of the chi-square test. The highest values of the relative frequencies of occurrence of letters for the literary works by L. Hlibov (Text 1, Text 2) are given in Table 1. The values show that letters A, O, B are the most frequently used in this sample. The results of the chi-square test (given below) confirm that the literary works by L. Hlibov are written by one author as the homogeneity hypothesis is not rejected. This proves that the chi-square test is powerful for identifying authorial styles.



```
> t3=c(115,65,88,40,1103-308)
> t4=c(47,35,48,21,654-151)
> chisq.test(cbind(t3,t4))

        Pearson's Chi-squared test

data:  cbind(t3, t4)
X-squared = 6.6044, df = 4, p-value = 0.1583
```

Application of the method of discriminant analysis allowed us to identify the authorial styles by Franko, O. Honchar and L. Hlibov. The method is efficient on the level of letters, stop words, punctuation marks and spaces. The results of calculations of the average number of the mentioned linguistic units are presented in Figure 4. The analysis of the average number of words and punctuation marks shows that the number of punctuation marks may be relatively greater if the number of words is not much greater. This is the case with L. Hlibov’s literary works: for the average number of words of

34.46, the average number of punctuation marks is 6.53 (in one literary work), and for the average number of words of 37.21, the average number of punctuation marks is 10.07 (in another literary work). This may be considered a characteristic feature of L. Hlibov's manner of writing.

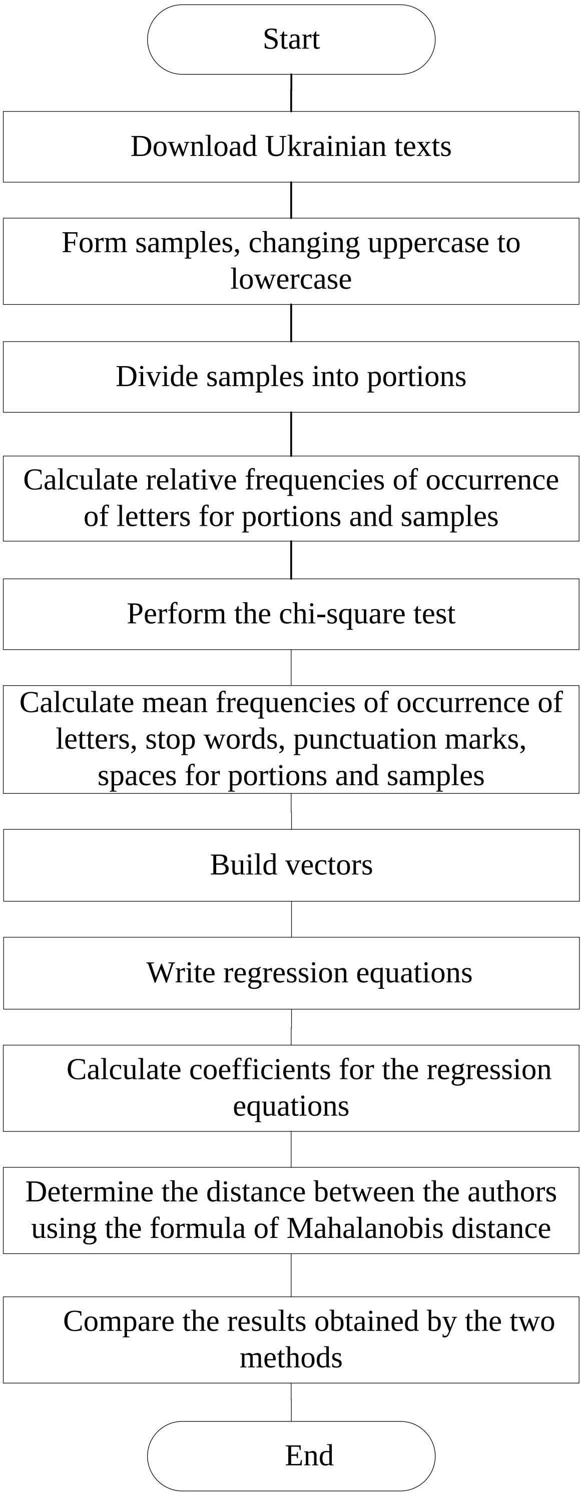

Start
Download Ukrainian texts
Form samples, changing uppercase to lowercase
Divide samples into portions
Calculate relative frequencies of occurrence of letters for portions and samples
Perform the chi-square test
Calculate mean frequencies of occurrence of letters, stop words, punctuation marks, spaces for portions and samples
Build vectors
Write regression equations
Calculate coefficients for the regression equations
Determine the distance between the authors using the formula of Mahalanobis distance
Compare the results obtained by the two methods
End

Figure 2: A flow chart of the algorithm for author identification by the chi-suare test and the discriminant analysis

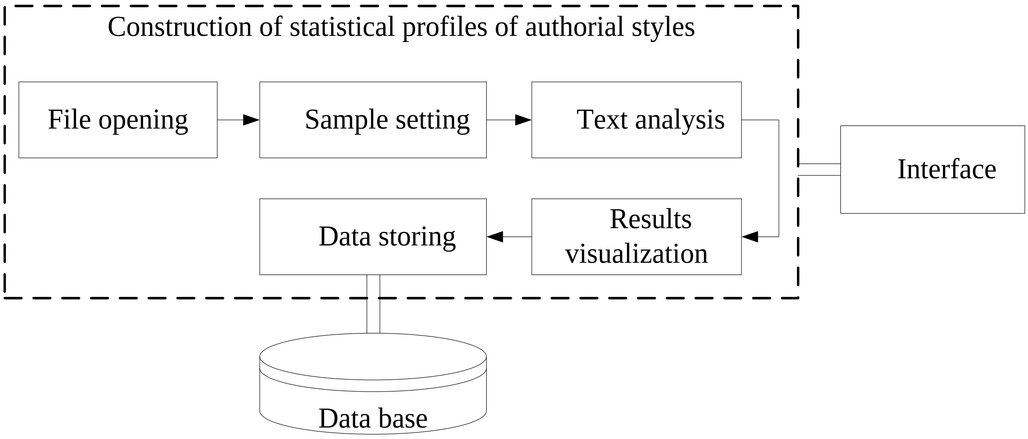


Figure 3: Graphical presentation of the structure diagram of the software system for author identification

In Figure 5, we see the results of the discriminant analysis determined by the squared Mahalanobis distances. The distances between the literary works of the same author are small if compared with the distances between the literary works of different authors. For the literary works by I. Franko, we have obtained the distances: 3.84, 1.64, 3.05, 3.09; for the works by O. Honchar: 0.60, 3.04, 2.36, 4.83; for the works by L. Hlibov: 2.84, 4.51, 2.78, 3.37. If we compare the literary works by I. Franko and l. Hlibov, the distance is much greater – 89.51. Consequently, the discriminant analysis is an efficient method for authorship attribution.

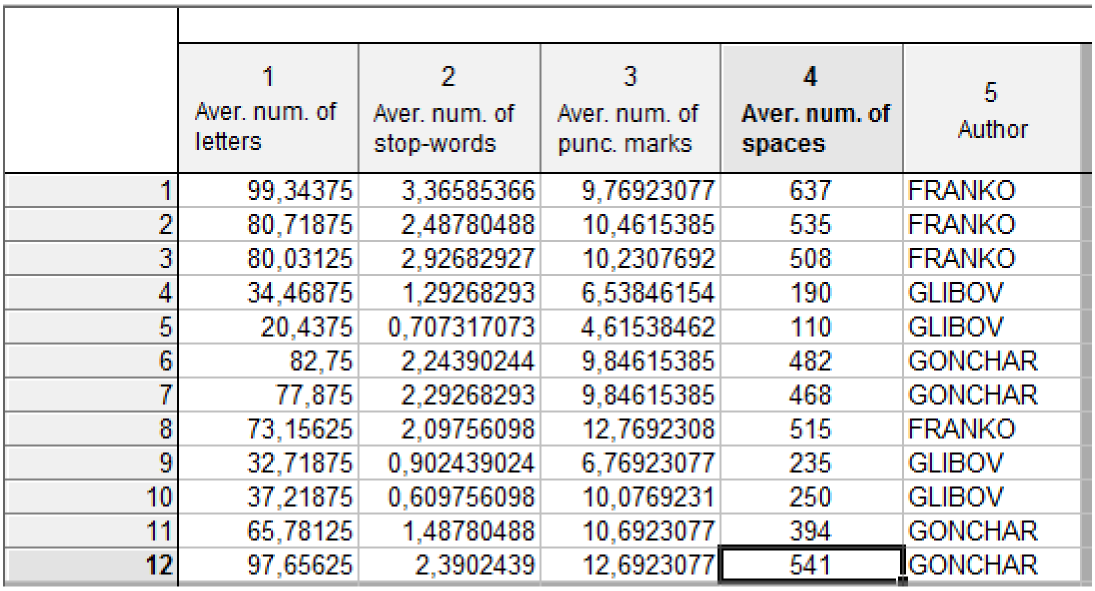
Table 1
The highest values of relative frequencies of occurrence of letters

Letters	Text 1	Text 2
А	10%	7%
В	6%	8%
Е	5%	8%
И	6%	5%
І	5%	5%
К	3%	5%
Л	5%	3%
Н	5%	5%
О	8%	7%
П	4%	3%
Р	4%	4%
С	4%	4%
Т	6%	5%
У	3%	5%

The analysis of the obtained results show that the chi-square test combined with the discriminnt analysis is an efficient combination for characterizing the authorial styles and performing the author identification.

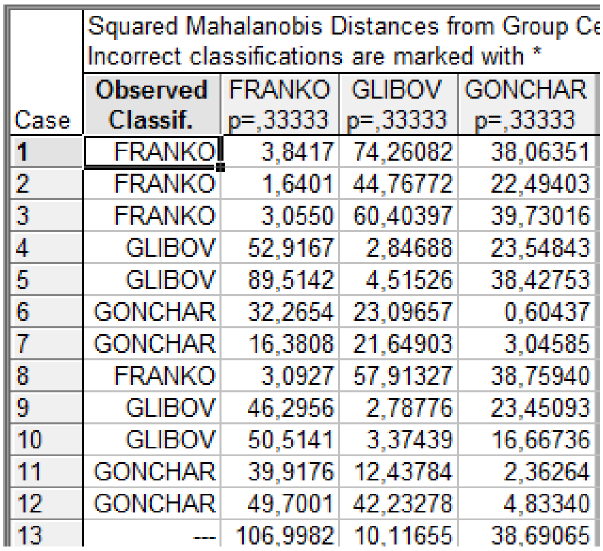
5. Discussions

This research is a continuation of testing the classical and machine leaning methods for efficiency in authorship attribution. In our earlier research, the statistical tests were tested on different language levels (phonological, lexical, syntactic) in two languages – English and Ukrainian. In a comparison to our previous research, we can state that the combination of the classical statistical method – the chi-square test and the machine learning method – the discriminant analysis is efficient for author identification in the Ukrainian language.



	1 Aver. num. of letters	2 Aver. num. of stop-words	3 Aver. num. of punc. marks	4 Aver. num. of spaces	5 Author
1	99,34375	3,36585366	9,76923077	637	FRANKO
2	80,71875	2,48780488	10,4615385	535	FRANKO
3	80,03125	2,92682927	10,2307692	508	FRANKO
4	34,46875	1,29268293	6,53846154	190	GLIBOV
5	20,4375	0,707317073	4,61538462	110	GLIBOV
6	82,75	2,24390244	9,84615385	482	GONCHAR
7	77,875	2,29268293	9,84615385	468	GONCHAR
8	73,15625	2,09756098	12,7692308	515	FRANKO
9	32,71875	0,902439024	6,76923077	235	GLIBOV
10	37,21875	0,609756098	10,0769231	250	GLIBOV
11	65,78125	1,48780488	10,6923077	394	GONCHAR
12	97,65625	2,3902439	12,6923077	541	GONCHAR

Figure 4: The average numbers of letters, stop words, punctuation marks and spaces



Squared Mahalanobis Distances from Group Ce
Incorrect classifications are marked with *

Case	Observed Classif.	FRANKO p=,33333	GLIBOV p=,33333	GONCHAR p=,33333
1	FRANKO	3,8417	74,26082	38,06351
2	FRANKO	1,6401	44,76772	22,49403
3	FRANKO	3,0550	60,40397	39,73016
4	GLIBOV	52,9167	2,84688	23,54843
5	GLIBOV	89,5142	4,51526	38,42753
6	GONCHAR	32,2654	23,09657	0,60437
7	GONCHAR	16,3808	21,64903	3,04585
8	FRANKO	3,0927	57,91327	38,75940
9	GLIBOV	46,2956	2,78776	23,45093
10	GLIBOV	50,5141	3,37439	16,66736
11	GONCHAR	39,9176	12,43784	2,36264
12	GONCHAR	49,7001	42,23278	4,83340
13	---	106,9982	10,11655	38,69065

Figure 5: The distances between literary works by the researched authors

In our earlier research, we tested the chi-square test in a combination with the other classical methods – the Student's t-test and the Kolmogorov-Smirnov test. In this combination, it was more powerful than the Student's t-test, but less powerful than the Kolmogorov-Smirnov test on the phonological level. In this research, the chi-square test is applied on the levels of letters and words showing good results.

The previously applied classical methods – the Lehmann-Rosenblatt test and the Wilcoxon test were tested on the levels of phonemes and word length. These methods were less powerful than the chi-square test. For the mentioned methods, the level of test validity was 95%.

The machine learning methods – the data clustering method and the method of discriminant analysis were previously tested in a combination with the chi-square test and the Student's t-test on the levels of words and phonemes. In this combination, the classical methods were more powerful [25].

According to the results of our earlier research, the method of discriminant analysis is more powerful than the method of data clustering. In this research, the method of discriminant analysis has shown good results obtained with the help of the squared Mahalanobis distances. The distances between the researched literary works by one author are small. This proves that the works have similar linguistic characteristics, typical of a certain authorial style. The distances for the works by I. Franko are: 3.84, 1.64, 3.05, 3.09; for the works by O. Honchar: 0.60, 3.04, 2.36, 4.83; for the works by L. Hlibov: 2.84, 4.51, 2.78, 3.37. Consequently, the method of discriminant analysis is rightly chosen for the language levels of letters and words, as it has given good results and solved the task of author identification.

In our research, we have developed a structure of the software system for author identification. The program is based on a modular principle, which allows us to quickly modify the program. The structure of the software system includes the following modules: a module of file opening, a module of sample setting, a module of text analysis, applying the chi-square test and the discriminant analysis, a module of results visualization, a module of data storing.

To improve the efficiency of the work, a module MainWindowUI has been developed. It imports all next modules and a file containing a code of the program interface. For the statistical analysis of letters, stop words, punctuation marks and spaces, the modules StatisticLetters, StatisticStopWords, StatisticPunctuationMarks, StatisticSpaces have been developed. The module StatisticTests is used for texts differentiation and author identification by the chi-square test and the discriminant analysis. The developed software system ensures quick and efficient work. The data obtained are reliable and can be used in our further research.

We consider it to be expedient to use the combination of the chi-square test and the method of discriminant analysis for authorship attribution on other language levels and in other languages. The level of test validity for the chi-square test is high – 95%. It is recommended to apply this test in a combination with other machine learning methods.

6. Conclusions

The purpose of the research has been achieved – the efficiency of the the chi-square test and the discriminant analysis has been proved on the levels of letters, stop words, punctuation marks and spaces in Ukrainian. The novel approach of the research consists in application of the developed structure of the software system based on a modular principle. The modular principle allows us to quickly modify the program system. The module StatisticTests based on the use of the chi-square test and the discriminant analysis has been applied for texts differentiation and author identification.

The results obtained by the chi-square homogeneity test with a test validity of 95%, show that the authorship of I. Franko, O. Honchar and L. Hlibov has been established for the researched literary works. For the comparisons of the literary works by each of the mentioned Ukrainian writers, the homogeneity hypothesis has not been rejected. This means for each author that the literary works were written by the same author. Therefore, for author identification, it is expedient to use the chi-square test, either alone or in a combination with other classical or machine learning methods. The combination of the chi-square test with the discriminant analysis in this research has given good results.

For the discriminant analysis, using the squared Mahalanobis distances, we have obtained the distances between the literary works by I. Franko, O. Honchar and L. Hlibov. The distances are small: for the literary works by I. Franko, the distances are: 3.84, 1.64, 3.05, 3.09; for the works by O. Honchar – 0.60, 3.04, 2.36, 4.83 and for the works by L. Hlibov – 2.84, 4.51, 2.78, 3.37. The established small distances testify that the researched literary works reflect the same authorial style, the linguistic features of the same manner of writing. Consequently, the discriminant analysis is an efficient method for author identification.

In the developed software system, standard functions and libraries of Python were used in the algorithm of performing the chosen mathematical tests. The Python tools were employed for different protocols. Two main stages of the process of automated identification of the authorial style included: the preparatory stage before the statistical calculations, and the stage of the statistical calculations. The following changes were made on the first stage: all the letters in uppercase were changed into the letters in lowercase, only one space was left between the words, a space was put at the beginning of a text. Then, all the linguistic units were sorted. The letters, stop words, punctuation marks and spaces were calculated. The chi-square test and the discriminant analysis were performed for differentiation of the authorial styles and the author identification.

In the algorithm of the software system functioning, there are the following steps: text files uploading, sample formation, sample division into portions, calculations of frequencies of occurrence of linguistic units in each portion and sample, application of the chi-square test and the discriminant analysis and analysis of the data obtained.

The structure of the developed software system includes a module MainWindowUI which shows the program main window, imports all next modules and a file containing a code of the program interface. Modules StatisticLetters, StatisticStopWords, StatisticPunctuationMarks, StatisticSpaces are responsible for the statistical analysis of letters, stop words, punctuation marks and spaces. Module StatisticTests is involved in the texts differentiation and the author identification by the chi-square test and the discriminant analysis.

The obtained results can be used in our future research aimed at testing statistical methods or their combinations for their efficiency in the author identification.

7. References

[1] M. A. da Rocha, P. S. G. de Morais, D. M. da Silva Barros, J. P. Q. dos Santos, S. Dias-Trindade, R. A. de Medeiros Valentim, A text as unique as a fingerprint: Text analysis and authorship recognition in a Virtual Learning Environment of the Unified Health System in Brazil. In: Expert Systems with Applications: An International Journal Volume 203 Issue COct 2022 https://doi.org/10.1016/j.eswa.2022.117280. (2022).

[2] 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, pp. 1–25. (2018).

[3] L. Muttenthaler, G. Lucas, J. Amann, Authorship Attribution in Fan-Fictional Texts Given Variable Length Character and Word N-Grams, Notebook for PAN at CLEF 2019. 9-12 September 2019, Lugano, Switzerland, vol. 2380. Paper 49. (2019).

[4] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Shared Tasks on Authorship Analysis at PAN 2020. In book: Advances in Information Retrieval, 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II, pp. 508–516. (2020) DOI: 10.1007/978-3-030-45442-5_66. (2020).

[5] E. Stamatatos, Authorship attribution using text distortion, in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1, 2017, pp. 1138–1149. (2017).

[6] E. Ferracane, S. Wang, R. Mooney, Leveraging discourse information effectively for authorship attribution, in Proceedings of the Eighth International Joint Conference on Natural Language Processing, vol. 1, 2017, pp. 584–593. (2017).

[7] U. Sapkota, T. Solorio, M. Montes-y Gomez, P. Rosso, The use of orthogonal similarity relations in the prediction of authorship, in Computational Linguistics and Intelligent Text Processing. Springer, 2013, pp. 463–475. (2013).

[8] Y. Seroussi, I. Zukerman, F. Bohnert, Authorship attribution with topic models, Computational Linguistics, vol. 40, no. 2, pp. 269–310, 2014. (2014).

[9] K. Sundararajan, D. Woodard, What represents style in authorship attribution? in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2814–2822. (2018).

[10] P. Plecha´c, K. Bobenhausen, B. Hammerich, Versification and authorship attribution. a pilot study on czech, german, spanish, and english poetry, Studia Metrica et Poetica, vol. 5, no. 2, pp. 29–54, 2018. (2018).

[11] R. Hou, C.-R. Huang, Robust stylometric analysis and author attribution based on tones and rimes, Natural Language Engineering, 2019, pp. 1–23. (2019).

[12] O. Aborisade, M. Anwar, Classification for authorship of tweets by comparing logistic regression and Naive Bayes classifiers, in: 2018 IEEE international conference on information reuse and integration, IEEE, 2018, pp. 269–276. (2018).

[13] C. Akimushkin, D.R. Amancio, O.N. Oliveira, On the role of words in the network structure of texts: Application to authorship attribution, Physica A: Statistical Mechanics and its Applications vol. 495, 2018, pp. 49–58, 10.1016/j.physa.2017.12.054. (2018).

[14] S. Voloshyn, V. Vysotska, O. Markiv, I. Dyyak, I. Budz and V. Schuchmann, Sentiment Analysis Technology of English Newspapers Quotes Based on Neural Network as Public Opinion Influences Identification Tool, 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 83-88. (2022).

[15] S. Albota, Modelling the impact of the pandemic on online communication: textual semantic analysis // CEUR Workshop Proceedings. – 2022. – Vol. 3171: Computational Linguistics and Intelligent Systems 2022: Proceedings of the 6th International conference on computational linguistics and intelligent systems (COLINS 2022). Vol. 1: Main conference, Gliwice, Poland, May 12-13, 2022, pp. 471–486. (2022).

[16] M. Karp, A. Burtnyk, I. Bekhta, N. Kunanets, O. Melnychuk, I. Shainer, Study of linking words in political speeches of Bill Clinton using Python // IEEE 17th International Conference on Computer Science and Information Technologies, CSIT 2022, November 10-12, 2022, Lviv, UKRAINE, pp. 77–82/83. (2022).

[17] M. Al-Ayyoub, Y. Jararweh, A. Rabab'ah, M. Aldwairi, Feature extraction and selection for Arabic tweets authorship authentication, Journal of Ambient Intellegence and Humanized Computing, 8 (3), 2017, pp. 383–393. (2017).

[18] S. Alanazi, Classical Arabic Authorship Attribution Using Simple Features, Project: Natural Language Processing, Jouf University, Saudi Arabia. September, (2018).

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

[20] A. Kornai, Mathematical Linguistics. Springer (2008).

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

[22] V. Turchyn, Matematychna statystyka. Navch. Posib. Vydavnychyj tsentr "Akademia": Kyiv, Ukraine, (1999). (in Ukrainian).

[23] V. Fetisov Paket statystychnoho analizu danyh STATISTICA, Nizhyn: NDU im. M. Hoholia, 2018, 114 s. (2018).

[24] V. Teslyuk, I. Kazymyra, Yu. Kordiiaka, I. Rybak, Modeli ta zasoby avtomatychnoho vyznachennia statystychnoho profiliu ukrainomovnyh tekstiv. Ukrainskyy zhurnal informatsiynyh tehnologiy. Tom 4. № 1. 2022, ss. 37 – 43. (2022)..

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