REPUBLIC OF AZERBAIJAN

On the rights of the manuscript

ABSTRACT

of the dissertation for the degree of doctor of sciences

DEVELOPMENT OF ALGORITHM AND SOFTWARE FOR A COMPUTER SYSTEM IDENTIFYING THE AUTHORSHIP OF TEXTS IN THE AZERBAIJAN LANGUAGE

Speciality:	1203.01- Computer sciences
Field of science:	Technical sciences
Applicant:	**Rustam Bakir Azimov**

Baku – 2024

The work was performed at the Institute of Control Systems of Ministry of Science and Education of the Republic of Azerbaijan.

Scientific supervisor: Corresponding member of ANAS, Doctor of Mathematical Sciences
Kamil Rajab Aida-zade

Official opponents: Doctor of Technical Sciences, Professor
Bayram Ganimat Ibrahimov

Doctor of Technical Sciences, Associate Professor
Mais Pasha Farhadov

Doctor of Philosophy on Engineering, Associate Professor
Shahla Surxay Huseynzade

Dissertation council ED 1.20 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Control Systems of Ministry of Science and Education of the Republic of Azerbaijan.

Chairman of the Dissertation Council: Academician of ANAS, Doctor of Technical Sciences, Professor
_________________________ **Ali Mahammad Abbasov**

Scientific Secretary of the Dissertation Council: Doctor of Philosophy on Engineering, Associate Professor
_________________________ **Tahir Ali Alizada**

Chairman of the Scientific Seminar: Doctor of Technical Sciences, Associate Professor
_________________________ **Rovshan Aghakishi Guliyev**

GENERAL DESCRIPTION OF THE WORK

The relevance and elaboration degree of the topic. The interest in the processing of texts written in natural language has led to the solution of various problems. One of the current directions in this field is text categorization. Texts can be categorized in terms of topic, sentiment, and writing style. Text categorization according to writing styles is widely used for document genre identification and recognition of authors of texts. It is assumed that text author recognition or text authorship identification is categorization of texts based on the writing styles of the authors.

Text author recognition is carried out on the basis of quantitative analysis of the writing styles of the authors in the texts. One of the first works dedicated to the quantitative analysis of the writing style of texts, i.e. stylometry, belongs to T.C. Mendenhol. In a study published at the end of the 19th century, T.C. Mendenhall showed that the number of words of different lengths (the number of one-letter words, the number of two-letter words, etc.) remained invariant in separate parts of a fiction literature work of Shakespeare and compared to the indicators of Bacon's works on the basis of a manual analysis without using any specific mathematical procedure. A.A. Markov in his study published in 1913 conducted a statistical analysis on the text of Pushkin's "Eugene Onegin" based on vowel-consonant counts. Later, in the 60s of the 20th century, F. Mosteller and D.L. Wallace used statistical analysis methods to identify the authorship of 12 out of 77 essays in New York newspapers in the 18th century about the first New York constitution. Authorship attribution established and developed in the works of scientists Mendenhol, F. Mosteller, D.L. Wallace, E. Stamatatos, P. Juola, J. Deiderich, Y. Zhao, V. Keselj, M. Kestemont, A.V. Anisimov, K.R. Aydazade, S.G. Talibov, etc.

As with other text categorization problem classes, text authorship recognition can be brought to pattern recognition. A lot of studies have been done in the field of pattern recognition in the Republic. In the 80s, under the leadership of Academician of ANAS T.A. Aliyev, G.G. Abdullayeva and N.H. Gurbanova created a system for recognizing writing written by hand in the Arabic-like alphabet using an electronic pen. In the 90s, under the leadership of corresponding member of ANAS R.A. Aliyev, a research was conducted in the field of fuzzy image recognition. Under the leadership of prof. A.K. Karimov, R.I. Davudova worked on the creation of optimal classifiers of dynamic systems, researche studies on state recognition of complex dynamic systems and cluster analysis were conducted. Under the leadership of prof. R.G. Mammadov, by A.S. Mutallimova technical vision systems were created. Under the leadership of Prof. O.G. Nusratov, a system for recognizing printed handwritten letters was developed. Under the leadership of the correspongin member of ANAS K.R. Aida-zade, by E.E. Mustafayev recognition of printed handwritten letters, by C.Z. Hasanov recognition of handwriting and by S.S. Rustamov speech recognition systems have been carried out. Under the leadership of A.N. Nasibov, work was carried out on the determination of cyber violence cases in Turkish language texts on social networks. Under the leadership of the corresponding member of ANAS R.M. Aliguliyev, scientific studies were conducted on the determination of plagiarism cases in texts and automatic text summarization.

The works done in the direction of text authorship recognition differ according to the text feature types (stylistic characteristics of the authors in the texts), text genre, volume, language, the authorship identification approach used, and so on. Among the scientific studies conducted on text authorship identification, one can find a number of works related to the recognition of newspaper column authors and other short texts and the recognition of the authors of fiction literature works.

Work was also done on the authorship recognition of the authors of texts in the Azerbaijani language. For example, one of their works, K.R. Aida-zade and S.G. Talibov used statistical methods and Support Vector Machine to recognize the authorship of small newspaper articles in the Azerbaijani language.

None of the existing text author recognition computer systems are designed to recognize the authors of texts in the Azerbaijani language. In the thesis, the characteristics of the texts in the

Azerbaijani language were comprehensively studied based on the lexicon of the language, the stylistic features of the texts were investigated, and a computer system for recognizing the authors of the texts in the Azerbaijani language was developed.

This shows that the topic of the dissertation is related to a relevant and important problem.

In the thesis, the principles of establishing a system for recognizing the authors of texts written in the Azerbaijani language were worked out, different types of features expressing the author's style in the text, effective text feature selection procedures, and the effectiveness of various machine learning methods in author recognition were studied on the considered author recognition problem.

Object and subject of the research. The research object of the thesis is the recognition of the authors of the texts written in the Azerbaijani language. The subject of the research is approaches based on known texts and their use with machine learning methods.

Goal and tasks of the research. The main goal of the thesis is to study the problems of authorship identification of texts in the Azerbaijani language, to develop solution algorithms and a computer system for recognizing the authors of the texts in the Azerbaijani language. In accordance with the purpose of the research, the following tasks were set in the thesis:

1. Development of author recognition methods that uses machine learning for Azerbaijani texts.
2. Development of procedures for calculating the values of text features of different types for use in recognizing the authors of texts in the Azerbaijani language.
3. Development of procedures for text feature selection for use in recognition of authors of texts in Azerbaijani language.
4. Development of principles and software of an author recognition computer system that allows author recognition of texts in the Azerbaijani language.

General research methodology. Machine learning methods, authorship recognition methodologies, information technologies and programming tools were used in the thesis.

The main provisions to be defended. The main provisions defended in the dissertation are the following:

1. Solution to the problem of author recognition of large and small literary fiction texts written in Azerbaijani language.
2. Development of text feature extraction algorithms for use in recognizing the authors of texts in the Azerbaijani language.
3. Development of procedures for text feature selection in recognition for use in recognizing the authors of texts in the Azerbaijani language.
4. Development of a computer system for authorship recognition of texts in the Azerbaijani language.

The scientific novelty. The scientific novelty of the thesis are as follows:

1. Procedures for selecting effective text features for use in recognizing the authors of texts in the Azerbaijani language have been proposed.
2. The proposed approach proposed for the combined processing of large documents (e.g. novels) and small documents (e.g. stories) for author recognition is justified using the results of two proposed empirical analyses.
3. Procedures for calculating the values of text features of different types have been proposed for use in recognizing the authors of texts in the Azerbaijani language.
4. Usage of different text feature types and different machine learning methods has been analyzed in the example of texts written in Azerbaijani language.

Theoretical and practical value of the research. Authorship identification is widely used to identify the author of an anonymous or disputed literary work, to verify the authorship of suicide letters, for information to determine whether an anonymous message or statement was written by a known terrorist, to identify the author of malicious computer programs, for example, computer viruses and malware, and to identify the authors of certain Internet texts.

Approbation and application. The main results of the thesis were presented at the following local and international conferences: 8th World Conference of Soft Computing (2021, Baku, Azerbaijan);

5th International Conference on Problems of Cybernetics and Informatics (2023, Baku, Azerbaijan); Международной научно-практической конференции с элементами научной школы - 2021, 2022, 2023 (Омск, Россия); Second International Bilateral Workshop on Science Between Dokuz Eylül University and Azerbaijan National Academy of Sciences (2022, İzmir, Türkiye); "Riyaziyyatın tətbiqi məsələləri və yeni informasiya texnologiyaları" adlı IV Respublika elmi konfransı (2021, Sumqayıt, Azərbaycan); "Riyaziyyatın fundamental problemləri və intellektual texnologiyaların təhsildə tətbiqi" mövzusunda II Respublika elmi konfransı (2022, Sumqayıt, Azərbaycan); Azərbaycan dilinin İKT problemləri, İKT-nin Azərbaycan dili problemləri (2023, Bakı, Azərbaycan); Tələbə və Gənc Tədqiqatçıların Beynəlxalq Elmi Konfransları - 2022, 2023 (Bakı, Azərbaycan); 2nd International Conference on Information Technologies and Their Applications (2024, Bakı Azərbaycan); Azərbaycan Respublikası Elm və Təhsil Nazirliyinin İdarəetmə Sistemləri İnstitutunun elmi seminarları, (2024, Bakı Azərbaycan). Reference was made to a work [15] in which some results of the thesis were published in a short period of time (see Alsanoosy, T., Shalbi, B., Noor, A. Authorship Attribution for English Short Texts // – Engineering, Technology & Applied Science Research – 2024. v. 14, no. 5, – pp. 16419-16426).

The software modules used for the structure and parametre identification of artificial neural networks in the thesis and the obtained results were used in the improvement of the "Form recognition system" application software package at the State Examination Center of the Republic of Azerbaijan, and the application proving documents (acts) were added in the appendix of the thesis.

Publication. 21 scientific works have been published on the thesis, 6 of them are articles [2, 4, 12, 19–21], 7 are conference papers [3, 8, 9, 15–18], 8 are abstracts [1, 5–7, 10, 11, 13, 14]. 2 articles were included in the Web of Science^{TM} ESCI international database of the Clarivate Analytics agency, 1 article in the РИНЦ database, and 1 article in the list of Ukrainian Supreme Attestation Committee. 4 of the conference papers and abstracts are included in the Scopus database, and 5 in the Russian Science Citation Index database.

The name of the institution where the thesis was performed. The thesis was performed at the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan.

Structure and volume of the dissertation. The dissertation consists of an introduction, 3 chapters, conclusion, a list of 110 references and an Appendix. The total volume of the thesis is 156 pages, and the main volume is 126 (166435 characters) pages, including 12 tables and 12 pictures. Including title page – 395 characters, table of contents – 1270 characters, introduction – 18230 characters, first chapter – 55921 characters, second chapter – 71491 characters, third chapter – 18190 characters.

CONTENT OF THE WORK

In the introduction the relevance of the topic is justified, the subject, purpose of the research, etc. is shown, the brief content and main results of the work are given.

In the first chapter the texts author recognition problems were analyzed.

In paragraph 1.1, the problem classes in the direction of text author analysis were studied, the place of author recognition problems in both author analysis of texts and texts categorization was analyzed. Then, the analysis of two paradigms widely used in solving author recognition problems was carried out.

Authorship analysis of texts can be used to identify the author of an anonymous or controversial fiction literature work (especially in copyright litigations), for authorship verification of suicide letters, in intelligence (for example, determining whether an anonymous message or statement was written by a known terrorist), it is widely used in the identification of the author of malicious computer programs (for example, computer viruses, malware), and in determining the authors of some texts on the Internet (e-mail letters, blog posts, texts on online forum pages).

In general, two main solution paradigms to author recognition are used to solve all author recognition problems:

- the author profile-based paradigm – the known texts of a certain author candidate are considered as a single unified text, in other words, during the solution, individual texts of each candidate author are combined and brought into a unified text, author recognition is done by comparing the characteristics in the unified texts of the authors with the characteristics in the given text which’s author is supposed to be recognized;
- text instance-based paradigm – each known author's text used in training is considered as an observation, the known texts of a certain candidate author being considered as individual text instances. Machine learning classification methods can be used with this paradigm.

In paragraph 1.2, A comparative analysis of some existing author recognition computer systems was performed.

For author recognition, many recognition systems have been developed. Among them, the most widespread computer systems are the Java Graphical Authorship Attribution Program developed in 2017 at Duquesne University, NeoNeuro Authorship Attribution from NeoNeuro and Flint AI Authorship Determination from Flint AI. Most of the mentioned systems only recognize authors of texts in English and some in some other languages.

None of the mentioned and many other author recognition computer systems are designed to recognize the authors of texts in the Azerbaijani language.

In paragraph 1.3, the characteristics of the problem of authorship recognition of the texts in the Azerbaijani language were mentioned, and the results of the computer experiments were analyzed for the evaluation of the effectiveness of the approach used in the joint processing of large-scale and small-scale literary works in the set of texts whose authors are known (dataset). In the third part of that paragraph, the stylistical characteristics of the authors' works were analyzed based on their statistical characteristics.

The dataset used in the considered authorship identification problem in the study was prepared as a result of acquisition of large-scale and small-scale works of 11 Azerbaijani writers. The number of large works per author in the dataset ranges from 1-5, and the number of small works varies from 0-46.

In author recognition, recognition of the authors of novel and short stories is done simultaneously. The fact that there are generally few texts with known authorship in the training set that will be used for effective feature selection in recognition and determining the parameters of mathematical models makes it even more difficult to solve the author recognition problem in a reliable and accurate way. A large number of candidate authors – writers in the considered problem have already died, therefore, it is not possible to naturally increase the number of texts with known authors that will be used to solve the author recognition problem. For a reliable and accurate solution to the problem under consideration, the approach of dividing large-scale works into a certain number of parts was used in order to overcome both the overall small number of works with unknown authors and the difficulties in processing large-scale and small-scale works together. However, it is questionable whether or not the individual texts obtained when dividing any large-scale work into parts can be considered separate examples (here literary works) from the point of view of authorship stylistics.

In order to consider these texts as separate literary work instances, two conditions must be met:

- when dividing a large-scale work, the texts obtained should be close to that large-scale text in terms of author stylistics,
- taking into account the author stylistics in the author's other literary works, the parts obtained by dividing a large-scale literary work should differ from each other stylistically.

In one of them, during the division of a large-volume work into separate texts, whether the author's stylistic indicators were preserved in the texts was analyzed in the form of character n-gram frequency features in the texts. A character n-gram is considered as a combination of certain n number of characters. Character n-gram frequency is the usage frequency of a certain character n-gram in a certain text. For this purpose, the text of one large work is divided into 10 texts (with approximately the same number of characters). The character 2-gram frequencies in each of these texts and the standard deviation of these frequency values were calculated. The presence of symbol 2-grams with small standard deviation values indicates that the authorship characteristics of that large text are preserved in the texts obtained when this text is divided. In other words, when a large-scale work is divided, the characteristics of the author in that large-scale work are preserved in the texts received (Table 1).

In Table 1, from left to right, 1st column is the row number, 2nd column is the character n-gram whose frequencies are used, 3rd column is the frequency in the first text part (i.e. one of 10 parts, where the text part number is conditionally given), 4nd column is the frequency in the second text part, 6th column is the frequency in the last, tenth text part, 7th column is the frequency in the large text itself, and 8th column is the standard deviation of the frequencies in the individual parts of the text. The symbols in the column names of the table and their descriptions are given below:

$\alpha_{\hat{\imath}_m^n}$ – character n-gram of certain n characters with an arbitrary multi-index $\hat{\imath}_m^n$, where $\alpha_{\hat{\imath}_m^n} = \left(\alpha_{i_1}, \dots, \alpha_{i_n}\right)$, $\hat{\imath}_m^n = (i_1, \dots, i_n)$, m is the number of characters in the character alphabet. Each of the indices i_v, $v = 1, 2, \dots, n$, in the mentioned multi-index can take the values in the given set $\{1,2, \dots, m\}$, where $1 \le i_v \le m,\ \ v = 1, 2, \dots, n$;

$\varphi_{j1\hat{\imath}_m^n}^i$ – the frequency of the character n-gram in the part number 1 of the i numbered text T_j^i of the j numbered author;

$\varphi_{j2\hat{\imath}_m^n}^i$ and $\varphi_{j10\hat{\imath}_m^n}^i$ – the frequencies of the character n-gram $\alpha_{\hat{\imath}_m^n}$ in the parts numbered as 2 and 10 of text T_j^i;

$f_{3j\hat{\imath}_m^n}^i$ – frequency of character n-gram $\alpha_{\hat{\imath}_m^n}$ in the text T_j^i;

$\sigma_{j\hat{\imath}_m^n}^i$ – the standard deviation of the character n-gram values $\varphi_{j1\hat{\imath}_m^n}^i,\ \varphi_{j2\hat{\imath}_m^n}^i, \dots, \varphi_{j10\hat{\imath}_m^n}^i$ in parts of the text T_j^i;
here $j = 1,2, \dots, l_i,\ i = 1,2, \dots, L,\ L$ is the number of candidate authors in the considered problem. Note that character n-gram frequencies in the other parts (3, 4, ..., 9) of the text are not given because they were enough to perform this analysis already.

As mentioned above, the frequency of the character n-gram $\alpha_{\hat{\imath}_m^n}$ in the given text T_j^i is denoted by $f_{3j\hat{\imath}_m^n}^i$. Here, the "3" in the index means that this feature belongs to the character n-gram frequency type. In the second chapter of the dissertation, the procedure for calculating the values of the five feature types used in the considered author recognition problem in the dissertation is described. Here – in the third paragraph of the first chapter, the frequency of the character n-gram $\alpha_{\hat{\imath}_m^n}$ in the text T_j^i, "3" in the symbol $f_{3j\hat{\imath}_m^n}^i$, in the second chapter, is used in the description of the rules for calculating the values of text features and the procedures for selecting effective features in recognition marked to be the same as the features.

The other analysis is based on the comparison of the stylistic characteristics in the texts obtained when the large-scale work is divided with the characteristics in the small-scale works, again in the example of character n-gram frequencies. In this proposed analysis, the texts of 1 large-scale work and more than 15 small-scale works are used.

Let us denote the indices set of texts of the author A^i as $J^i = \overline{J}^i \cup \underline{J}^i$, where $\overline{J}^i = \{\bar{j}_k^i : k = 1,2,\dots,\overline{\xi}_i\}$ and $\underline{J}^i = \{\underline{j}_k^i : k = 1,2,\dots,\underline{\xi}_i\}$ are the sets of numbers of large and small texts, respectively.

The average degree of difference between the frequencies $\varphi_{\bar{j}_1^i 1\hat{\imath}_m^n}^i$, $\varphi_{\bar{j}_1^i 2\hat{\imath}_m^n}^i, \dots,$ $\varphi_{\bar{j}_1^i \mathcal{N}_{\bar{j}_1^i}^i \hat{\imath}_m^n}^i$ of the given character n-gram $\alpha_{\hat{\imath}_m^n}$ in the texts $T_{\bar{j}_1^i 1}^i$, $T_{\bar{j}_1^i 2}^i, \dots,$ $T_{\bar{j}_1^i \mathcal{N}_{\bar{j}_1^i}^i}^i$ obtained when dividing the large-scale work of a certain author is as follows:

$$\mathcal{F}_{\hat{\imath}_m^n}^{1,1} = \frac{2}{\left(\mathcal{N}_{\bar{j}_1^i}^i\right)^2} \sum_{x=1}^{\mathcal{N}_{\bar{j}_1^i}^i} \sum_{y=1}^{\mathcal{N}_{\bar{j}_1^i}^i / 2} abs\left(\varphi_{\bar{j}_1^i x\hat{\imath}_m^n}^i - \varphi_{\bar{j}_1^i y\hat{\imath}_m^n}^i\right),$$

$abs(\cdot)$ is the absolute value of the number, $\mathcal{N}_j^i$ is the number of parts into which the j numbered text T_j^i of the author A^i is divided.

The average degree of difference among the frequencies $f_{3\underline{j}_1^i\hat{\imath}_m^n}^i, f_{3\underline{j}_2^i\hat{\imath}_m^n}^i, \dots, f_{3\underline{j}_{\underline{\xi}_i}^i\hat{\imath}_m^n}^i$ of the given character n-gram $\alpha_{\hat{\imath}_m^n}$ in the $\underline{j}_1^i, \underline{j}_2^i, \dots, \underline{j}_{\underline{\xi}_i}^i$ numbered small-scale texts $T_{\underline{j}_1^i}^i, T_{\underline{j}_2^i}^i, \dots, T_{\underline{j}_{\underline{\xi}_i}^i}^i$ of a certain author A^i is as follows:

$$\mathcal{F}^{2,2}_{\hat{\imath}^n_m} = \frac{2}{\left(\underline{\xi}_i\right)^2} \sum_{v=1}^{\underline{\xi}_i} \sum_{z=1}^{\underline{\xi}_i/2} abs\left(f^i_{3\underline{j}^i_v \hat{\imath}^n_m} - f^i_{3\underline{j}^i_z \hat{\imath}^n_m}\right).$$

The average degree of difference among large-scale work’s text parts $T^i_{\bar{\jmath}^i_1 1}$, $T^i_{\bar{\jmath}^i_1 2}$,…, $T^i_{\bar{\jmath}^i_1 \mathcal{N}^i_{\bar{\jmath}^i_1}}$ and small-scale texts $T^i_{\underline{j}^i_1}, T^i_{\underline{j}^i_2}, \dots, T^i_{\underline{j}^i_{\underline{\xi}_i}}$ of a given author A^i by a given character n-gram $\alpha_{\hat{\imath}^n_m}$ is as follows:

$$\mathcal{F}^{1,2}_{\hat{\imath}^n_m} = \frac{2}{\mathcal{N}^i_{\bar{\jmath}^i_1} \underline{\xi}_i} \sum_{x=1}^{\mathcal{N}^i_{\bar{\jmath}^i_1}} \sum_{v=1}^{\underline{\xi}_i} abs\left(\varphi^i_{\bar{\jmath}^i_1 x \hat{\imath}^n_m} - f^i_{3\underline{j}^i_v \hat{\imath}^n_m}\right).$$

The degree of stylistic difference between a large-scale work’s parts on a certain character n-gram $\alpha_{\hat{\imath}^n_m}$ is considered not to be close to the degree of stylistic difference of small-scale works, nor to the degree of stylistic difference of a large-scale work’s parts, if the following $\dot{\mathcal{F}}_{\hat{\imath}^n_m}$ is small:

$$\dot{\mathcal{F}}_{\hat{\imath}^n_m} = abs\left(\mathcal{F}^{1,2}_{\hat{\imath}^n_m} - \frac{\mathcal{F}^{1,1}_{\hat{\imath}^n_m} + \mathcal{F}^{2,2}_{\hat{\imath}^n_m}}{2}\right).$$

For a simple interpretation of the value of a certain character n-gram $\alpha_{\hat{\imath}^n_m}$, the following quantity $\dot{\mathcal{F}}_{\hat{\imath}^n_m}$ is included:

$$\ddot{\mathcal{F}}_{\hat{\imath}^n_m} = \frac{\dot{\mathcal{F}}_{\hat{\imath}^n_m}}{abs\left(\mathcal{F}^{1,1}_{\hat{\imath}^n_m} - \mathcal{F}^{2,2}_{\hat{\imath}^n_m}\right)} \times 100.$$

Table 1

Results of the comparison of individual parts of the large volume literary work

N	$\alpha_{\hat{\imath}^n_m}$	$\varphi^i_{j1\hat{\imath}^n_m}$	$\varphi^i_{j2\hat{\imath}^n_m}$	...	$\varphi^i_{j10\hat{\imath}^n_m}$	$f^i_{3j\hat{\imath}^n_m}$	$\sigma^i_{j\hat{\imath}^n_m}$
1	an	0.0140	0.0127	...	0.0128	0.0125	0.0020
2	da	0.0129	0.0128	...	0.0126	0.0122	0.0018
3	in	0.0092	0.0118	...	0.0113	0.0121	0.0017
4	ar	0.0131	0.0126	...	0.0121	0.0119	0.0015
5	la	0.0125	0.0122	...	0.0105	0.0110	0.0015

From the small percentage values in the first column from the right in Table 2, it can be seen that such text features can be found that when dividing a large-scale work, the texts obtained are not close to that large-scale work or to small-scale works in terms of author's stylistics. Table 2-5 shows only 5 rows due to space limitation.

Different text feature types were used in the thesis. Among the text feature types used in the study, character n-gram frequency and word frequency type are distinguished by the large number of possible features (for example, the number of words whose frequencies in a text can be used as the text features). Therefore, it is necessary to choose the features that will be effective in recognition among the features of these types. In the study, four different feature selection procedures were used to select the features of these two types. The mathematical description of these feature selection procedures was given and their effectiveness in recognition was analyzed based on the results of computer experiments conducted on the considered author recognition problem. In the third paragraph of the first chapter, the frequencies of character n-grams with $n = 2$ characters and the frequencies of the most commonly used words by the authors on the works of the two writers in the considered problem were analyzed. Table 3 shows the average value and variance of the frequencies of some character n-grams in the texts of two authors - Suleyman Rahimov and Ismayil Shikhli, where

$$\dot{\mu}^{i}_{\hat{i}^{n}_{m}} = \frac{1}{l_i}\sum_{j=1}^{l_i} f^{i}_{3j\hat{i}^{n}_{m}},$$

$$\dot{\sigma}^{i}_{\hat{i}^{n}_{m}} = \frac{1}{l_i}\sum_{j=1}^{l_i}\left(f^{i}_{3j\hat{i}^{n}_{m}} - \dot{\mu}^{i}_{\hat{i}^{n}_{m}}\right)^2.$$

Table 2

Results of the comparison of individual parts of a large literary work with small literary works

N	$\alpha_{\hat{i}^{n}_{m}}$	$\mathcal{F}^{1,1}_{\hat{i}^{n}_{m}}$	$\mathcal{F}^{1,2}_{\hat{i}^{n}_{m}}$	$\mathcal{F}^{2,2}_{\hat{i}^{n}_{m}}$	$\ddot{\mathcal{F}}_{\hat{i}^{n}_{m}}$
1	ae	0	2.86197E-06	5.34234E-06	3.57
2	vn	9.63262E-06	7.64103E-06	5.34234E-06	3.58
3	vg	4.81631E-06	9.79213E-05	0.000178458	3.62
4	lz	4.81631E-06	6.60234E-05	0.000118915	3.64
5	lh	4.81631E-06	2.16771E-05	3.61856E-05	3.75

The dot over the sigma ("σ") in the notation $\dot{\sigma}_{\hat{\imath}_m^n}^{i}$ above is intended to conveniently distinguish it from the notation $\sigma_{j\hat{\imath}_m^n}^{i}$, where $\sigma_{j\hat{\imath}_m^n}^{i}$ is the standard deviation of the frequencies $\varphi_{j1\hat{\imath}_m^n}^{i}$, $\varphi_{j2\hat{\imath}_m^n}^{i}$, $\varphi_{j10\hat{\imath}_m^n}^{i}$ of a character n-gram $\alpha_{\hat{\imath}_m^n}$ in the text parts of T_{jk}^{i}, $k = 1, 2, \dots, \mathcal{N}_j^{\,i}$, of the text T_j^i, but $\dot{\sigma}_{\hat{\imath}_m^n}^{i}$ is the standard deviation of the frequencies of a character n-gram $\alpha_{\hat{\imath}_m^n}$ in the text T_j^i, $j = 1,2,\dots,l_i$, of an author A^i.

It can be seen from Table 3 that the statistical characteristics of different authors for different character n-grams can differ from each other. Obviously, the variance values in this table vary in scale with respect to the mean value. Therefore, we use the following coefficient of variation to make the results for all authors easier to interpret:

$$d_{\hat{\imath}_m^n}^{i} = \begin{cases} \dfrac{\dot{\sigma}_{\hat{\imath}_m^n}^{i}}{\dot{\mu}_{\hat{\imath}_m^n}^{i}}, & if\, \mu_k^j \geq \rho, \\ \varrho, & else, \end{cases}$$

where ρ is some small number and ϱ is some very large number, for example, $\rho = 10^{-4}$, $\varrho = 10^{12}$ and is chosen based on the results of experiments. The values of these $d_{j\hat{\imath}_m^n}^{i}$ quantities for some character n-grams are given in table 4. From the values in this table, it can be seen that the statistical indicators of different authors on different features (here, character n-gram frequency type of features) vary.

Table 5 shows the usage frequency of the five most frequently used words in the texts of the author candidates in the training set.

From the character n-grams and word frequencies and other statistical indicators (listed in Tables 3-5), it is clear that the indicators of usage of different features differ from each other by different candidate authors. Therefore, feature selection procedures were used in the study.

Table 3

Statistical indicators on the character n-gram frequencies on the texts of two authors

N	$\alpha_{\hat{\imath}_m^n}$	$\dot{\mu}^{i}_{\hat{\imath}_m^n}$		$\left[\dot{\sigma}^{i}_{\hat{\imath}_m^n}\right]^2$	
		Süleyman Rəhimov ($i = 9$)	İsmayıl Şıxlı ($i = 10$)	Süleyman Rəhimov ($i = 9$)	İsmayıl Şıxlı ($i = 10$)
1	aa	0.00025	0.00042	3.20254E-08	2.74829E-08
2	ab	0.00235	0.00146	9.76353E-07	5.05273E-08
3	ac	0.00125	0.00135	3.12249E-08	2.30432E-08
4	aç	0.00103	0.00092	7.07818E-08	3.59291E-09
5	ad	0.00374	0.00380	4.79537E-07	1.15944E-07

Table 4

Statistical indicators on the frequencies of some character n-grams on the texts of different authors

N	$\alpha_{\hat{\imath}_m^n}$	$d^{1}_{\hat{\imath}_m^n}$	$d^{2}_{\hat{\imath}_m^n}$	$d^{3}_{\hat{\imath}_m^n}$	…	$d^{9}_{\hat{\imath}_m^n}$	$d^{10}_{\hat{\imath}_m^n}$	$d^{11}_{\hat{\imath}_m^n}$
1	aa	0.55	0.78	0.55	…	1.02	0.60	0.96
2	ab	0.97	0.45	0.36	…	0.59	0.29	0.70
3	ac	0.55	0.77	0.23	…	0.28	0.17	0.42
4	aç	0.66	0.86	0.34	…	0.43	0.19	0.66
5	ad	0.23	0.39	0.16	…	0.20	0.13	0.49

Table 5

Usage frequencies of the 5 most frequently used words in the training set in authors' texts (in percentage)

v	ω_v	$\phi_v^{5,1,1}$	$\phi_v^{5,1,2}$	$\phi_v^{5,1,3}$	…	$\phi_v^{5,1,10}$	$\phi_v^{5,1,11}$	$\phi_v^{5,2}$
1	bir	2.0	2.7	1.9	…	1.4	2.7	0.0
2	bu	2.1	1.6	1.6	…	0.9	1.5	0.0
3	ki	1.9	2.7	1.4	…	0.8	1.2	0.0
4	də	1.5	1.1	1.3	…	1.1	0.4	0.0
5	və	2.6	3.1	1.0	…	0.8	2.4	0.0

In the second chapter the methodology used to recognize the authors of texts in the Azerbaijani language was studied.

In paragraph 2.1, it was mentioned about the recognition methods used in the study.

Support vector machine (SVM), random forest (RF) methods, and multilayer feedforward artificial neural network (ANN) models of machine learning were used in the considered author recognition problem.

In paragraph 2.2, a brief literature review of known text features was carried out, the extraction rules of text features related to the types used in the considered author recognition problem and the proposed procedures for selecting effective features in recognition were described.

In author recognition, any given text is characterized by certain text features in a certain feature group. The features in this feature group can belong to different feature types. In the study, feature groups consisting of features related to the sentence length frequency, word length frequency, character n-gram frequency, variance of character n-gram frequencies (proposed in the work), and word frequency types were used. The values of features in these feature groups in a certain text constitute the feature vector of that text. Mathematical descriptions of the procedures for calculating text feature values and selecting effective text features in recognition using the features are provided using the following notations.

Consider the set of given author candidates $A = \left\{A^i : i = 1, 2, \dots, L\right\}$, where L is the number of author candidates (here and after $|D|$ denote the number of elements of an arbitrary set D). Let's denote set of texts of author A^i by $T^i = \left\{T_j^i : j = 1, 2, \dots, l_i\right\}$, here l_i denote the number of texts of A^i, $i = 1, 2, \dots, L$.

In the authorship recognition problem, let's denote the set of texts written by candidate authors (or dataset) by $T = \left\{T^i : i = 1, 2, \dots, L\right\}$. Any text T_j^i is a set consisting of sentences s_{jk}^i, $k = 1, 2, \dots, N_j^{si}$, any sentence s_{jk}^i is a set consisting of words $w_{jk\nu}^i$, $\nu = 1, 2, \dots, N_{jk}^{wi}$, any word $w_{jk\nu}^i$ is a set of letters $c_{jk\nu\eta}^i$, $\eta = 1, 2, \dots, N_{jk\nu}^{ci}$, and they are denoted as follows:

$$T_j^i = \{s_{jk}^i : k = 1, 2, \dots, N_j^{si}\},\ \ j = 1, 2, \dots, l_i,\ \ i = 1, 2, \dots, L,$$
$$s_{jk}^i = \{w_{jk\nu}^i : \nu = 1, 2, \dots, N_{jk}^{wi}\},\ \ k = 1, 2, \dots, N_j^{si},\ \ j = 1, 2, \dots, l_i,\ \ i = 1, 2, \dots, L,$$
$$w_{jk\nu}^i = \{c_{jk\nu\eta}^i : \eta = 1, 2, \dots, N_{jk\nu}^{ci}\},\ \ \nu = 1, 2, \dots, N_{jk}^{wi}, k = 1, 2, \dots, N_j^{si},\ \ j = 1, 2, \dots, l_i,\ \ i = 1, 2, \dots, L.$$

Here T_j^i is the j^{th} text of author A^i; $s_{jk}^i,\ k = 1, 2, \dots, N_j^{si}$, are the sentences in text T_j^i; $w_{jk\nu}^i,\ \nu = 1, 2, \dots, N_{jk}^{wi}$, are the words in the k^{th} sentence of text T_j^i, where $k = 1,2, \dots, N_j^{si}$; $c_{jk\nu\eta}^i \in \alpha^1,\ \eta = 1, 2, \dots, N_{jk\nu}^{ci}$, are characters in the ν^{th} word of the k^{th} sentence of text T_j^i, where $\nu = 1, 2, \dots, N_{jk}^{wi}, k = 1,2, \dots, N_j^{si}$, and N_j^{si}, N_{jk}^{wi}, $N_{jk\nu}^{ci}$ are the number of sentences in T_j^i, the number of words in the k^{th} sentence of text T_j^i, the number of characters in the ν^{th} word of the k^{th} sentence of text T_j^i, respectively, α^1 is character alphabet set, $|\alpha^1|$ is the number of characters in this character alphabet set.

The set of texts with known authors used in the solution of the author recognition problem – the dataset is divided into two non-intersecting subsets $T = \bar{T} \cup \tilde{T},\ \bar{T} \cap \tilde{T} = \emptyset$. One of them $\bar{T}$ is the training set, the other is the test set $\tilde{T}$.

Sentence length frequency. By sentence length, we mean the number of words in the sentence, by word length, we mean the number of letters in the word. The number of occurrence of sentences with sentence length v in j^{th} text T_j^i of i^{th} author A^i is as follows:

$$b_{1jv}^i = \left|\{s_{jk}^i : N_{jk}^{wi} = v,\ \ k = 1,2, \dots, N_j^{si}\}\right|,$$

where $v \in I = \{1, 2, \dots\}, j = 1,2, \dots, l_i,\ \ i = 1,2, \dots, L.$

Frequency of occurrence of sentences with sentence length v in j^{th} text T_j^i of i^{th} author A^i is calculated as follows:

$$f_{1jv}^i = \frac{b_{1jv}^i}{N_j^{si}},$$

where $v \in I$.

Word length frequency. The number of occurence of words with word length v in j^{th} text T_j^i of i^{th} author A^i is as follows:

$$b_{2jv}^{i} = \left|\left\{s_{jk}^{i} \colon N_{jk}^{wi} = v, \;\; k = 1,2,\dots,N_{j}^{si}\right\}\right|,$$

where $v \in I$.

Frequency of occurrence of words with word length v in jth text T_j^i of ith author A^i is calculated as follows:

$$f_{2jv}^{i} = \frac{b_{2jv}^{i}}{N_j^{si}},$$

where $v \in I$.

Character n-gram frequency. In the scientific literature, the character n-gram is considered to mean any given combination of the given n $(n = 1,2,\ldots)$ characters. For a given n, let us denote the character alphabet set, i.e., the set of all character 1-grams from which characters in any character n-gram are selected as follows:

$$\alpha^1 = \{\alpha_k \colon k = 1, 2, \dots, |\alpha^1|\}.$$

Hereafter in the article, $\hat{\imath}_m^n = (i_1, \dots, i_n)$ notation will be used for n dimensional multi-index $\hat{\imath}_m^n$. Here everyone of the i_v, $v = 1, 2, \dots, n$, indices can take values in the given set $\{1,2,\dots,m\}$, hence $1 \le i_v \le m, \;\; v = 1, 2, \dots, n$.

A character n-gram with an arbitrary multi-index $\hat{\imath}_m^n$ whose characters are chosen from the character alphabet set α^1 will be used as $\alpha_{\hat{\imath}_m^n} = \left(\alpha_{i_1}, \dots, \alpha_{i_n}\right)$, where $m = |\alpha^1|$.

Let us denote the set of all character n-grams as follows:

$$\alpha^n = \left\{\alpha_{\hat{\imath}_m^n} \colon 1 \le i_v \le m, \;\; v = 1, 2, \dots, n, \;\; m = |\alpha^1|\right\}.$$

To depict the rule for calculating the frequencies of character n-grams in an arbitrary text T_j^i, let us denote this text as a set of $\tilde{c}_{jk}^i, \;\; k = 1, 2, \dots, \widetilde{N}_j^i$, characters as follows:

$$T_j^i = \left\{\tilde{c}_{j1}^i, \dots, \tilde{c}_{j\widetilde{N}_j^i}^i\right\},$$

where $\widetilde{N}_j^i$ is the number of characters in the text T_j^i, $j = 1, 2, \dots, l_i, \;\; i = 1, 2, \dots, L$.

The number of occurrence of an arbitrary character n-gram $\alpha_{\hat{\imath}_m^n} \in \alpha^n$ in the text T_j^i is as follows:

$$b_{3j\hat{\imath}_m^n}^{i} = \left|\left\{\left\{\tilde{c}_{jk}^i, \dots, \tilde{c}_{jk+n-1}^i\right\} \colon \left\{\tilde{c}_{jk}^i, \dots, \tilde{c}_{jk+n-1}^i\right\} = \alpha_{\hat{\imath}_m^n}, \;\; k =\right.\right.$$

$$1, 2, \dots, \dot{N}_j^i\}\Big|, \text{ where } m = |\alpha^1|,$$

$\dot{N}_j^i$ is the number of character n-grams that can be extracted from the text T_j^i, $j = 1, 2, \dots, l_i$, $i = 1, 2, \dots, L$.

Frequency of occurrence of an arbitrary character n-gram $\alpha_{\hat{\imath}^n_m} \in \alpha^n$ in the text T_j^i is calculated as follows:

$$f^i_{3j\hat{\imath}^n_m} = \frac{b^i_{3j\hat{\imath}^n_m}}{\dot{N}_j^i}, \text{ where } m = |\alpha^1|.$$

Variance of character n-gram frequencies. Statistical characteristics (e.g., variances) of character n-gram frequencies in separate parts of an arbitrary given text can be used as text features. In the study, the analysis of the recognition effectiveness of this type of features was carried out.

If we calculate the frequencies of an arbitrary character n-gram in the separate parts of a given text and find the variance of these frequencies, this variance shows the fulfillment of this character n-gram in the separate parts—at the beginning, in the middle, and at the end of that text, i.e., the stability of the character n-gram in the text. In other words, if the variance of frequencies of a character n-gram in a text, which is calculated as aforementioned, is small, this n-gram can be considered stable in that text. Even if we calculate the frequency of an arbitrary character n-gram in each of the texts of a certain author candidate in the training set and find the variance of these frequencies, the small value of this variance indicates that the character n-gram characterizes the texts of that author well. Or if we combine the known texts of each of the author candidates in the considered problem and get a single text for each author candidate (for L number of authors, we obtain L texts), if the variance of an arbitrary character n-gram in one of these texts is small, and the variances in the others' are large, this character n-gram characterizes one author candidate better than others, so this character n-gram can be effectively used in the considered author recognition problem. It turns out that the variance values, which indicate how character n-grams are distributed in a certain text or the stability of character n-grams, are informative in terms of author recognition. Hence, we considered that along with the frequencies of character n-grams in a given text, their variances in the text can be used as text features as well. In other words, by dividing the given text into a certain number of separate, non-intersecting parts and calculating the frequencies of an arbitrary character n-gram in these parts, the variance of these frequencies can be used as some of the features of that text.

For an arbitrary n in order to calculate the values of the statistical characteristics of an arbitrary character n-gram frequencies in a given text, at first the text have to be divided into certain parts, the frequency of this n-gram in each of these text parts should be found, and the variance value of these frequencies should be calculated.

Separate, non-intersecting $\mathcal{N}_j^i$ number of parts of any text T_j^i are shown in the following:

$$T_j^i = \left\{T_{jk}^i\colon\ k = 1, 2, \dots, \mathcal{N}_j^i,\ \ T_{jk_1}^i \cap T_{jk_2}^i = \emptyset,\ \ 1 \le k_1, k_2 \le \mathcal{N}_j^i,\ \ k_1 \ne k_2\right\},$$

where T_{jk}^i is the k^{th} part of the text T_j^i, $\mathcal{N}_j^i$ is the number of parts of the text T_j^i, $j = 1, 2, \dots, l_i,\ \ i = 1, 2, \dots, L$.

Let us denote the frequency of an arbitrary character n-gram $\alpha_{\hat{\imath}_m^n} \in \alpha^n$ in T_{jk}^i – the k^{th} part of the text T_j^i with $\varphi_{jk\hat{\imath}_m^n}^i$, where $m = |\alpha^1|,\ j = 1, 2, \dots, l_i,\ \ i = 1, 2, \dots, L$.

The variance of frequencies $\varphi_{jk\hat{\imath}_m^n}^i,\ k = 1, 2, \dots, \mathcal{N}_j^i$, of an arbitrary character n-gram $\alpha_{\hat{\imath}_m^n} \in \alpha^n$ in the seperate parts $T_{jk}^i,\ k = 1, 2, \dots, \mathcal{N}_j^i$, of the text T_j^i is calculated as follows:

$$f_{4j\hat{\imath}_m^n}^i = \sigma_{j\hat{\imath}_m^n}^i = \frac{1}{\mathcal{N}_j^i} \sum_{k=1}^{\mathcal{N}_j^i} \left(\varphi_{jk\hat{\imath}_m^n}^i - \mu_{j\hat{\imath}_m^n}^i\right)^2,$$

where $\mu_{j\hat{\imath}_m^n}^i = \frac{1}{\mathcal{N}_j^i} \sum_{k=1}^{\mathcal{N}_j^i} \varphi_{jk\hat{\imath}_m^n}^i$ is the average value of frequencies $\varphi_{jk\hat{\imath}_m^n}^i,\ k = 1, 2, \dots, \mathcal{N}_j^i$, of an arbitrary character n-gram $\alpha_{\hat{\imath}_m^n} \in \alpha^n$ in the seperate parts $T_{jk}^i,\ k = 1, 2, \dots, \mathcal{N}_j^i$, of the text T_j^i, $m = |\alpha^1|,\ j = 1, 2, \dots, l_i,\ \ i = 1, 2, \dots, L$.

Word frequency. Let us denote non-repeated word set of words

$w_{jkv}^{i},\ v = 1,2,\dots,N_{jk}^{wi},\ k = 1,2,\dots,N_{j}^{si},\ T_{j}^{i} \in \bar{T},\ j = 1,2,\dots,l_{i},\ i = 1,2,\dots,L,$ in texts $T_{j}^{i} \in \bar{T},\ j = 1,2,\dots,l_{i},\ i = 1,2,\dots,L,$ in training set $\bar{T}$ as follows:

$$W = \left\{\omega_{p}\colon\ T_{j}^{i} \in \bar{T},\ \ \omega_{p} \in s_{jk}^{i},\ \ k = 1,2,\dots,N_{j}^{si},\ \ j = 1,2,\dots,l_{i},\ \ i = 1,2,\dots,L,\ \ \omega_{p_1} \neq \omega_{p_2}\ if\ p_1 \neq p_2\right\},$$

where $\omega_{p} = \left\{a_{r}^{p}\colon\ a_{r}^{p} \in \alpha^{1},\ r = 1,2,\dots,\left|\omega_{p}\right|\right\}, \alpha^{1}$ is the character alphabet set.

The number of occurrence of an arbitrary word $\omega_{v} \in W$ in the text T_{j}^{i} is as follows:

$$b_{5jv}^{i} = \left|\left\{w_{jkv}^{i}\colon w_{jkv}^{i} = \omega_{v},\ \ N_{jkv}^{ci} = |\omega_{v}|,\ \ v = 1,2,\dots,N_{jk}^{wi},\ \ k = 1,2,\dots,N_{j}^{si}\right\}\right|,$$

where $|\omega_{v}|$ is the number of letters in the word ω_{v}, $v \in \{1,\dots,|W|\}$, $j = 1,2,\dots,l_{i},\ \ i = 1,2,\dots,L.$

Frequency of occurrence of an arbitrary word $\omega_{v} \in W$ in the text T_{j}^{i} is calculated as follows:

$$f_{5jv}^{i} = \frac{b_{5jv}^{i}}{\widetilde{N}_{j}^{i}},$$

where $\widetilde{N}_{j}^{i} = \sum_{k=1}^{N_{j}^{si}} N_{jk}^{wi}$ is the number of words in text T_{j}^{i}, here N_{jk}^{wi} is the number of words in the kth sentence of text T_{j}^{i}, $v \in \{1,\dots,|W|\}$, $j = 1,2,\dots,l_{i},\ \ i = 1,2,\dots,L.$

The results of our research show that if the frequency of an arbitrary word in the given text are used together with the frequencies of the unified texts of the author candidates (here, by the united text of a candidate author, we mean the text obtained merging of the texts of the author in the training set) among the text features it positively influences in terms of recognition effectiveness. In addition to the frequencies of a given word in the unified texts of the author candidates, the frequency of this word in the unified text obtained as the result of merging all the texts in the training set can be used among the text features. The rules for calculating the frequencies of an arbitrary word in the unified texts of the candidate authors and in the text obtained by merging all the texts in the training set are given below.

Set of texts of author A^i in training set is as follows:

$$\bar{T}^i = T^i \cap \bar{T},$$

where T^i is the set of texts of author A^i in the dataset T, $\bar{T} \subset T$ is set of all texts in the training set, $i = 1,2, \dots, L$.

The frequency of an arbitrary word $\omega_v \in W$ in the unified text obtained by merging the texts of an arbitrary author candidate A^i in the training set is as follows:

$$\phi_v^{51i} = \frac{\sum_{j=1}^{l_i} b_{5jv}^i}{\sum_{j=1}^{l_i} \tilde{N}_j^i}, \text{ where } T_j^i \in \bar{T},$$

b_{5jv}^i is the number of word ω_v in text T_j^i (formula (5)), $v \in \{1, \dots, |W|\}$, $i = 1,2, \dots, L$.

The frequency of an arbitrary word $\omega_v \in W$ in the unified text obtained by merging all the texts in the training set is as follows:

$$\phi_v^{52} = \frac{\sum_{i=1}^{L} \sum_{j=1}^{l_i} b_{5jv}^i}{\sum_{i=1}^{L} \sum_{j=1}^{l_i} \tilde{N}_j^i}, \text{ where } T_j^i \in \bar{T},$$

b_{5jv}^i is the number of word ω_v in text T_j^i (formula (5)), $v \in \{1, \dots, |W|\}$.

In the problems of author recognition of texts, it is necessary to describe the texts with such features that the texts of different authors differ from each other, consequently, the features should have a discriminatory nature among the authors. Since some features can be more discriminatory among authors than others, it is necessary to choose the most discriminatory ones among possible features. Certain feature selection procedures can be used to select those features which are more discriminative for a certain feature type. Such feature selection procedures are typically used to select features for feature types that are too many for a human to select based on his or her own heuristics or small-scale manual (computable without a computer in a reasonable time) analyses. In the considered author recognition problem in the study, some proposed feature selection procedures were used to select words and character n-grams.

In the considered author recognition problem in the study, two different feature selection procedures described below are proposed for selecting words, and one feature selection procedure for selecting character n-grams (recall that these feature selection procedures are carried out based on the texts whose authors are known in the training set) whose frequencies will be used as text features in author recognition. In one of the feature selection procedures used to select words (the procedure for separately selecting words frequently used by authors (procedure one)), the mostly (or averagely or least) used words by each of the authors in their texts are determined separately per authors, followed by selecting some of the most (or average or least) occurring words in each author's lexicon, and a combined word set consisting of the majority of words in the authors' lexicons is created. In the other procedure used for word selection (procedure for selecting frequently used words in the training set (second procedure)), the texts of different authors are considered together, not separately, (regardless of their authors) in the texts in the training set and the most (or average or least) occurring words are selected. The feature selection procedure used for selecting character n-grams (the procedure for selecting character n-grams that is often used in the training set (the third procedure)) is the character n-gram analogue of the second (last) procedure used for word selection: (without taking into account authors of the texts) the most (or average or least) occurring character n-grams in the texts in the training set are selected.

Using these feature selection procedures, the recognition effectiveness of the feature groups was analyzed (recognition effectiveness means the adequacy of the model or method used with a certain feature group).

In paragraph 2.3, in the considered author recognition problem, the effectiveness of using machine learning methods with different feature groups consisting of feature related to the types of text features was evaluated.

The two feature groups with the highest recognition accuracy on dataset-0 consisting of large and small works of the authors in the author recognition problem considered in the study (one of these two feature groups contains features related to the character n-gram frequency type, and the other feature group contains features related to the word frequency type) from the results given in Figure 1, it is clear that although the recognition accuracies obtained on dataset-0 with these two feature groups are the same, author recognition can be done with a greater degree of confidence with the feature group consisting of character n-gram frequencies.

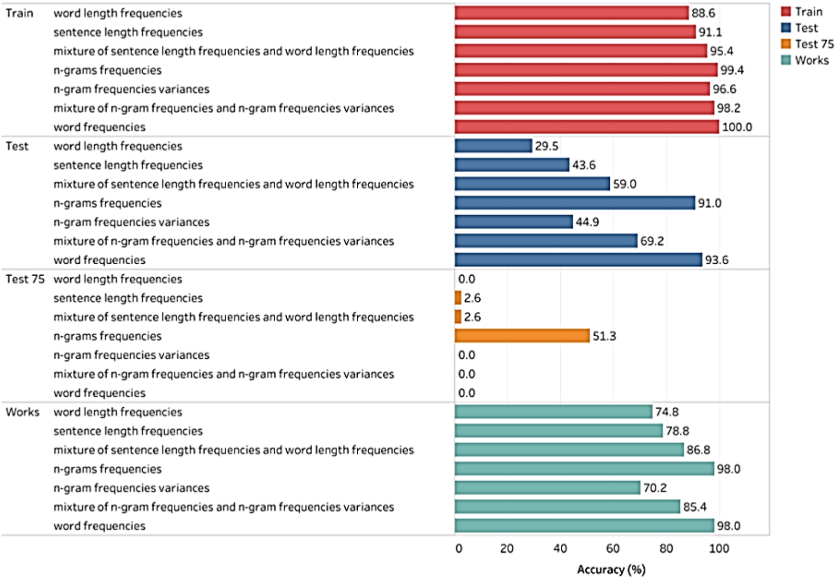

Figure 1. Maximum recognition accuracies obtained using features belonging to different feature types

In the third chapter The description of the software for recognizing the authors of texts in the Azerbaijani language is provided.

In paragraph 3.1, the structure of the developed system and the work stages on the system are briefly reviewed.

In paragraph 3.2, the process of introducing texts whose author is known, as well as texts whose author should be recognized (the author of which is considered unknown) into the system, system training process, that is, preparing the system for author recognition, recognition process and results demonstration process are described.

In paragraph 3.3, group decision-making approaches that are possible to be implemented in the system for the collective use of individual author recognition methods were analyzed.

In the appendix, the application proving documents and the text of the codes of the main program modules of the software for authorship identification of the texts in the Azerbaijani language is given.

CONCLUSION

1. Author recognition methods using machine learning have been developed for texts in the Azerbaijani language.
2. Procedures for extraction of features of different types have been developed for use in recognizing the authors of texts in the Azerbaijani language.
3. Procedures of effective text feature selection for use in recognizing the authors of texts in the Azerbaijani language were proposed and developed.
4. The use of different text feature types with different methods of machine learning for use in recognizing the authors of texts in the Azerbaijani language has been studied.
5. The software of the author recognition computer system, which allows recognition of the authors of texts in the Azerbaijani language, has been developed.
6. The software modules used for structure and parameter identification of Artificial Neural Networks and the obtained results were used in the improvement of the "Form recognition system" application software package at the State Examination Center of the Republic of Azerbaijan.

The main content of the dissertation is published in the following scientific works:

1. Мустафаев, Э.Э., Ахмедлы, Н.А., Азимов, Р.Б. Сравнительный анализ применения нейронных и сверточных нейронных сетей распознаванания рукопечатных букв Азербайджанского алфавита // Информационный бюллетень Омского научно-образовательного центра ОмГТУ и ИМ СО РАН в области математики и информатики, – Омск, Россия:– 22-29 апреля 2021, т.5, № 1, – с.94-95. **(РИНЦ)**
2. Mustafayev, E.E., Azimov, R.B. Comparative analysis of the application of multilayer and convolutional neural networks for recognition of handwritten letters of the Azerbaijani alphabet // – Kiev, Ukraine: Cybernetics and Computer Technologies, – 2021. №3, – pp. 65–73. **(the journal is included to the Ukraine Supreme Attestation Committee)** https://doi.org/10.34229/2707-451X.21.3.6.
3. Əzimov, R.B. Azərbaycan dilinin çap əlyazma hərflərinin tanınmasında bəzi əlamətlərin müqayisəsi // “Riyaziyyatın tətbiqi məsələləri və yeni informasiya texnologiyaları” adlı IV Respublika elmi konfransı, – Sumqayıt, Azərbaycan: 9-10 dekabr 2021, № 9, s.79-84. **(РИНЦ)**
4. Мустафаев, Э.Э., Азимов, Р.Б. Использование многослойных и сверточных нейронных сетей для распознавания рукопечатных букв на примере азербайджанского алфавита // – Омск, Россия: Прикладная математика и фундаментальная информатика, – 2022. т.8, №2, – с. 38-45. **(РИНЦ)**
5. Azimov, R.B., Mustafayev, E.M. Comparison of SVM and ANN methods for recognition of authorship of texts // Applied Mathematics and Fundamental Informatics: Proceedings of the XII Intern. youth scientific-practical. conf. with elements of science. schools, – Омск: 16-21 may 2022, – pp. 60-61. **(РИНЦ)**
6. Əzimov, R.B. Azərbaycan çapəlyazma əlifbasının tanınmasına

yanaşmaların müqayisəli təhlili // Tələbə və Gənc Tədqiqatçıların III Beynəlxalq Elmi Konfranslarının “Proseslərin avtomatlaşdırılması və informasiya təhlükəsizliyi-2022” konfransı, – Bakı, Azərbaycan: – 26-27 aprel 2022, – s. 157-159.

7. Aida-zade, K.R., Mustafayev, E.M., Azimov, R.B. Features analysis for application in a computer recognition systems of Azerbaijani texts authorship // Second International Bilateral Workshop on Science Between Dokuz Eylül University and Azerbaijan National Academy of Sciences, – İzmir, Türkiye: – 18 november 2022, – p. 11.
8. Əzimov, R.B., Mustafayev, E.M. Azərbaycan dilindəki mətnlərin müəlliflərini tanıyan kompüter sistemində istifadə üçün müxtəlif əlamət qruplarının müqayisəli təhlili // “Riyaziyyatın fundamental problemləri və intellektual texnologiyaların təhsildə tətbiqi” mövzusunda II Respublika elmi konfransı, – Sumqayıt, Azərbaycan: – 15-16 Dekabr 2022, s. 34-39.
9. Azimov, R.B., Aida-zade, K.R. Analysis of features of texts for use in an author recognition system // ICT problems of the Azerbaijani language, Azerbaijani language problems of the ICT, – Baku, Azerbaijan: – 21-22 Fevral 2023, s. 22-25.
10. Əzimov, R.B. Müəllif Tanıma Sistemində Mətn Əlamətləri Siniflərinin İstifadəsinin Təhlili // Tələbə və Gənc Tədqiqatçıların IV Beynəlxalq Elmi Konfranslarının “Rəqəmsal Transformasiya-2023” konfransı, – Baku, Azerbaijan: – 18-19 aprel 2023, s. 286-288.
11. Azimov, R.B. Analysis of use of text feature classes in an author recognition system // Applied Mathematics and Fundamental Informatics: Proceedings of the XII Intern. youth scientific-practical. conf. with elements of science, – Omsk, Russia: – 15-20 may 2023, s. 88-89. **(РИНЦ)**
12. Azimov, R.B. Approaches to the recognition of handwritten letters of the Azerbaijani language and their analysis // Informatics and Control Problems, – 2023, v. 43, no. 3, – pp. 32-40. https://doi.org/10.54381/icp.2023.1.05.

13. Azimov, R.B. Analysis of the Use of Methods and Feature Groups for Author Recognition on the Example of Texts in the Azerbaijani Language // Abstracts of V International Conference on “Problems of Cybernetics and Informatics” (PCI 2023), – Baku, Azerbaijan: – p. 90.
14. Khalilov, C.J., Mustafayev, E.E., Mahmudov, I.M., Azimov, R.B. Computer System of Analysis of the Mass Exam Results // Abstracts of V International Conference on “Problems of Cybernetics and Informatics” (PCI 2023), – Baku, Azerbaijan: – pp. 29-30.
15. Azimov, R.B. Analysis of the Use of Methods and Feature Groups for Author Recognition on the Example of Texts in the Azerbaijani Language // 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI), – Baku, Azerbaijan: – 2023, – pp. 1-4. **(Scopus)** https://doi.org/10.1109/PCI60110.2023.10325956.
16. Khalilov, C.J., Mustafayev, E.E., Mahmudov, I.M., Azimov, R.B. Computer System of Analysis of the Mass Exam Results // 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI), – Baku, Azerbaijan: – 2023, – pp. 1-4. **(Scopus)** https://doi.org/10.1109/ PCI60110.2023.10326014.
17. Azimov, R.B. Comparison of artificial and convolutional neural networks in recognition of handwritten letters of Azerbaijani alphabet // Recent Developments and the New Directions of Research, Foundations, and Applications. Studies in Fuzziness and Soft Computing (8th World Conference of Soft Computing), – Baku, Azerbaijan: Springer, – 3-5 February, – 2022, – v. 422, pp. 377-385. **(Scopus)** https://doi.org/10.1007/978-3-031-20153-0_31.
18. Mustafayev, E.E., Azimov, R.B. Computer System of Evaluation of the Mass Exam Results Based on Recognition of Handprinted Azerbaijani Characters // Proceedings of the Information Technologies and Its Applications Conference, – Baku, Azerbaijan: Springer, – 23-25 April, – 2024, – v. 2, pp. 171-183. **(Scopus)**

19. Azimov, R.B., Efthimios, P. A Comparative Study of Machine Learning Methods and Text Features for Text Authorship Recognition in the Example of Azerbaijani Language Texts // Algorithms, – 2024. V. 17, no. 6, – 242 (25 pages). https://doi.org/ 10.3390/a17060242. **(Web of Science, ESCI)**
20. Aida-zade, K.R., Azimov, R.B.. Analysis of the use of text features in the authorship identification of literary works in the Azerbaijani language // Informatics and Control Problems, – Baku, Azerbaijan: – 2024, v. 44, no. 1, – pp. 51-58.
21. Azimov, R.B. Comparative Analysis of using Different Text Features, Models, and Methods in Text Author Recognition // Cybernetics and Systems Analysis, – Kiev, Ukraine: – 2024. V. 60, no. 5, – pp. 711-725. https://doi.org/10.1007/s10559-024-00709-z. **(Web of Science, ESCI)**

The defense will be held on 6 December 2024 at 14:00 at the meeting of the Dissertation council ED 1.20 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan.

Address: AZ1141, Baku, St. B. Vahabzade, 68.

Dissertation is accessible at the Library of the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan.

Electronic versions of dissertation and its abstract are available on the official website (http://www.isi.az) of the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan.

Abstract was sent to the required addresses on 4 November 2024.

Signed for print: 01.11.2024

Paper format: A5

Volume: 37076

Number of hard copies: 20