The Approach to Minimize the Impostor Method Errors in the Author Identification Open Problem*

Anastasia Iskhakova [4]
shumskaya.ao@gmail.com

Svetlana Kruglova[1]
ms.kr666@mail.ru

Sergey Melnikov [2]
melnikov@linfotech.ru

Evgeniy Sidorov[3]
sidorov@linfotech.ru

[1] Moscow State University,
[2] Linguistic and Information technologies, LLC
[3] Bauman Moscow State Technical University,
[4] V. A. Trapeznikov Institute of Control
Moscow, Russian Federation

Abstract

The article discusses the open problem of authorship identification, when the analyzed text can be written by one of known authors or an unknown author. For identification, the impostor method with a large number of external texts is used. A practical approach to achieve the minimum arithmetic mean of errors of the first and second kinds based on joint selection of the values of two parameters of the method is proposed. The experimental results are given for texts in Russian and Arabic. In the experimental results, the minimum arithmetic mean of errors is achieved when the length of the texts-impostors is equal to the average length of the texts of the author's collections.

Keywords: *authorship identification, verification problem, open problem, impostor method, author's collection.*

1 Introduction

A significant amount of research has been devoted to the development of automatic methods for text authorship identification (see the review of statistical approaches in [Malyutov, 2006] and the review of existing engineering solutions in [Romanov, Shelupanov, Meshcheryakov, 2011]). In most works, the closed identification problem is solved when the analyzed text is written by one of the authors known to the system. The open problem, when the analyzed text can be written both by a known author and by an unknown one, is more complicated.

The need for automatic identification of text authorship in an open problem arises in a number of modern applications, including:

- identification of authorship of literary and historical documents by computational text analysis ([Mamajev, Marusenko, Piotrowska, Ronzhin, 2017],

- identification of compromised accounts in social networks ([Barbon, Igawa, Zarpelão, 2017],

- construction of filters for phishing attacks analysis ([Duman etc., 2016],

- continuous user identification ([Brocardo etc., 2017]),

- detection of fake reviews of goods and services ([Layton, Watters, Ureche, 2013]),

- building opinion mining systems and analyze of user reviews by personalizing approach ([Panicheva, Cardiff, Rosso, 2010]),

- detection of attacks on web resources using automatically generated texts with minor content changes ([Shahid etc., 2017]),

- the author identification in the problems of detection of destructive orientation in electronic materials ([Iskhakova, Iskhakov, Meshcheryakov, 2019]),

- a range of problems associated with the detection of artificially generated texts, including the problem of plagiarism detection ([Shumskaya, 2016]).

The basic particular case of an open identification problem is the verification problem, that is, checking whether a given text is written by the only known author.

Let A be a collection of texts by a known author, and X be a text which authorship is unknown. When solving the verification problem, the value $p\,(X|A)$ is evaluated that shows how likely it is that the text X is written by the same author as the texts in the collection A.

In [Potha, Stamatatos, 2019], the conventional classification of methods for solving the problem of text authorship verification is proposed according to the following criteria:

- only the text X and texts from the collection A are used, or external texts are involved;

- all texts from A are combined and the so-called "Author profile" is formed, or texts from A participate in the calculations separately.

At present the most effective verification methods include the method of impostors and its modifications ([Koppel, Winter, 2014], [Seidman, 2013], [Khonji, Iraqi, 2014], [Potha, Stamatatos, 2017]). According to the above classification, the method of impostors uses external texts, and the texts of the author's collection participate in the calculations separately.

The variants of the method described in the above works use various approaches to obtain a set of external texts (impostors), the choice of which affects the efficiency. In different variations of the impostor method, different sets of characteristics of the author's style and different values of the parameters (the number of iterations, the number of impostors at each iteration, thresholds of criteria, the proportion of randomly selected components of the feature vector) can be used. The values of these parameters significantly affect the accuracy and computational complexity of the method. The choice of these values should be carried out depending on the language of the texts and the characteristics of the author's collections and the analyzed text.

In this paper, we consider the following problem. There are several authors represented by their own collections of texts. The text received for analysis can be written by either one of the known authors or an unknown one. It is necessary to verify the authorship of the incoming text for each available author. The objectives of this work are:

- to propose a method for determination of threshold values of criteria for the impostor method, at which the minimum of arithmetic mean values of errors of the first and second kinds, calculated on existing collections of authors, is achieved,

- to study the influence of the lengths of impostors on the accuracy of the method,
- to study the influence of the parameter “proportion of randomly selected components of the feature vector” on the accuracy of the impostor method.

2 Implementation of the impostor method

The analyzed text X is associated with the feature vector $\mathbf{x}$ consisting of frequencies of N-grams of letters of the language alphabet ($N = 1, 2, 3, 4$), words ($N = 1, 2$), functional words (in the subsequence of functional words obtained after deleting all other words of the text ([Germanovich, Melnikov, Khvostenko, 2017])) ($N-1, 2$), words shape patterns, the proportion of different words in the text. Word shapes used following replacements: upper case letters are replaced with the letter “C”, lowercase letters are replaced with the letter “c”, the numbers in the text are replaced with the letter “N”. For Arabic texts, word shape patterns were not used. NLTK files were used for functional words lists. The listed features are considered to be effective for authorship identification.

The degree of proximity $sim(X, Y)$ of texts X and Y is calculated as the minimum-maximum measure ([Koppel, Winter, 2014]) between their feature vectors,

$$sim(X, Y) = minmax(\mathbf{x}, \mathbf{y}) = \frac{\sum min(x_i, y_i)}{\sum max(x_i, y_i)}.$$

The degree of authorial proximity $IM(X, Y)$ of texts X and Y is calculated using a large set S of external extraneous texts, the so-called impostors, as follows.

Input: X, Y - texts; S - impostor set.

Algorithm parameters: k, n - integer; $\theta, \Delta, rate > 0$.

Output: $IM(X, Y)$.

1. *Score*=0.

2. Repeats k times

   a. The coordinates of the vector $\mathbf{x}$, that will be used to calculate the degree of proximity are randomly selected. The fraction of the set of selectable coordinates is equal to $rate$;

   b. Texts $I_1 \dots I_n \in S$ are randomly selected;

   c. $Score = Score + \frac{sim(X,Y)*sim(Y,X)}{Max_{1\leq j\leq n} sim(X,I_j)*Max_{1\leq j\leq n} sim(Y,I_j)}$.

3. $IM(X, Y) = \frac{1}{k} Score$.

Let M authors be represented by collections $A_i = \{T_{ij}, j = 1, \dots, \mid A_i \mid\}, i = 1, \dots, M$ of their texts. For the analyzed text T, the following values are calculated:

$$GIM_\Delta(T, A_i) = \frac{1}{\mid A_i \mid} \sum_j Ind_\Delta(T, T_{ij}),$$

where

$$Ind_\Delta(T, T_{ij}) = \left\{ \begin{array}{l} 1, IM(T, T_{ij}) \geq \Delta \\ 0, IM(T, T_{ij}) < \Delta \end{array} \right., i = 1, \dots, M.$$

In the case $GIM_\Delta(T, A_i) \geq \theta$, we will assume that the text T is written by the i-th author. Otherwise, we will assume that the text T is written by another author.

3 Selection of the parameters values and the study of their influence on the method accuracy

An error of the first kind occurs when the text T was written by the i-th author, but the inequality $GIM_\Delta(T, A_i) < \theta$ is satisfied. An error of the second kind occurs if the text T is written by another author, but the inequality $GIM_\Delta(T, A_i) \geq \theta$ is satisfied.

In this paper, we consider the accuracy measure equal to the average over all texts of all authors of arithmetic mean errors of the first and second kinds. At the text authorship identification this text was deleted from the collection of its author. The proposed approach can be easily extended to the case when the accuracy measure is an arbitrary smooth function of errors, for example, a weighted sum.

3.1 Selection of the values of the parameters Δ and θ

The accuracy of the method depends significantly on the choice of the Δ and θ parameters values. For different versions of the author identification problem (language, text size, genre, subject, etc.) this choice can be made differently. For the identification problem considered at the CLEF PAN'13 conference in [Seidman, 2013] (under the conditions of the specific method of choice of impostors) the optimized values of these parameters were given for English, Greek and Spanish.

In this paper, to find out how the accuracy of the method depends on the Δ and θ parameters values, the following method is proposed.

The ranges of the parameter $\Delta \in (0, Max_{i,j,T} IM(T, T_{ij}))$ and $\theta \in (0, 1)$ changes are divided equally into several segments, and the errors of the first and second kinds are calculated at the nodes of the resulting grid. All other parameters of the method are considered as fixed. Each text T from the author's collection $\{T_{ij}\}$ acts as the analyzed one at a time, while the text itself is deleted from the author's collection. The best pair of the Δ and θ values is considered to be the one for which the arithmetic mean of errors of the first and second kinds for all $\sum_{i=1}^{M} \mid A_i \mid$ analyzed texts is minimal.

3.2 Selection of the impostor set

The conclusion about the authorship of the considered text is made on the basis of statistics, during the calculation of which the measure of proximity between the analyzed text and the author's texts, as well as with texts from the set of impostors, is repeatedly calculated. The set of impostors can be chosen in many ways.

In [Koppel, Winter, 2014], three ways to select a set of impostors were considered:

- fixed in advance set of texts that are not connected with collections of author's texts,

- the set of obviously external author's texts of the same genre as the analyzed ones (the author's blogs texts were considered),

- use of queries to the Google search engine. Small sets of random words were selected in a random subset of the author's collection. Each such set is used as a search query for Google. The first few found search results, after extraction of the text component from the found documents, were used to form impostors (in [Siedman, 2013] it was proposed to take 1500 first words of each of them). The process continued until a sufficiently large text corpus was accumulated. This method allows to get a large number of impostors, lexically connected with author's texts, but requires access to online resources.

In this work we used large corpora of texts accumulated in advance. Fragments similar in lexical composition to the author's texts were selected from them as impostors. All texts of the selected corpora were "glued" into one text, which was then cut into fragments of equal length $L_{impostor}$. The last word of the fragment, if it turned out to be incomplete, was deleted. A random subset of author's texts was selected several times from the author's collection. A list of not function words that appeared in the texts of the selected subset was compiled. Sets of words were randomly selected from this list, which were then searched in all the "sliced" fragments. 10 fragments were selected for which the sum of the frequencies of the found words was maximum. The process ended when 1500 fragments were found. These found fragments were used as impostors.

To study the dependence of the accuracy of the method on the value of the parameter $L_{impostor}$, using $L_{impostor} = ILR * L_{author}$, where L_{author} is the average length of the author's texts, we will test several values of the ILR (Impostors Length Ratio) coefficient at fixed $k, n, rate$. For each variant of the parameter values, we will calculate the Δ and θ thresholds at which the minimum of the arithmetic mean of the errors is achieved, as well as the value of this average.

3.3 The choice of the rate value (the proportion of randomly selected components of the feature vector).

The $rate \in (0, 1)$ parameter determines the proportion of randomly selected components of the feature vector when calculating the degree IM of author proximity. In different implementations of the impostor method, its values were chosen differently. In [Siedman, 2013] $rate = 0.4$ was chosen for English and Spanish, and $rate = 0.6$ - for Greek. In [Khonji, Iraqi, 2014] and [Potha, Stamatatos, 2017] $rate = 0.4$ and $rate = 0.5$ were used respectively.

In this paper, to study the dependence of the accuracy of the method on the value of parameter $rate$ we used total testing of values of this parameter in a given range with a given step at fixed $k, n, L_{imposter}$. For each variant of the parameter values, we will calculate the Δ and θ thresholds at which the minimum of the average errors is achieved, as well as the value of this average.

4 Data for experiments

The Russian-language author's collection consisted of 536 author's texts in Russian and contained the texts of 29 authors (from 7 to 25 texts per author), the average length of the text was $L_{author} = 3543$ characters, including numbers, spaces, and punctuation marks. The collection was drawn from the publications of the top 100 popular bloggers in LiveJournal for 2018-2019 on socio-political issues.

The Arabic-language author's collection consisted of 403 author's texts in Arabic and contained the texts of 20 authors (from 19 to 22 texts per author), the average text length was $L_{author} = 5447$ characters, including numbers, spaces and punctuation marks. The collection was drawn from texts of analytical articles (author's columns) of a socio-political orientation, published in the largest electronic media of Algeria, Egypt, Iraq, Lebanon, Palestine, Saudi Arabia, Syria and Tunisia in 2015-2018.

The texts of the author's collections were viewed visually by linguists, medium-sized texts which did not include volumetric citation were selected.

The collected texts of Russian-language and Arabic-language news agencies for 2015-2018 were used for the formation of impostor sets. The volume of each of the text corpus was approximately 350 MB. The accumulated corpora passed the procedure of cleaning from non-text fragments, texts not in target languages and fuzzy duplicates. The procedures developed in [Belozerov, etc., 2016] were used to collect and clean the corpora of texts.

The normalization of texts of author's collections and impostors was not carried out.

5 Description of software components for research

The developed software components, as well as the stages of the calculations, are divided into three parts, the first of which receives the texts of the author's collections as input, and all subsequent ones - the results of the calculations of the previous software component.

Processing consists of the following steps.

5.1 Stage of preliminary calculations (Python)

This stage is implemented in Python3 with the use of the NLTK library for natural language text processing. At this stage, the collected texts for the formation of impostors are glued and divided into fragments of the required length. Lists of functional words are formed. The following data is calculated: the average length of the texts, the feature vectors and statistics for both author's texts and impostors. Texts are divided into words (tokenized), functional words are deleted and the necessary statistics are calculated, the necessary statistics are also calculated on the subsequences of functional words. All results are saved in protobuf files and transferred to the next stage.

5.2 Stage of the method of impostors (C++)

At this stage, the results of the previous stage, as well as ranges and steps of changing of $k, n, rate$ parameters are loaded into the memory. For each pair of texts and for each value from the parameter range with a given step, the impostor method is used. The results (Δ and θ arrays), as well as application parameters, are saved in a json file, which is convenient for human reading on the one hand, and is a computer-readable format on the other.

This stage has high computational complexity and therefore it is implemented in C++. The procedure for selection of impostors also uses lists of functional words that match the lists of the preliminary calculation stage. At this stage of the calculation, such lists are the parts of the application code.

5.3 Stage of analysis of the results (Python)

At this stage, the json file is loaded into the memory and the minimum error values are calculated. The application is implemented in Python 3; to simplify the calculations the NUMPY library is used. As a result of this stage, the final data of experiments are calculated, according to which tables and graphs are built.

The above division into stages has significantly reduced the time to study the characteristics of the method, since the stage of preliminary calculations allows to get all the necessary data for the impostor method in advance and not to calculate them again each time.

To plot the surfaces in Fig. 1 and Fig. 2 and the lines in Fig. 3, the ListPlot3D and ListLinePlot functions of Mathematica software package ver.10 were used.

6 The results of the experiments

6.1 The influence of the Δ and θ parameters values on the identification accuracy

The experiments on calculation of the arithmetic mean of the errors were carried out with the following values of the method parameters: $k = 10, n = 20, rate = 0.4$. Impostors were built in the manner described above with the condition $L_{impostor} = L_{author}$. Fig. 1 and Fig. 2 show the images of surfaces corresponding to average errors depending on the Δ and θ parameters values for the Russian and Arabic languages, respectively.

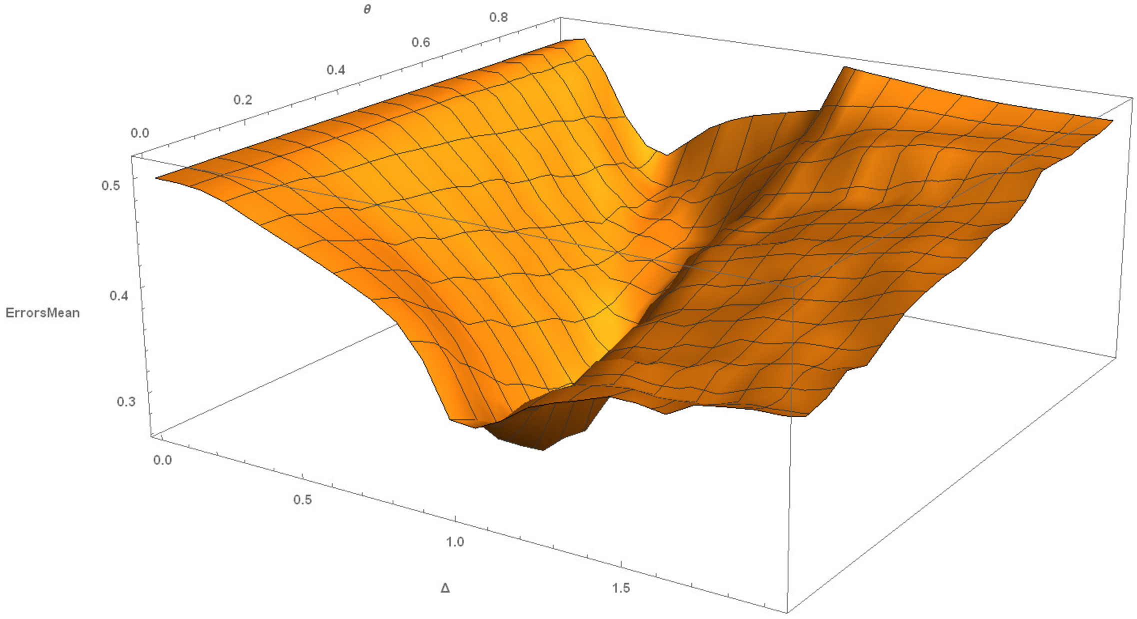

Figure 1: Arithmetic mean of the errors depending on the Δ and θ parameters, Russian language.

The minimum value of the arithmetic mean of the errors on the Russian data is 0.26 and achieved with the values $\Delta = 0.96, \theta = 0.3$ (Fig. 1), on the Arabic data - 0.16 and achieved with the values $\Delta = 0.96, \theta = 0.45$ (Fig. 2).

6.2 Effect of impostor length on identification accuracy

The impostors of the length $L_{impostor} = ILR * L_{author}$ were used, where L_{author} is the average length of author's texts, ILR is a coefficient. The calculations were carried out at the Russian and Arabian corpora. Table 1 and Fig. 3 show the arithmetic means of the errors of the first and second kinds and the corresponding values of θ and Δ. The values $k = 10, N = 20, rate =$

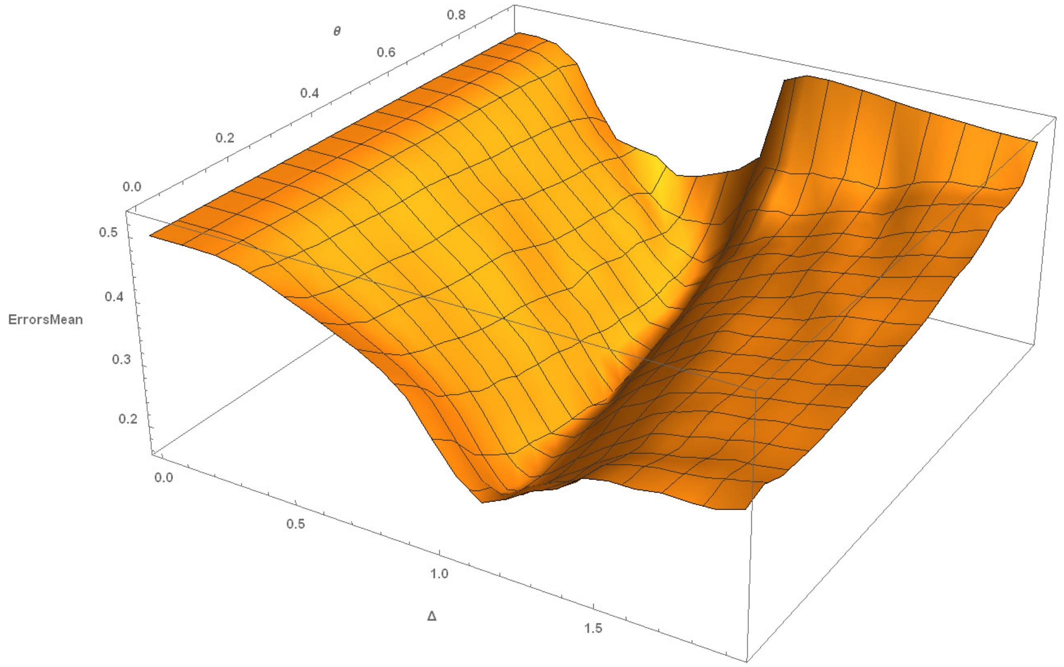

Figure 2: Arithmetic mean of the errors depending on the Δ and θ parameters, Arabic language.

0.4 were used. The values of pairs of the θ and Δ parameters which minimize the arithmetic mean of the errors, for each variant of the ILR coefficient were calculated as described above.

6.3 The influence of the *rate* parameter values (proportions of randomly selected components of the feature vector) on the identification accuracy

The parameter determines the proportion of coordinates in the feature vector that are used to calculate IM distances. The calculations were carried out at the Russian corpus with $k = 10, N = 20, ILR = 1$. The values of pairs of the θ and Δ parameters which minimize the arithmetic mean of the errors for each variant of the $rate = 0.2, \ldots, 0.6$ were calculated as described above.

Table 2 presents the values of the arithmetic mean of the errors of the first and second kinds and the corresponding values of θ and Δ for the five *rate* parameter values.

7 Results analysis and future research

The results of the experiments show that the value of the *rate* parameter in the range $0.3 \leq rate \leq 0.6$ practically does not affect the method accuracy.

The length $L_{impostor}$ of the used impostors significantly affects the accuracy of the method that was not noted earlier in the cited publications. Calculations performed on the developed software for two different languages show that the lengths of impostors should be approximately equal to the lengths of author's texts. Therefore, the methods of selection of impostors,

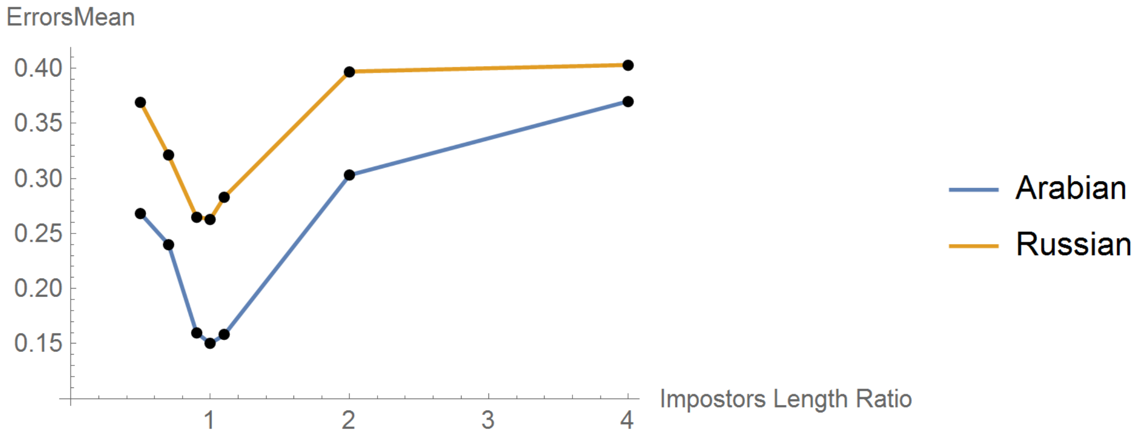

Figure 3: Arithmetic means of the errors of the first and second kinds depending on the ratio of the impostor text lengths to the author's text lengths, Russian and Arabic languages.

Table 1: The dependence of the identification accuracy on the value of the ILR, Russian and Arabic languages.

ILR	The arithmetic mean of errors (Russian/Arabian)	Δ (Russian/ /Arabian)	θ (Russian/ /Arabian)
0.5	0.369/0.268	0.96/0.72	0.05/0.9
0.7	0.321/0.24	0.88/0.96	0.2/0.75
0.9	0.265/0.16	0.88/0.96	0.35/0.55
1	0.263/0.15	0.96/0.96	0.3/0.4
1.1	0.283/0.158	0.88/0.96	0.3/0.4
2	0.397/0.303	2.75/1.96	0.4/0.65
4	0.403/0.37	5.75/3.0	0.6/0.65

in particular, using search engines, should take into account their length.

An important result is obtaining experimental dependences of errors on the threshold values of the criteria (Fig. 1 and Fig. 2). The graphs of the obtained surfaces show that the form of the functional dependence of errors on the θ and Δ thresholds does not allow to carry out consecutive one-dimensional minimization on these parameters.

The differences in the identification errors achieved for Russian and Arabic are likely to be related both to the peculiarities of the languages and sets of used author's features, and to the difference in the average lengths of the author's texts (3543 and 5447 characters, respectively). Probably, the same reasons are responsible for differences in the smoothness of the surfaces depicted in the Fig. 1 and Fig. 2.

In our opinion, promising areas of research include:

- study of the dependence "accuracy" - "computational complexity" in a wide range of parameters and of the method,

- expansion of the research into other languages, in particular, Romance, Germanic and Turkic and languages of Southeast Asia.

Table 2: Identification accuracy dependence on the *rate* parameter. Russian language.

rate	Arithmetic mean of errors	Δ	θ
0.2	0.255	0.96	0.3
0.3	0.259	0.96	0.2
0.4	0.263	0.96	0.3
0.5	0.263	0.96	0.3
0.6	0.267	0.96	0.3

8 Conclusion

A practical method for calculation of the impostor method parameters values to minimize the arithmetic mean of errors of the first and second kinds is presented. Errors are calculated on the author's texts collections. Large text corpora are used to form impostors.

The three-stage structure of software developed for carrying out of experiments is described.

The results of experiments for the author's texts collections in Russian and Arabic show that the used criteria thresholds, as well as the impostor length, significantly affect the identification accuracy. The best accuracy is achieved when the impostor length is approximately equal to the average length of the texts in the author's collections. The proportion of randomly selected components of the feature vector in the studied range does not significantly affect the accuracy of the method.

Acknowledgements

The reported study was funded by RFBR according to the research project # 18-29-22104.

References

[Malyutov, 2006] Malyutov, M. (2006). Authorship Attribution of Texts: A Review. *Information Transfer and Combinatorics*, LNCS 4123, pp. 362–380.

[Romanov, Shelupanov and Meshcheryakov, 2011] Romanov, A., Shelupanov, A., and Meshcheryakov, R. (2011) Development and research of mathematical models, methods and software of information processes in authorship identifying. V-Spektr, Tomsk, Russia. 188 p. (In Rus.) = Razrabotka i issledovanie matematicheskih modelej metodiki programmnyh sredstv informacionnyh processov pri identifikacii avtora teksta. Tomsk.: Izdatelstvo "V-Spektr", 2011. - 188s.

[Mamajev, Marusenko, Piotrowska, Ronzhin, 2017] Mamajev, N., Marusenko, M., Piotrowska, X., Ronzhin, A. (2017). Burrows's Delta in Authorship Attribution of Russian Literary Texts // *Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. – S.Petersburg, Russia, November 27, 2017.* CEUR Workshop Proc. Vol. 2233, pp.107-119.

[Barbon, Igawa and Zarpelao, 2017] Barbon, S., Igawa, R., and Zarpelao, B. (2017). Authorship verification applied to detection of compromised accounts on online social networks: A continuous approach. *Multimedia Tools and Applications.* 76(3), 3213–3233. DOI: 10.1007/s11042-016-3899-8.

[Duman etc., 2016] Duman, S., Kalkan-Cakmakci, K., Egele, M., Robertson, W., and Kirda, E. (2016). EmailProfiler: Spearphishing filtering with header and stylometric features of emails. *In Proceedings of the 2016 IEEE 40th Annual Computer Software and Applications Conference.* (Vol. 1, pp. 408–416). DOI:10.1109/compsac.2016.105.

[Brocardo etc., 2017] Brocardo, M., Traore, I., Woungang, I., and Obaidat, M. (2017). Authorship verification using deep belief network systems. *International Journal of Communication Systems,.* 30(12), e3259. DOI:10.1002/dac.3259.

[Layton, Watters, Ureche, 2013] Layton, R., Watters, P., and Ureche, O. (2013). Identifying faked hotel reviews using authorship analysis. *In Proceedings - 4th Cybercrime and Trustworthy Computing Workshop,* CTC '13, pp. 1–6. DOI: 10.1109/CTC.2013.8.

[Panicheva, Cardiff, Rosso, 2010] Panicheva, P., Cardiff, J., and Rosso, P. (2010). Personal sense and idiolect: Combining authorship attribution and opinion analysis. *In Proceedings of the International Conference on Language Resources and Evaluation,* LREC.

[Shahid, etc., 2017] Shahid, U., Farooqi, S., Ahmad, R., Shafig, Z., Srinivasan, P., and Zaffar, F. (2017). Accurate detection of automatically spun content via stylometric analysis. *In Proceedings of the 2017 IEEE International Conference on Data Mining* (ICDM) (pp. 425–434). DOI: 10.1109/ICDM.2017.52

[Iskhakova, Iskhakov and Meshcheryakov, 2019] Iskhakova, A., Iskhakov, A., and Meshcheryakov, R. (2019). Research of the estimated emotional components for the content analysis. *Journal of Physics: Conference Series,* 1203. 012065. DOI: 10.1088/1742-6596/1203/1/012065.

[Shumskaya, 2016] Shumskaya, A. (2016). The method of determining artificial texts based on the calculation of measures of belonging to the invariants. SPIIRAS Proceedings 6(49), 104-121. (In Rus.) = Metod opredeleniya iskusstvennyh tekstov na osnove rascheta mery prinadlezhnosti k invariantam. Trudy SPIIRAN, N. 6. ,V. 49. - s. 104-121. DOI: 10.15622/sp.49.6.

[Potha and Stamatatos, 2019] Potha, N. and Stamatatos, E. (2019). Improving author verification based on topic modeling. *Journal of the Association for Information Science and Technology (JASIST),* October 2019, 10(70), 1074-1088. DOI: 10.1002/asi.24183.

[Koppel and Winter, 2014] Koppel, M., and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology,* 65(1), 178–187. DOI: 10.1002/asi.22954.

[Seidman, 2013] Seidman, S. (2013) Authorship verification using the impostors method // *In: CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.* 65(1), 178–187. DOI: 10.1002/asi.22954.

[Khonji and Iraqi, 2014] Khonji, M., and Iraqi, Y. (2014). A slightly-modified GI-based author-verifier with lots of features (ASGALF) // In: CLEF 2014 Labs and Workshops - Notebook Papers.

[Potha and Stamatatos, 2017] Potha, N. and Stamatatos, E. (2017). An Improved Impostors Method for Authorship Verification. CLEF 2017, LNCS 10456, pp. 138–144, 2017. DOI: 10.1007/978-3-319-65813-1_14.

[Germanovich, Melnikov and Khvostenko, 2017] Germanovich, A., Melnikov, S., and Khvostenko, V. (2017). On the selection of words sets characterizing the author's style of the Arabic text. Review of applied and industrial mathematics,4(24), 324-325. (In Rus.) = O vybore mnozhestva slov harakterizuyushchih avtorskij stil arabskogo teksta. Obozrenie prikladnoj i promyshlennoj matematiki, V.24, N.4, s.324-325.

[Belozerov etc., 2016] Belozerov, A., Vahlakov, D., Melnikov, S., Peresypkin, V., and Sidorov, E. (2016). Technological aspects of building a system for collecting and preprocessing news text corps for language models creation. *Izvestiya SFedU.* Engineering sciences 12(185), 29-42. (In Rus.) = Tekhnologicheskie aspekty postroeniya sistemy sbora i predobrabotki korpusov novostnyh tekstov dlya sozdaniya modelej yazyka. Izvestiya YuFU. Tekhnicheskie nauki, N.12, V.185. - s. 29-42. DOI: 10.18522/2311-3103-2016-12-2942.