Machine Learning Methods in the Problem of Attribution of Publicistic Texts of the XIX Century

Alexander Rogov, Nikolai Moskin, Kirill Kulakov
Petrozavodsk State University
Petrozavodsk, Russia
{rogov, moskin}@petrsu.ru, kulakov@cs.karelia.ru

Roman Abramov
ITMO University
Saint Petersburg, Russia
monset008@gmail.com

We consider in this work linguostatistical methods that were used for attribution (establishing authorship) of publicistic articles of the XIX century. At that time, F. M. Dostoevsky edited and headed three journals: "Time", "Epoch" and "Citizen", where there are about 500 unattributed texts. Samples from texts were compiled, their characteristics were studied, and a comparative analysis of the classification results based on various machine learning methods (decision trees, recurrent networks, parallel recurrent networks, transformer model) was carried out. The input of texts, their processing and the calculation of linguostatistical parameters were carried out using an updated version of the SMALT information system.

I. Introduction

One of the urgent tasks in the field of classification and clustering of objects is text attribution. This direction is well known in applied linguistics [1]. Similar problems have arisen long ago, starting from the problems of determining the authorship of the Old and New Testaments, the works of Plato, Aristotle, Homer, W. Shakespeare, etc. In Russia one of the first works (even before the advent of computer technology) belongs to N. A. Morozov, who wrote in 1915 the article "Linguistic spectra: a means for distinguishing plagiarism from the true works of one or another author. A stylistic etude" [2]. Later scientific works appeared that studied the authorship of well-known literary works by A. Pushkin, N. V. Gogol, M. A. Sholokhov and others [3, 4, 5].

In general, *attribution* in philology means the definition of the attributes of a text work (establishing authorship, time and place of creation, etc.). By conducting attribution of texts, it is possible to investigate various factors influencing the text [6]. For example, there are known studies of the gender attribution of texts (often they are manifested at the lexical level: some words are more common in texts written by women, others in texts by men) [7]. In [8] as markers of gender differences lexemes are studied that express various types of semantics of expressiveness:

- Components of rational and emotional assessment.
- Components of the intensity of the manifestation of the feature.
- Morphemic markers of emotional and expressive coloring.
- Graphic marking of the expression of an emotional attitude.

One of the possible areas of application of the results of such studies is forensic linguistics (for example, for conducting an author's expertise in court cases on the protection of copyright and related rights, protection of honor and dignity of citizens, protection of trademark rights, etc. [9]). The objects of author's expertise are not only ordinary texts, but also texts of mass media, Web-communications, as well as texts written in various programming languages [10].

Currently, an important task is to identify borrowings and plagiarism, which has become a widespread phenomenon in the field of science and education, including due to the rapid development of the Internet [11]. A close task is the identification of an artificially generated text [12]. For example, such texts can be created automatically by synonymization, when the text is formed by replacing individual lexemes with similar ones in meaning, using special dictionaries of synonyms [13].

Classical Russian philology also set itself tasks of this type. For example, the famous Russian linguist and literary critic, the founder of the largest scientific school in linguistics V. V. Vinogradov in [14] proposed a typology of text attribution factors, dividing them into subjective and objective. Moreover, all subjective factors, according to the scientist, should be carefully critically studied.

The focus of our work is on linguostatistical methods that were used to analyze publicistic articles of the XIX century [15]. During that period F. M. Dostoevsky edited and headed three journals that had a great influence on Russian social thought: «Time» (1861-1863), «Epoch» (1864-1865) and «Citizen» (1873-1874). They contain about 500 unattributed texts. Some of the articles were published anonymously (either without a signature, or under pseudonyms). The question of belonging them to Fedor Mikhailovich has been of interest to researchers since the beginning of the XX century and is still open, despite the fact that a lot of work has been done in this area [16, 17].

II. Works on related topics

First of all, we should mention one of the most detailed studies on the attribution of anonymous and pseudonymous publicistic articles from the journals «Time» and «Epoch», which was carried out by G. Kjetsaa [18]. However, the publication of his results caused many critical comments from

both philologists and mathematicians. In his research, the scientist identified 15 linguistic statistical parameters [18]:

- General distribution of parts of speech in the first two and in the last three positions of the sentence.
- Distribution of parts of speech in the first position of the sentence.
- Distribution of parts of speech in the second position of the sentence.
- Combination of parts of speech in the first two positions of the sentence.
- Distribution of parts of speech in the third position from the end of the sentence.
- Distribution of parts of speech in the penultimate position of the sentence.
- Distribution of parts of speech in the last position of the sentence.
- Combination of parts of speech in the last three positions of the sentence.
- Average word length in letters, calculated based on samples of 500 text words.
- General distribution of sentence length.
- Average length of a sentence in words, calculated on the basis of samples of 30 sentences.
- General distribution of the length of the sentence.
- Lexical spectrum of the text at the dictionary level.
- Lexical spectrum of the text at the text level.
- Vocabulary diversity index.

Thus, each text was assigned a vector of its characterizing numbers. However, the decision on whether the text belongs to F. M. Dostoevsky was made separately for each attribute, using methods of mathematical statistics. The main disadvantage of this study is the increased type II error.

Another study using mathematical methods was carried out under the supervision of M. A. Marusenko [19]. To determine the belonging of the articles to F. M. Dostoevsky and other authors, a set of 51 parameters was used, of which the most informative ones were selected. The purpose of this study was the attribution of twelve controversial articles from the list provided by G. Kjetsaa. A significant disadvantage of this study is the imbalance of the samples. It was expressed in the significant predominance of a group of texts by other authors over the texts of F. M. Dostoevsky. In this case, the F-measure was not calculated.

Currently, the following mathematical methods are used to establish authorship of works: neural networks, QSUM method, decision trees, support vector machine (SVM), k-means method, Bayesian classifier, Markov chains, principal component analysis, discriminant analysis, genetic algorithms, statistical criteria (Pearson's chi-square, Student's test, Kolmogorov-Smirnov's test) and others [20], [21], [22], [23], [24]. The entropy classification method (using various data compression algorithms) is also of interest, when a fragment of an unknown author is added to a text with a known author and how well this additional fragment is compressed is analyzed [25]. Where the compression ratio shows the best result, that class of documents is recognized as correct. Here a advantage is that there is no need for text pre-processing. We also note the following works on this topic [26], [27], [28], [29], [30], [31], [32], [33], [35].

III. Machine learning methods

Let's take a closer look at the three methods that were used in this study: decision tree, neural networks and transformer.

1. *Decision tree*. The simplicity and effectiveness of some machine learning models make them so popular for text attribution. The decision tree in this case is a powerful tool not only because of the ability to separate classes, but also the ability to interpret the model (which is important when explaining the results to philologists). In the study of text fragments *N*-grams (sequences of *N* elements - encoded parts of speech) were constructed, which proved to be good in problems of attribution [24, 36, 37, 38]. The frequencies of occurrence of *N*-grams were sent to the input of the model, which based on them determined the authorship of the text fragment.

2. *Neural networks*. The growth in the amount of data and the increase in computing power have led to rapid growth in the field of neural networks. Over the past decades, many architectures have been developed to deal with different types of data. Recurrent networks and their modifications are typical solutions for natural language processing problems. In this study, two different models and their modifications are considered: an ordinary recurrent network and a parallel recurrent network, constructed specifically for this problem. Neural networks also require transformation of the source texts. Text vectorization algorithms have gained popularity due to their simplicity and ability to capture the context of words. For this study, GloVe [39] was chosen, which can search for relationships between words. A vocabulary for pre-revolutionary writing was created from accessible texts before the training of models began. A recurrent network without modifications and with LSTM cells was chosen as models for the experiments.

3. *Parallel recurrent neural network* (PRNN) [40]. It was created to establish the fact of common authorship of two texts and showed good results in the original article. The model transforms two texts using a recurrent network into vectors, which are subsequently compared with seven different metrics. The results of the differences are sent to the input of the direct distribution network, the output of which is the probability of the source texts having a common author. For the current work, some details of the architecture have been changed taking into account the constraints of the task. The transformation of the model takes place through one network, the vocabulary does not change during training. This is due to the small size of the training sample, and without these innovations, the model quickly begins to retrain. We also tried LSTM cells instead of usual cells of the recurrent network.

4. *Transformer*. Over the past few years, a notable architectural advance in natural language processing has been the transformer model [41]. Such a simple, but at the same time effective solution has taken the NLP field to the next level. One of the main advantages is the ability to capture long-term dependencies in the data by processing the data in

its entirety, rather than sequentially, as happens in recurrent networks. Together with the ability to parallelize computations, this gives a tool that can learn from millions of texts and then generate new and meaningful ones [30]. The limitations in the number of texts for this study suggests the possibility of using a less complex model than the original transformer. The encoder unit, also described in [30], was chosen as such a model. The hyperparameters of the block were also changed: the size of the hidden layer was set to 32, and the number of heads in the attention block was reduced to two.

The main problem of using mathematical methods in attribution of texts is the lack of proof of their stability. As a result, these methods cannot be applied in wide practice, and attributed articles cannot be used in scientific circulation. The use of various techniques for attribution of texts in the case of unambiguous attribution increases confidence in the correct establishment of authorship. It requires the development of various attribution methods and proof of their applicability (i.e. searching for parameters when the results obtained are stable). One approach to solving this problem is described in this article.

IV. Software

The various methods of text attribution are based on a laborious procedure of formal-grammatical and syntactic analysis of the text content. As a rule, during the content analysis, each word is compared to a certain set of characteristics (attributes), for example, the paragraph number, the sentence number, the position of the word in the sentence, part of speech [18]. The process of comparing the characteristics to words is carried out by specialists-philologists. The result is a large array of data that can be processed manually or using software tools.

The SMALT information system («Statistical methods of analysis of literary texts») was developed at Petrozavodsk State University and is intended for organizing work on attribution of texts and carrying out research [42]. The main functions of the SMALT information system are:

- Import of texts with the selection of paragraphs, sentences and words.
- Verification and attribution of texts by specialists.
- Performing statistical analysis of text fragments.
- Presentation of the results of attribution of texts in the form of tables and reports.

The SMALT information system is implemented as a web application (url: http://smalt.karelia.ru/shower), located on the Internet, and has a high-level architecture shown in Fig. 1.

The core module contains the basic methods used in other modules, connection to databases, and basic text display functions. The SmaltPaper module implements the text management logic (displaying the text content, changing the description, etc.). The SmaltWord module contains the logic of working with words and the markup of words (search, editing, displaying forms, etc.). The SmaltUser module implements the logic of user management (registration, authorization, rights verification). Text import is implemented in the import module. Statistical analysis algorithms are implemented in the research module. The database is presented in the MySQL format and contains a dictionary and an information system corpus. The main task of the import function is to prepare the text for markup. During the import, the text is divided into components: sections, paragraphs, sentences and words. After that, the search for words in the dictionary of the SMALT information system is performed and the selection of frequently encountered markup options for each word is performed. To search for a word in the database, preliminary normalization is performed using stemming. For words that are missing in the dictionary, preparation for attribution is performed by creating a template.

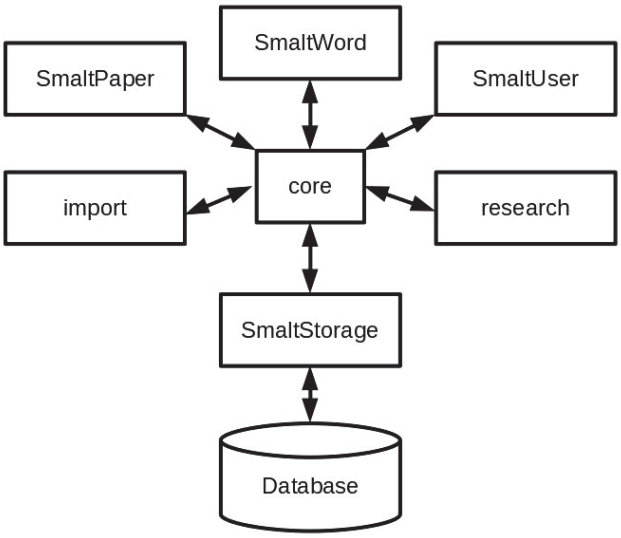

Fig. 1. High-level architecture of SMALT

As a result of the import, records are formed in the following tables:

- *Text*: general information about a text.
- *Word*: words of a text with an indication of the position in the text.
- *Entries*: text markup [43].

The text markup obtained during the import is not final and may contain inaccuracies [44]. Therefore, for further use, it is required to check and clarify the markup of the text with the help of a specialist philologist. The specialist has access to the following operations:

- Changing general information about the text.
- Changing the spelling of the word, the initial form of the word and/or the modern spelling.
- Changing the markup of the word.
- Replacing the markup of a word with another existing one.
- Deletion of a word or sentence
- Combining proposals.
- Adding a word.

Changes are tracked using entries in the log tables:

- *Log_text* - changes in general information about the text.
- *Log_word* - changes in text structure, word spelling, initial word form and/or modern spelling, used word markup.
- *Log_entries* - word markup changes.

The resulting markup can be used for statistical analysis and export to other programs and information systems. SMALT implements a number of statistical methods used in research [2, 43]:

- Text analysis by the strong graph method [45].
- Calculation of Kjetsaa metrics (including lexical spectra) [18].
- Search for n-grams (sequences of *n* elements - encoded parts of speech).

During the analysis of texts by the strong graph method, 2 or more texts are selected and the values for each pair of works are calculated using the Euclidean metric or the metric of city quarters. As a result of the algorithm, a table of measures of the proximity of texts to each other is formed. The calculation of Kjetsaa metrics and the search for n-grams is performed both for the entire text and for a fragment. The size of the fragment is determined in words. As a result of the algorithm, a set of calculated Kjetsaa metrics is formed for each fragment and the found n-grams are highlighted. To perform research, the SMALT information system allows you to get a representation of the results of text attribution in the form of tables and reports. The researcher can search for words, search for the use of word markup in texts, get the results of statistical analysis by the selected method in the form of html-pages. For a more detailed analysis, the SMALT information system allows you to upload text attribution and the results of calculating Kjetsaa metrics for a text in the form of Excel files.

V. Experiment data

As data for comparing various methods were used 20 articles of the magazines «Time» and «Epoch» (1861 - 1865), the authors of which are F. M. Dostoevsky, M. M. Dostoevsky, N. N. Strakhov, O. N. Shill, A. A. Grigoriev, A. U. Poretsky and Y. P. Polonsky. The list of reference texts was prepared based on philological analysis. Using the SMALT system, the following parts of speech were identified: noun, adjective, numeral, pronoun, verb, participle, gerund, preposition, conjunction, modal word, interjection, onomatopoeic word, adverb, category of state, particle, foreign word, quote, introductory word, old slavicism, part of phraseology, a non-linguistic symbol, an abbreviated word, part of a polynomial name.

Each article was divided into parts with a length of 1000 words. But the limited number of such fragments forces us to resort to data expansion techniques. A sliding window was used as a similar technique. The step of the sliding window is the number of words from the beginning of the current and the next fragment (Fig. 2). Note that earlier G. Kjetsaa used disjoint fragments [18]. The size of the sliding window and its step have a significant impact on the obtained result; therefore, finding their optimal parameters is a separate task.

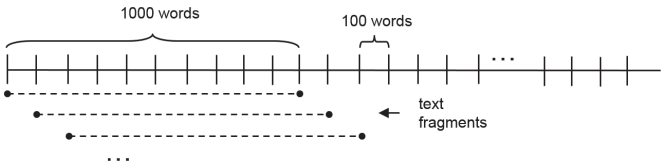

Fig. 2. Dividing texts into fragments for analysis with a step of 100 words

For the training and test samples, 18 and 2 articles were used, respectively. For the test sample, the articles of two authors were selected with a size of approximately 10000 words, which divides the number of fragments used for the training and test set in the ratio of 80% by 20%. One of the authors was not represented in the training set, which makes the task even more difficult.

The statement of the research task is to determine the texts under the authorship of F. M. Dostoevsky, which is actually a binary classification task. The author of interest was taken for the first class, and all the others were assigned to the second class. The first experiments were conducted for the decision tree model. Different step sizes were used to form a text fragment and the length of an *N*-gram. It is assumed that with the growth of *n*, the model will be able to better determine the authorship of the fragment due to more specific information. The step size increases the sample for training, which can have a positive effect on the metrics of the model.

The next stage was the use of models of recurrent neural networks. The starting point was the preparation of a word2vec dictionary – the available corpus of pre-revolutionary texts was trained using the Stanford GloVe model for 15 iterations and with the preservation of all the encountered words. Then the text fragments were transformed using a dictionary and sent to the input models. Experiments were carried out for different step sizes. The latest model is a transformer. It also works with a vector representation of words, but instead of a pre-radiated dictionary, a dictionary that is trained together with a transformer was chosen, which allows the model to be more flexible.

VI. Characteristics of lexical spectra

As it was shown earlier, lexical spectrum (at the dictionary level *f* and at the text level *mf*) are significant characteristics for text attribution [18]. However, the use of some machine learning methods (for example, decision trees) requires the representation of the spectrum as a single number that would adequately reflect its structure. The approximation of lexical spectrum by three types of curves was considered, as a result of which two characteristics for each curve are obtained (they were found using the least squares method after preliminary data normalization):

- Hyperbolic curve $y = a/x+b$.
- Exponential curve $y = c \cdot e^{-\lambda x}$.
- Power curve $y = p \cdot x^n$.

Based on the material of 149 text fragments from the articles of the pre-revolutionary magazine «Time» (1861-1863), it was shown that the coefficients of hyperbolic regression approximate the data much better than the

coefficients of the exponential and power curve. To do this, the Pearson coefficients were found between the modules of the difference of the regression coefficients (for example, $d^1_{ij} = |a_i - a_j|$) and the distance χ^2 between the diagrams (d^2_{ij}). Using z_{ij}, we denote j value of the spectrum for the text fragment i, if $j < 10$, and the sum of the remaining values, if $j = 10$. To calculate the distance χ^2, the following formula was used [46]:

$$d_{ij}^2 = n_{ij}\left(\sum_{k=1}^{10}\left(\frac{z_{ik}^2}{z_i(z_{ik}+z_{jk})}+\frac{z_{jk}^2}{z_j(z_{ik}+z_{jk})}\right)-1\right),$$

where $n_{ij} = \sum_{k=1}^{10}(z_{ik}+z_{jk})$, $z_i = \sum_{k=1}^{10} z_{ik}$ $(i, j = 1, 2, ..., 149)$.

Table I shows the obtained Pearson correlation coefficients for each of the six regression coefficients (here the text fragments consisted of 500 words). All values showed statistical significance at the level of 0,05. According to the Chaddock scale, the relationship with the coefficients λ, c, n, p is weak (unlike a and b), so a hyperbolic curve is better suited for data approximation.

TABLE I. PEARSON CORRELATION COEFFICIENTS

Coefficients	Lexical spectrum at the vocabulary level (*f*)	Lexical spectrum at the text level (*mf*)
a	0,681808	0,757219
b	0,681808	0,757219
λ	0,206101	0,168497
c	0,18129	0,082559
n	0,199566	0,162318
p	0,150804	0,123265

VII. COMPARATIVE ANALYSIS OF MODELS

The experiments were carried out for different hyperparameters of the models. In addition to the common *step* – step of the sliding window, there were also unique ones for each model. For a decision tree this was the N – number of a sequence of elements. The models were compared using the accuracy metric:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN},$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative outcomes.

TABLE II. EXPERIMENTAL RESULTS FOR DIFFERENT MODELS

Model	Accuracy
Tree (step = 100, N = 2)	71%
Tree (step = 500, N = 2)	76%
Tree (step = 1000, N = 2)	62%
Tree (step = 100, N = 3)	43%
Tree (step = 500, N = 3)	53%
Tree (step = 1000, N = 3)	67%
RNN (step = 100)	80%
RNN (step = 1000)	67%
LSTM (step = 100)	89%
LSTM (step = 1000)	89%
PRNN (step = 1000)	75%
Transformer (step = 100)	98%
Transformer (step = 500)	100%
Transformer (step = 1000)	98%

The decrease in accuracy with an increase in the splitting step for some models can be explained by a decrease in the amount of data for training. The decision tree showed comparable results, reaching 76% accuracy. Its essential advantage for wide application is the possibility of interpreting the obtained results. In addition, it allows you to work with texts of unproductive authors and avoid retraining. The use of other machine learning methods (recurrent networks and parallel recurrent networks) showed results comparable to decision trees. Their advantage is that there is no need for manual markup of the text, but when trying to explain the model, difficulties arise (networks do not allow you to clearly demonstrate the characteristics for attribution without which philologists do not trust the results). The transformer model showed the highest accuracy, but it requires a large amount of data for training and also there are difficulties in interpreting the results. Note that the limits of applicability of the transformer model require significantly more research and this analysis is planned in the future.

VIII. APPLICATION OF DECISION TREES IN TEXT ATTRIBUTION

When solving the problem of text attribution, the problem arises of determining the author's style of a writer who has created a smaller number of texts (both quantitatively and in terms of the total volume of words) in comparison with other authors. Possible solutions to this problem were considered using the example of determining the style of Apollon Grigoriev [47]. A data set for training was compiled (118 fragments – Apollon Grigoriev, 899 – the rest). The fragment size is 1000 words, step – 100. The texts from which the data were prepared are presented in Table III. In this case, fragments of texts by Apollon Grigoriev are the objects of the minority class, and all the others are of the majority class.

TABLE III. THE STUDIED TEXTS

Name	Author
Pis'mo k redaktoru	Polonsky Y. P.
Zhukovskij i romantizm	Fyodor Dostoevsky
Literaturnaya isterika	Fyodor Dostoevsky
Odin iz proektov chudesnago obogashcheniya Rossii	Schill I. N.
Pis'mo k izdatelyu "Vremeni"	Polonsky Y. P.
Podpiska na 1863 god	Mikhail Dostoevsky
Ryad statej o russkoj literature. Vvedenie	Fyodor Dostoevsky
Slavyanofily, chernogorcy i zapadniki	Fyodor Dostoevsky
Ryad statej... G. -bov i vopros ob iskusstve	Fyodor Dostoevsky
Knizhnost' i gramotnost'. Stat'ya pervaya	Fyodor Dostoevsky
Knizhnost' i gramotnost'. Stat'ya vtoraya	Fyodor Dostoevsky
Poslednie literaturnye yavleniya. Gazeta "Den"	Fyodor Dostoevsky
Neobhodimoe literaturnoe ob'yasnenie, po povodu ra…	Fyodor Dostoevsky
Politicheskoe obozrenie	Golovachev A. A.
Lermontov i ego napravlenie. Stat'ya vtoraya	Apollon Grigoriev
Oppoziciya zastoya. Cherty iz istorii mrakobesiya	Apollon Grigoriev
Nashi domashnie dela 12.1864	Poretsky A. U.
DURNYE PRIZNAKI	Strakhov N. N.
ESHCHE O PETERBURGSKOJ LITERATURE	Strakhov N. N.
Vsyo-li na Rusi tak ploho, kak kazhetsya?	Meshchersky V. P.

As a method of constructing an ensemble of classifiers, bootstrap aggregating was used in the work. As a result of calculations, it turned out that the relative frequency of the bigram «particle-adjective» greater than 6,5 is a distinctive feature of the publicistic style of Apollon Grigoriev.

A study of the article «Poems of A. S. Khomyakov» was conducted, which confirms the earlier conclusion that there is no reason to consider it as belonging to Apollon Grigoriev [47]. It should be noted that for the first time information about the affiliation of this article to F. M. Dostoevsky using the method of G. Kjetsaa was obtained in [45]. However, this result caused even greater distrust among philologists of mathematical methods of text attribution. Only 15 years later, this fact was confirmed by other methods and introduced into scientific circulation.

One of the possible directions related to the text attribution is the analysis of strong positions in the text. Studying them, it is possible to identify the introduction of editorial changes in the texts of the original authors. The analysis of 19 texts was carried out (among the authors: F. M. Dostoevsky, M. M. Dostoevsky, N. N. Strakhov, A. A. Golovachev, I. N. Shill, A. Grigoriev, A. U. Poretsky, Ya. P. Polonsky). 12 of them belong to F. M. Dostoevsky, the rest – to other authors. The analysis was carried out on text fragments of 10, 20 and 30 sentences in size, sentences of less than 3 words were discarded. To increase the sample size, a step equal to 5 sentences was used to count the beginning of the next fragment. The following features were used: the frequency of the part of speech located at the 1st, 2nd and 3rd place of the sentence, a bigram at the 1st and 2nd place of the sentence and a trigram (1st, 2nd and 3rd place of the sentence). There are only 5 types of features. Analyzing the results of the experiment, you can see that fragments of 30 sentences are more informative. The most informative feature was «the frequency of occurrence of a noun in the 3rd place of a sentence». A similar study was conducted on the last words of the sentence. They were less informative. It should be noted that the obtained results were highly appreciated by philologists.

TABLE IV. FRAGMENTS OF 30 SENTENCES

Depth	Accuracy (by the maximum number of fragments)	Accuracy of the combined sample (by fragments)
1	0,7632	74%
2	0,8094	84%
3	0,8443	89%
10	0,9858	100%

IX. CONCLUSION

The paper presents the results of authors on the attribution of publicistic articles of the XIX century. They were performed using the information system SMALT, which has long been positively proven in the study of the creative heritage of such writers as F. M. Dostoevsky, A. Grigoriev, etc. To establish the authorship of a number of works, machine learning methods were used. Experiments have shown that the accuracy of classification of recurrent networks and parallel recurrent networks is comparable to decision trees. The transformer model showed the highest accuracy, but it requires a large amount of data for training and is difficult for philologists to interpret. The article also shows that the coefficients of hyperbolic regression approximate the data of lexical spectra much better than the coefficients of the exponential and power curve.

The stability of the method is achieved by using a larger number of text fragments. Each article was divided into parts with a length of 1000 words. Unlike G. Kjetsaa, who used non-overlapping fragments, this study used the sliding window technique, which allowed increasing the number of analyzed fragments. It allowed the use of machine learning methods. As a result of experiments, it was found that the most stable size for building a decision tree is a fragment size of 1000 words and a step of 100 words. As a result of using the decision tree, interesting results were obtained. For example, the relative frequency of the bigram «particle-adjective» greater than 6.5 is a distinctive feature of the publicistic style of Apollon Grigoriev. The study of strong positions in the texts confirms the early hypothesis about the introduction of edits by F. M. Dostoevsky, who was the editor of the journals «Time», «Epoch» and «Citizen», in the texts of articles by other authors. The task of searching for other heterogeneous fragments in the articles of the XIX century (i.e. fragments that differ significantly from the rest of the text in terms of a set of characteristics) seems promising. All this proves the importance of an integrated approach in such studies, when mathematical methods complement (confirm or refute) traditional linguistic analysis. The obtained results were presented for further consideration to specialists of the Institute of Philology of Petrozavodsk State University. More details about the results of the experiments can be found on the website: http://smalt.karelia.ru/.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research, project no. 18-012-90026. The authors also thank Petrozavodsk State University for its support.

REFERENCES

[1] A.A. Lebedev, *Introduction to applied linguistics*. Petrozavodsk: PetrSU Publ., 2019.

[2] A.A. Rogov, A.V. Sedov, Y.V. Sidorov and T.G. Surovceva, *Mathematical methods for text attribution*. Petrozavodsk: PetrSU Publ., 2014.

[3] A.A. Markov, “About one application of the statistical method”, *Bulletin of the Imperial Academy of Sciences*, Series 6, No. 4, 1916, pp. 239-242.

[4] N.P. Velikanova, B.V. Orehov, “Digital textology: text attribution on the example of M. A. Sholokhov’s novel "Quiet Don"”, *World of Sholokhov*, No. 1, 2019, pp. 70-82.

[5] D.V. Khmelev, “Recognition of the author of the text using A. A. Markov’s chains”, *Moscow University Bulletin*, *Series 9: Philology*, No. 2, 2000, pp. 115-126.

[6] M.B. Malutov, “Review of methods and examples of text attribution”, *Review of Applied and Industrial Mathematics*, vol. 12(1), 2005, pp. 41-78.

[7] A.P. Vasilevich, M.M. Mamaev, “The problem of identifying gender-significant parameters of the text”, *Bulletin of the Moscow State Regional University, Linguistics series*, No. 2, 2014, pp. 17-24.

[8] A.A. Stepanenko, Z.I. Rezanova, “Expressiveness as a marker of gender differences in computer communication (to the problem of automatic gender attribution of text)”, *Bulletin of Tomsk State University*, No. 433, 2018, pp. 38-46.

[9] A.Y. Khomenko, “Algorithm for automatic identification of the author of a written speech work in judicial authorship”, *Legal Linguistics*, vol. 3 (14), 2014, pp. 83-93.

[10] T.N. Romanchenko, “Methods of attribution in the author’s expertise”, *Bulletin of the Saratov State Law Academy*, vol. 2(91). 2013, pp. 228-233.

[11] A.V. Nikitov, O.A. Orchakov, Y.V. Chekhov, “Plagiarism in the works of students and postgraduates: the problem and methods of counteraction”, *University management: practice and analysis*, vol. 5(81), 2012, pp. 61-68.

[12] A.O. Iskhakova, *Method and software tool for determining artificially created texts*. Tomsk: Tomsk State University of Control Systems and Radio Electronics, 2016.

[13] A.O. Shumskaya, “Method for determining artificial texts on the basis of calculating the measure of belonging to invariants”, *Proceedings of SPIIRAS*, vol. 6 (49), 2016, pp. 104-121.

[14] V.V. Vinogradov, *The problem of authorship and the theory of styles*. Moscow: State Publishing House of Fiction, 1961.

[15] A.A. Rogov, N.D. Moskin, R.V. Abramov, K.A. Kulakov, “Possibilities of using decision trees in the problem of attribution of publicistic texts of the 19th century”, *Abstracts of the 13th International Conference "Intellectualization of Information Processing",* 2020, pp. 336-340.

[16] L.V. Alekseeva, “Problems of attribution in studies about F. M. Dostoevsky (a review of the proposed solutions), *Unknown Dostoevsky*, No. 4, 2015, pp. 3-10.

[17] E.I. Gurova, “Methods of attribution of authorship in modern Russian philology”, *New philological bulletin*, vol. 3 (38), 2016, pp. 29-44.

[18] G. Kjetsaa, *Attributed to Dostoevsky: The Problem of attributing to Dostoevsky anonymous articles in Time and Epoch*. Oslo: Solum Forlag A. S., 1986.

[19] M.A. Marusenko, E.S. Rodionova, E.E. Melnikova, “About the authorship of anonymous and pseudonymous articles attributed to F. M. Dostoevsky (magazines "Time" and "Epoch", 1861-1865)”, *Bulletin of St. Petersburg university. Series 9. Philology. Oriental studies. Journalism*, No. 3, part 1, 2008, pp. 48-56.

[20] R.V. Meshcheryakov, A.S. Romanov, “Identification of the author of the text using the support vector machine in the case of two possible alternatives”, *Computational linguistics and intellectual technologies: materials of the international conference "Dialog-2009"*, vol. 8(15), 2009, pp. 432-437.

[21] A.S. Romanov, *Methodology and software package for identifying the author of an unknown text*. Tomsk, 2010.

[22] J. Calle-Martin, A. Miranda-Garcia, “Stylometry and Authorship Attribution: Introduction to the Special Issue”, *English Studies*, vol. 93(3), 2012, pp. 251-258.

[23] J.M. Farringdon, A.Q. Morton, M.G. Farringdon, M.D. Baker, *Analyzing for Authorship*. Cardiff: University of Wales Press, 1996.

[24] E. Stamatatos, “A Survey of Modern Authorship Attribution Methods”, *Journal of the American Society for Information Science and Technology*, vol. 6(3), 2009, pp. 538-556.

[25] T.V. Batura, “Formal methods for determining the authorship of texts”, *Bulletin of the Novosibirsk State University. Series "Information Technologies"*, vol. 10(4), 2012, pp. 81-94.

[26] H. Ramnial, S. Panchoo, S. Pudaruth, “Authorship attribution using stylometry and machine learning techniques”, *Intelligent Systems Technologies and Applications*, 2016, pp. 113-125.

[27] S. Petrovic, G. Berton, S. Campbell, L. Ivanov, “Attribution of 18th century political writings using machine learning”, *Journal of Technologies in Society*, vol. 11(3), 2015, pp. 1-13.

[28] U. Stańczyk, K. A. Cyran, “Machine learning approach to authorship attribution of literary texts”, *International journal of applied mathematics and informatics*, vol. 1(4), 2007, pp. 151-158.

[29] P. Shrestha, S. Sierra, F. A. Gonzalez, M. Montes, “Convolutional neural networks for authorship attribution of short texts”, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, 2017, pp. 669-674.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, “Language models are unsupervised multitask learners”, *OpenAI Blog*, vol. 1(8), 2019, 9 p.

[31] I. Sutskever, O. Vinyals, Q. V. Le, “Sequence to sequence learning with neural networks”, *Advances in neural information processing systems*, 2014, pp. 3104-3112.

[32] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, “End-to-end memory networks”, *Advances in neural information processing systems*, 2015, pp. 2440-2448.

[33] N. Schaetti, R. Emile-Argand, *Author Verification in Stream of Text with Echo State Network-based Recurrent Neural Models*. SwissText, 2019.

[34] M. Kestemont, E. Stamatatos, E. Manjavacas, W. Daelemans, M. Potthast, B. Stein, “Overview of the Cross-domain Authorship Attribution Task at PAN 2019”, *CLEF (Working Notes)*, 2019.

[35] E. Stamatatos, F. Rangel, M. Tschuggnall, B. Stein, M. Kestemont, P. Rosso, M. Potthast, “Overview of PAN 2018. Author Identification, Author Profiling, and Author Obfuscation”, *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, Cham, 2018, pp. 267-285.

[36] A.A. Kotov, Z.I. Mineeva, A.A. Rogov, A.V. Sedov, Y.V. Sidorov, *Linguistic Corpuses*. Petrozavodsk: PetrSU Publ., 2014.

[37] V. Kešelj, F. Peng, N. Cercone, C. Thomas, “N-gram-based author profiles for authorship attribution”, *Proceedings of the conference pacific association for computational linguistics PACLING-2003*, vol. 3, 2003, pp. 255-264.

[38] A. Zečević, “N-gram based text classification according to authorship”, *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, 2011, pp. 145-149.

[39] J. Pennington, R. Socher, C. D. Manning, “Glove: Global vectors for word representation”, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[40] M. Hosseinia, A. Mukherjee, “Experiments with neural networks for small and large scale authorship verification”, *arXiv preprint arXiv:1803.06456*, 2018.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need”, *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[42] A.A. Rogov, K.A. Kulakov, N.D. Moskin, “Software support in solving the problem of text attribution”, *Software Engineering*, vol. 10(5), 2019, pp. 234–240.

[43] K.A. Kulakov, A.A. Lebedev, A.A. Rogov, T.G. Surovtsova, N.D. Moskin, “Attribution of texts using mathematical methods and computer technologies”, *Materials of the XIII conference «Digital technologies in education, science, society»*, 2019, pp. 121-125.

[44] A.A. Lebedev, A.A. Rogov, K.A. Kulakov, N.D. Moskin, “On the problem of creating marked-up text corpora in 19th century graphics”, *Proceedings of the international conference "Corpus linguistics-2019"*, 2019, pp. 296-302.

[45] Y.V. Sidorov, *Mathematical and information support of literary texts processing methods based on formal grammatical parameters*. Petrozavodsk: PetrSU Publ., 2002.

[46] S.A. Ayvazyan, I.S. Enyukov, L.D. Meshalkin, *Applied statistics: foundations of modeling and primary data processing.* Moscow: Finance and Statistics, 1983.

[47] A.A. Rogov, R.V. Abramov, A.A. Lebedev, K.A. Kulakov, N.D. Moskin, “Text Attribution in Case of Sampling Imbalance by the Method of Constructing an Ensemble of Classifiers Based on Decision Trees”, *Data Analytics and Management in Data Intensive Domains: XXII International Conference DAMDID/RCDL'2020: Extended Abstracts of the Conference*, 2020, pp. 185-188.