ИНФОРМАТИКА

UDC 004.8
MSC 68T50

Research of features of Dostoevsky's publicistic style by using n-grams based on the materials of the "Time" and "Epoch" magazines*

R. V. Abramov, K. A. Kulakov, A. A. Lebedev, N. D. Moskin, A. A. Rogov

Petrozavodsk State University, 33, pr. Lenina, Petrozavodsk,
185910, Russian Federation

For citation: Abramov R. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Rogov A. A. Research of features of Dostoevsky's publicistic style by using n-grams based on the materials of the "Time" and "Epoch" magazines. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2021, vol. 17, iss. 4, pp. 389–396.
https://doi.org/10.21638/11701/spbu10.2021.407

The paper is devoted to the study of the publicity style of F. M. Dostoevsky on the basis of publications in the journals "Time" and "Epoch" (1861–1865). For this, fragments of texts (including other authors: M. M. Dostoevsky, N. N. Strakhov, A. A. Golovachev, etc.) were selected in sizes of 500, 700 and 1000 words, on which the occurrence of bigrams and trigrams (encoded sequences of parts of speech) were counted. Decision trees were built on their basis and an analysis of the accuracy of text recognition was performed. If we consider the class cation at the rest level of the tree (fragment size 1000), then the accuracy was on average 87 resulting decision trees.

Keywords: publicity style, text attribution, decision tree, n-gram, F. M. Dostoevsky, information system "Statistical methods for analyzing literary texts", tree matching.

1. Introduction. The paper is devoted to the research of the publicistic style of F. M. Dostoevsky based on articles in the magazines "Time" and "Epoch" (1861–1865). The list of studied texts, the authors of which in addition to F. M. Dostoevsky are M. M. Dostoevsky, N. N. Strakhov, A. A. Golovachev, I. N. Shill, A. Grigoriev, A. U. Poretsky and Ya. P. Polonsky, is presented in Table 1. There are 19 texts in total, 12 of them belong to F. M. Dostoevsky and 7 to other authors. The size of a single text is at least 1300 words. Note that some other articles from these journals still do not have an author, therefore, this work should be considered in the context of a broader problem, namely attribution of texts (establishment of authorship of anonymous texts). This problem, which has been declared in the philological community for a long time, is still far from being solved, and the ownership of certain articles by F. M. Dostoevsky continues to cause lively

* This work was supported by the Russian Foundation for Basic Research (project N 18-012-90026).

discussions among experts. It is for this reason that there is a need to use methods that are not quite traditional for literary studies as a scientific field for consideration of the required issues [1].

Mathematical methods and computer technologies can be used to solve this problem [2–8]. Note some authors who used mathematical methods to solve the problem of text attribution: A. Q. Morton, T. C. Mendenhall, J. M. Farringdon, B. Efron, R. Thisted, W. J. Teahan, C. E. Chaski, E. Stamatatos, P. Juola, R. D. Peng, T. Joachims, J. J. Diederich, C. Apte, D. Lowe, R. Matthews, F. J. Tweedie, O. de Vel, S. Argamon, S. Levitan, R. Zheng. Among other methods of data mining, decision trees are distinguished by the fact that they are easy to understand and interpret and also do not require special preliminary data processing. The specificity of this study is that the analyzed texts are presented in a pre-reform orthography, which makes it necessary to create new or modernize existing approaches. By itself, the orthography in the middle of the 19^{th} century implies a certain variability of spellings, which complicates the automatic processing of sources. But at the same time it better allows to trace the features of the individual author's style, demonstrated in the construction of the text at its different levels.

Usually "author's style" means three levels: syntactic, lexical-phraseological and stylistic, although texts can also be studied at the punctuation and orthography levels. Our proposed analysis is based on taking into account grammatical and morphological features (namely, the distribution of parts of speech within the analyzed texts), which should also be considered a distinguishing feature of the work of a particular author.

2. Research methods (description of the experiment). In our research of texts n-grams (sequences of n elements — encoded parts of speech) were constructed, which proved themselves well in solving problems of determining authorship. For example, in [9] it is shown how to identify non-uniform fragments in the text using 4-grams and counting χ^2 statistics. In [10] a nonlinear method for constructing n-grams based on syntactic dependency trees using the Stanford analyzer is considered.

In our case the SMALT information system ("Statistical methods for analyzing literary texts") developed at the Petrozavodsk State University was used [11]. Specialists in philology carried out grammatical markup of texts, which took into account 14 parts of speech (noun, adjective, numeral, pronoun, adverb, category of state, verb, participle, gerund, preposition, conjunction, particle, modal word, interjection). It also made possible to mark the encountered words as quotes, foreign words, introductory words, abbreviated words and non-linguistic symbols.

The texts were analyzed using bigrams as follows: the fragment size (500, 700 and 1000 words) and the step (100, 200 words, etc.) were allocated to select the beginning of the next fragment. Then the frequency of occurrence of n-grams in texts was calculated.

The constructed fragments were analyzed using the decision tree. At the same time, the texts of F. M. Dostoevsky were considered against other authors. In this context, we can mention the well-known machine learning algorithm called random forest, which was proposed in 2001 [12].

A decision tree is an acyclic graph that classifies objects (in our case, texts) described by a set of features. Each node in the tree contains a branching condition for one of the attributes. In the classification process sequential transitions are performed from one node to another in accordance with the values of the object's attributes (in practice binary trees are usually used).

As a result of their consideration, it turned out that a tree of depth 3 is sufficient to determine the morphological features of the publicistic style of F. M. Dostoevsky and

Table 1. **Source texts for analysis**

Code	Name	Author	Journal	Year	Number of	
					journal	words
2	Fires	F. M. Dostoevsky	Time	1862	1	1943
11	Taras Shevchenko	A. Grigoriev	Time	1861	4	1724
13	Letter to the editor	Y. P. Polonsky	Time	1863	3	2303
34	Literary hysteria	F. M. Dostoevsky	Time	1861	7	2808
35	Young pen	F. M. Dostoevsky	Time	1863	2	1872
40	Subscription for 1863 year	M. M. Dostoevsky	Time	1863	1	2541
42	A number of articles about Russian literature. Introduction	F. M. Dostoevsky	Time	1861	1	12508
43	Slavophiles, montenegrins and westernizers	F. M. Dostoevsky	Time	1862	9	2058
75	A number of articles about Russian literature. G.-bov and the question of art	F. M. Dostoevsky	Time	1861	2	11053
77	Bookness and literacy. Article two	F. M. Dostoevsky	Time	1861	8	14210
78	Recent literary phenomena. The newspaper "Day"	F. M. Dostoevsky	Time	1861	11	4323
82	The necessary literary explanation about the work...	F. M. Dostoevsky	Time	1863	1	3680
86	To finish. The last explanation from "Modern..."	F. M. Dostoevsky	Epoch	1864	9	1378
87	Political review	A. A. Golovachev	Epoch	1864	8	10309
89	Lermontov and his direction. Article two	A. Grigoriev	Time	1862	11	7480
92	Our household chores	A. U. Poretsky	Epoch	1864	12	8602
96	Voice for the Petersburg Don Quixote (Regarding the articles by G. Teatrin)	F. M. Dostoevsky	Time	1862	10	1334
97	Remark	F. M. Dostoevsky	Epoch	1864	9	1599
116	Bad signs	N. N. Strakhov	Time	1862	11	6331

other writers (Figure, *a*, *b* show fragments of graphs obtained as a result of calculations with a step of 100 words and a fragment size of 1000 words for bigram and trigram).

3. Tree comparison. As a result of the analysis of different text fragments, we get a set of decision trees $G_1, G_2, \ldots, G_n$ that differ from each other. At the same time, they are united by the fact that they are oriented binary trees, where each vertex is associated with a certain bigram or trigram (for example, at the root vertex you can observe "Adjective-Noun" or "Noun-Adjective-Noun"), threshold value and Gini index. To research the features of Dostoevsky's publicistic style it is important to understand how text fragments and decision trees derived from them are related, i. e. to switch from abstract models to their philological interpretation. For this we suggest to carry out a comparative analysis of trees $G_1, G_2, \ldots, G_n$ (this direction is known as *graph matching*). You can establish a measure of similarity $d(G_i, G_j)$ on a set of trees (in particular *tree edit distance* [13]). After the

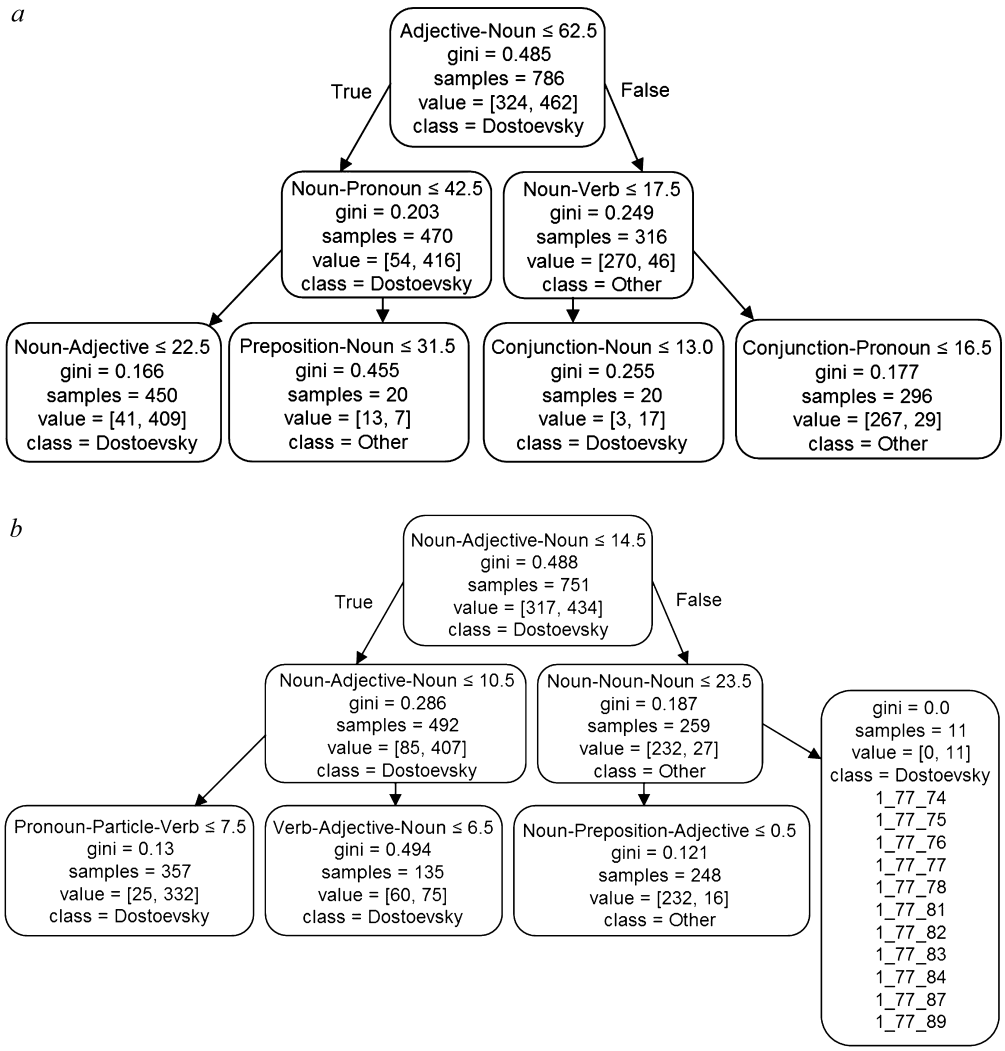

Figure. Decision tree (step 100 words, fragment 1000 words)
a — bigram; *b* — trigram.

necessary calculations a matrix of distances between trees is obtained. After classifying the decision trees you can trace how much the distance between the trees is related to the conditions for the generating of text fragments.

The first experiments in this direction were carried out using the information system "Folklore", which was previously used to solve the problem of recognizing folklore and author's texts [14]. They showed that the trees obtained from close attributes are similar to each other (the differences are manifested in the threshold value for the same features). At this moment we can say that this method is quite stable. However, additional experiments are required to get recommendations for choosing the step size and fragment size.

4. Conclusion. Among the experiments the weakest results were shown by decision trees with a fragment size of 500 words (Tables 2, 3). On decision trees with a depth 3 the accuracy was always less than 89 %. The most stable results showed fragments of

Table 2. **Classification accuracy for different fragments**

Depth of the tree	Size of sliding window	Number of fragments	Fragment size	Accuracy, %
1	100	816	500	79.0
2	100	816	500	79.0
3	100	816	500	83.0
1	100	790	700	83.0
2	100	790	700	84.0
3	100	790	700	88.0
1	100	751	1000	87.0
2	100	751	1000	90.0
3	100	751	1000	91.0
1	200	405	500	77.0
2	200	405	500	79.0
3	200	405	500	86.0
1	200	392	700	82.0
2	200	392	700	83.0
3	200	392	700	89.0
1	200	372	1000	88.0
2	200	372	1000	91.0
3	200	372	1000	93.0
1	300	268	500	80.0
2	300	268	500	80.0
3	300	268	500	86.0
1	300	260	700	82.0
2	300	260	700	83.0
3	300	260	700	88.0
1	300	247	1000	87.0
2	300	247	1000	91.0
3	300	247	1000	91.0
1	400	199	500	81.0
2	400	199	500	81.0
3	400	199	500	88.0
1	400	193	700	81.0
2	400	193	700	86.0
3	400	193	700	91.0
1	400	183	1000	87.0
2	400	183	1000	88.0
3	400	183	1000	93.0
1	500	158	500	81.0
2	500	158	500	81.0
3	500	158	500	89.0
1	500	153	700	84.0
2	500	153	700	85.0
3	500	153	700	91.0

1000 words (the most frequency accuracy is 93 %). The step size significantly influenced the accuracy, however, this is rather due to the number of analyzed fragments.

So with a step of 1000 words and a fragment size of 1000 words we get text splitting without intersections and only 70 fragments, which are easier to divide into 8 groups than 751 fragments into the same 8 groups (step 100 words, fragment size 1000 words). In the first case we get an accuracy of 97 %, in the second 91 %. In addition, this suggests that the obtained features predominate in the size of the article, however, they may be unstable in separated fragments. The proposed method allows them to be identified and subjected to additional analysis.

If we consider the classification at the first level of the tree (fragment size 1000), then the accuracy was on average 87 %. This feature is the percentage of the presence of

Table 3. **Classification accuracy for different fragments (continuation)**

Depth of the tree	Size of sliding window	Number of fragments	Fragment size	Accuracy, %
1	500	145	1000	87.0
2	500	145	1000	91.0
3	500	145	1000	93.0
1	600	127	700	81.0
2	600	127	700	82.0
3	600	127	700	90.0
1	600	120	1000	88.0
2	600	120	1000	89.0
3	600	120	1000	93.0
1	700	107	700	84.0
2	700	107	700	86.0
3	700	107	700	96.0
1	700	101	1000	85.0
2	700	101	1000	88.0
3	700	101	1000	93.0
1	800	88	1000	89.0
2	800	88	1000	91.0
3	800	88	1000	97.0
1	900	77	1000	87.0
2	900	77	1000	91.0
3	900	77	1000	95.0
1	1000	70	1000	89.0
2	1000	70	1000	93.0
3	1000	70	1000	97.0

Table 4. **The value of the Gini index for various features at the first level**

Gini index	Feature	Gini index	Feature
0.1769	Adjective-Noun	0.18322343	Noun-Adjective-Noun
0.1659	Noun-Adjective	0.120744636	Participle-Noun-Adjective
0.1577	Participle-Noun	0.114925099	Noun-Noun-Adjective
0.1212	Noun-Preposition	0.113225856	Preposition-Participle-Noun
0.0953	Verb-Participle	0.111462412	Adjective-Noun-Preposition
0.0946	Noun-Noun	0.108502517	Adjective-Noun-Verb
0.0911	Conjunction-Particle	0.106896858	Preposition-Noun-Adjective
0.0898	Quote-Quote	0.105382403	Noun-Verb-Adjective
0.0898	Noun-Quote	0.098399618	Noun-Verb-Participle
0.0898	Pronoun-Conjunction	0.098296557	Noun-Preposition-Noun

the "Adjective-Noun" bigram. Table 4 provides information about the value of the Gini index for bigrams and trigrams (step 100 words, fragment size 1000 words) for various features at the first level. In fragments of texts belonging to F. M. Dostoevsky quite often the percentage of presence of the bigram "Adjective-Noun" is less than 62.5 % (step — 100, fragment — 1000, accuracy — 90 % for fragments belonging to F. M. Dostoevsky and 83 % for other authors). When constructing decision trees for other steps, the value of this attribute changed insignificantly. In the second and subsequent steps the decision trees used different bigrams after changing the step. This indicates their less stability for solving the classification problem.

Note that among the incorrectly classified fragments of the articles that do not belong to F. M. Dostoevsky, the first and last fragments predominated, which may indicate the editorial correction of Fedor Mikhailovich.

Another trend was observed in the analysis of trigrams. They shared the texts worse.

To achieve the same accuracy, a tree of greater depth was required. At the same time, better results were obtained on trees with a fragment size of 500 words. Thus, increasing the size of the fragment worsened the result. The most significant feature at the first level was the "Noun-Adjective-Noun" trigram.

References

1. Kjetsaa G. *Attributed to Dostoevsky: The problem of attributing to Dostoevsky anonymous articles in Time and Epoch*. Oslo, Solum Forlag A. S. Publ., 1986, 82 p.

2. Batura T. V. Formalnye metody opredeleniya avtorstva tekstov [Formal methods for determining the authorship of texts]. *Bulleten' Novosibirskogo gosudarstvennogo universiteta. Seria informatsionniye tehnologii* [*Novosibirsk State University Bulletin. Series Information Technology*], 2012, vol. 10(4), pp. 81–94. (In Russian)

3. Lebedev A. A. *Vvedenie v prikladnuyu lingvistiku* [*Introduction to applied linguistics*]. Petrozavodsk, Petrozavodsk State University Press, 2019, 48 p. (In Russian)

4. Malyutov M. B. Obzor metodov i primerov atribucii tekstov [Overview of methods and examples of text attribution]. *Review of Applied and Industrial Mathematics*. Moscow, 2005, vol. 12(1), pp. 41–78. (In Russian)

5. Rogov A. A., Sedov A. V., Sidorov Y. V., Surovceva T. G. *Matematicheskie metody atribucii tekstov* [*Mathematical methods for text attribution*]. Petrozavodsk, Petrozavodsk State University Press, 2014, 96 p. (In Russian)

6. Calle-Martin J., Miranda-Garcia A. Stylometry and authorship attribution: Introduction to the special issue. *English Studies*, 2012, vol. 93(3), pp. 251–258.

7. Farringdon J. M. *Analyzing for Authorship*. Cardiff, University of Wales Press, 1996, 324 p.

8. Stamatatos E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 2009, vol. 60(3), pp. 538–556.

9. Kotov A. A., Mineeva Z. I., Rogov A. A., Sedov A. V., Sidorov Y. V. *Lingvisticheskie korpusy* [*Linguistic corpuses*]. Petrozavodsk, Petrozavodsk State University Press, 2014, 140 p. (In Russian)

10. Sidorov G., Velasquez F., Stamatatos E., Gelbukh A., Chanona-Hernández L. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 2014, vol. 41(3), pp. 853–860.

11. Rogov A. A., Kulakov K. A., Moskin N. D. Programmnaya podderzhka v reshenii zadachi atribucii tekstov [Software support in solving the problem of text attribution]. *Software Engineering*, 2019, vol. 10(5), pp. 234–240. (In Russian)

12. Breiman L. Random forests. *Machine Learning*, 2001, vol. 45(1), pp. 5–32.

13. Isert C. The editing distance between trees. *Ferienakademie Bäume: Algorithmik und Kombinatorik* [*Holiday Academy Trees: Algorithmics and Combinatorics*]. Sarntal, Italy, 1999, pp. 1–13.

14. Shchegoleva L. V., Lebedev A. A., Moskin N. D. Metody analiza dannykh v zadache razgranicheniya fol'klornykh i avtorskikh tekstov [Methods of data mining in the task of distinguishing between folklore and author's texts]. *Voprosy Jazykoznanija* [*Questions of Linguistics*], 2020, vol. 2, pp. 61–74. (In Russian)

Received: December 25, 2020.
Accepted: October 13, 2021.

Authors' information:

Roman V. Abramov — Student; monset008@gmail.com

Kirill A. Kulakov — PhD in Physics and Mathematics, Associate Professor; kulakov@cs.karelia.ru

Alexander A. Lebedev — PhD in Philology, Senior Lecturer; perevodchik88@yandex.ru

Nikolai D. Moskin — PhD in Technics, Associate Professor; moskin@petrsu.ru

Alexander A. Rogov — Dr. Sci. in Technics, Professor; rogov@petrsu.ru

Исследование особенностей публицистического стиля Ф. М. Достоевского с помощью n-грамм по материалам журналов «Время» и «Эпоха»*

Р. В. Абрамов, К. А. Кулаков, А. А. Лебедев, Н. Д. Москин, А. А. Рогов

Петрозаводский государственный университет, Российская Федерация,
185910, Петрозаводск, пр. Ленина, 33

Для цитирования: *Abramov R. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Rogov A. A.* Research of features of Dostoevsky's publicistic style by using n-grams based on the materials of the "Time" and "Epoch" magazines // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2021. Т. 17. Вып. 4. С. 389–396. https://doi.org/10.21638/11701/spbu10.2021.407

Работа посвящена изучению публицистического стиля Ф. М. Достоевского на материалах статей в журналах «Время» и «Эпоха» (1861–1865 гг.). Для этого были выбраны фрагменты текстов (в том числе М. М. Достоевского, Н. Н. Страхова, А. А. Головачева и др.) размером 500, 700 и 1000 слов, на которых выполнялся подсчет встречаемости би- и триграмм, представляющих собой закодированные последовательности частей речи. Далее на их основе были построены деревья решения и выполнен анализ точности распознавания текстов. Если рассмотреть классификацию на первом уровне дерева (размер фрагмента 1000), то точность в среднем была равна 87 %. Этим признаком выступает процент наличия биграммы «прилагательное — существительное». При анализе триграмм наиболее значимым признаком на первом уровне была последовательность «существительное — прилагательное — существительное». Также в статье рассмотрена задача сравнения полученных деревьев решений.

Ключевые слова: публицистический стиль, атрибуция текстов, дерево решений, n-грамма, Ф. М. Достоевский, сравнение деревьев, информационная система «Статистические методы для анализа литературных текстов».

Контактная информация:

Абрамов Роман Владимирович — студент; monset008@gmail.com

Кулаков Кирилл Александрович — канд. физ.-мат. наук, доц.; kulakov@cs.karelia.ru

Лебедев Александр Александрович — канд. филол. наук, ст. преп.; perevodchik88@yandex.ru

Москин Николай Дмитриевич — канд. техн. наук, доц.; moskin@petrsu.ru

Рогов Александр Александрович — д-р техн. наук, проф.; rogov@petrsu.ru

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 18-012-90026).