Text Attribution in Case of Sampling Imbalance by the Method of Constructing an Ensemble of Classifiers Based on Decision Trees *

Alexander Rogov[1], Roman Abramov[1], Alexander Lebedev[1][0000−0001−9939−9389], Kirill Kulakov[1][0000−0002−0305−419X], and Nikolai Moskin[1][0000−0001−5556−5349]

Petrozavodsk State University, Petrozavodsk, Russia
rogov@petrsu.ru, monset008@gmail.com, perevodchik88@yandex.ru,
kulakov@cs.karelia.ru, moskin@petrsu.ru
https://petrsu.ru/

Abstract. When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises. In this paper we consider possible solutions to this problem by the example of determining the style of Apollon Grigoriev. As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*). The SMALT information system ("Statistical methods for analyzing literary texts") was used to determine the frequency characteristics of the texts and Python 3.6 was used to build decision trees. As a result of calculations we can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. There also was a study of the article "Poems by A. S. Khomyakov", which confirms the previously conclusion that there is no reason to consider it as belonging to Apollon Grigoriev.

Keywords: Text attribution · F. M. Dostoevsky · Apollon Grigoriev · Poems by A. S. Khomyakov · sampling imbalance · decision tree · software complex "SMALT".

1 Introduction

Authorship identification of anonymous texts (attribution of texts) is one of most urgent problem for the philological community; however, there are no universal mechanisms for its solution [10]. Specialists in study of literature use methods that are often somewhat unusual for the humanitarian sphere to answer such questions, including mathematical methods of analysis. One of the issues, which is far from its final decision, is the affiliation of anonymous articles published in the magazines "Time" and "Epoch" (1861-1865). The authorship of some

* Supported by the Russian Foundation for Basic Research, project no. 18-012-90026.

of these articles has been established, while the authorship of other materials causes a lot of controversy and discussion in the philological field [6]. The solution to this problem is additionally hampered by the uneven amount of available textual material: there are many articles owned by F. M. Dostoevsky, while the remaining authors published in these journals (for example, A. Grigoriev, N. N. Strakhov, Ya. P. Polonsky, etc.), don't have so many texts that are uniquely attributed to them.

The following mathematical methods are used to establish authorship of works: neural networks, QSUM method, decision trees, support vector machine (SVG), k-means method, Bayesian classifier, Markov chains, principal component analysis, discriminant analysis, genetic algorithms, statistical criteria (χ^2 test, Student's t-test, Kolmogorov-Smirnov criterion), etc. 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. Note some authors who used mathematical methods to solve the problem of text attribution: Morton A. Q., Mendenhall T. C., Farringdon J. M., Efron B., Thisted R., Teahan W. J., Chaski C. E., Stamatatos E., Juola P., Peng R. D., Joachims T., Diederich J. J., Apte C., Lowe D., Matthews R., Tweedie F. J., de Vel O., Argamon S., Levitan S., Zheng R. [3], [5], [11], [13]. It should be noted that Russian language differs significantly from English, so the methods of analysis of texts in English is often not suitable for Russian language.

When solving the problem of classification into two classes, the problem of sampling imbalance often arises, i.e. when the number of objects of one class significantly exceeds the number of objects of another class. In this case the first class is called the majority class and the second class is called the minority class. In such samplings classifiers are configured for objects of the majority class, i.e. high accuracy of the classifier can be obtained without selecting objects of the minority class. When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises. Let's consider possible solutions to this problem by the example of determining the style of Apollon Grigoriev. The authors do not know any analogs of such research of Russian-language texts except for the works of G. Kjetsaa and M. A. Marusenko [4], [10].

2 Construction and Analyzing Decision Trees

An overview of the types of sampling imbalance and the methods used in such cases can be found in [8]. In this work we will use sampling, namely *Undersampling*. In this method the balance of sampling elements is achieved by removing objects of the majority class. The authors think that this method is more appropriate for the task than *Oversampling* (the sampling balance is achieved by duplicating objects of the minority class) or *SMOTE* (by generating new objects of the minority class).

As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*) [2]. The idea of this method is to train several models on random subsamples of the original sample (using *Bootstrap*) with further averaging. The authors believe that it meets the meaning of the task better than *Boosting*. During previous studies in determining the features of the journalistic style of F. M. Dostoyevsky we found that the constructed decision trees based on bigrams well reflect the author's style. In the experiments the best results were shown by decision trees with a fragment size of 1000 words. The optimal step size for choosing the beginning of the next fragment is 100 words. The same parameters were used in this work. The SMALT information system ("Statistical methods for analyzing literary texts") developed at Petrozavodsk State University was used to determine the frequency characteristics [9]. 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) and also allowed to mark the quotes, foreign words, introductory words, abbreviated words and non-linguistic symbols. A set of data for training was compiled (118 fragments – Apollon Grigoriev, 899 – the rest). The texts from which the data were prepared are presented in Table 1. In this case fragments of the texts of Apollon Grigoriev are objects of the minority class and all the others are from the majority class. The text size is quite small (from 2000 to 7000 words).

Table 1. Source texts for analysis.

Name	Author
Pismo k redaktoru	Y. P. Polonsky
Zhukovskij i romantizm	F. M. Dostoevsky
Literaturnaya isterika	F. M. Dostoevsky
Odin iz proektov chudesnago obogasheniya Rossii	I. N. Shill
Pismo k izdatelyu "Vremeni"	Y. P. Polonsky
Podpiska na 1863 god	M. M. Dostoevsky
Ryad statej o russkoj literature. Vvedenie	F. M. Dostoevsky
Slavyanofily, chernogorcy i zapadniki	F. M. Dostoevsky
Ryad statej o russkoj literature. G. -bov i vopros ob iskusstve	F. M. Dostoevsky
Knizhnost i gramotnost. Statya pervaya	F. M. Dostoevsky
Knizhnost i gramotnost. Statya vtoraya	F. M. Dostoevsky
Poslednie literaturnye yavleniya. Gazeta "Den"	F. M. Dostoevsky
Neobhodimoe literaturnoe obyasnenie, po povodu ra...	F. M. Dostoevsky
Politicheskoe obozrenie	A. A. Golovachev
Lermontov i ego napravlenie. Statya vtoraya	A. Grigoriev
Oppoziciya zastoya. Cherty iz istorii mrakobesiya	A. Grigoriev
Nashi domashnie dela	A. U. Poretsky
Durnye priznaki	N. N. Strakhov
Eshe o Peterburgskoj literature	N. N. Strakhov
Vsyo-li na Rusi tak ploho, kak kazhetsya?	V. P. Mesherskij

Python 3.6 was used to build decision trees (libraries: *scikit-learn* – for tree implementation, *pandas* – for data reading). The original data set was divided into 7 parts. In each part all fragments of Apollon Grigoriev were taken as a class with a label "1", the same number of fragments of other authors were taken randomly as a class with a label "0". Repetitions of fragments of other authors were not allowed.

A decision tree was trained on each part of data. The training continued until accuracy reached 100% (tree depth). The fragment of one of the trained trees is shown in Fig. 1. All trees formed an ensemble. The decision was accepted by a majority vote. Accuracy was calculated on the entire data set using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP is true-positive, TN is true-negative, FP is false-positive and FN is false-negative predicted class. The experimental results are presented in Table 2.

Table 2. Classifier accuracy

Depth	Accuracy
1	0,8628
2	0,9592
3	0,9841
4	0,9891
5	0,992
6	0,9901

In total 7 decision trees were built. A fragment of one of the trees is shown in Fig. 1. Note that on the third level there are two leaves that contain a small number of fragments (summary from 12 to 27, on average less than 8%). You should take into account the possible inaccuracy of the source data. The texts of Apollon Grigoriev could be edited by F. M. Dostoevsky. In addition there is a slight volatility in the parameters of the author's style depending on external factors (such as mood, health status, etc). Therefore, when solving the problem of text attribution, you should limit yourself to the first level or at most the first two levels of decision trees. As you can see from Table 2, the accuracy of the ensemble at the second level already falls into the generally accepted 5% significance level. Analyzing the decision trees contained in the ensemble, it can be noted that in 4 of them the first attribute was the "particle-adjective" bigram less than or equal to 6.5. In two cases the same attribute is found, but with a different threshold (less than or equal to 7.5). Only one tree had a different first attribute ("adjective-particle") less than or equal to 2.5. We can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. The proposed algorithm allows to solve the problem of text attribution.

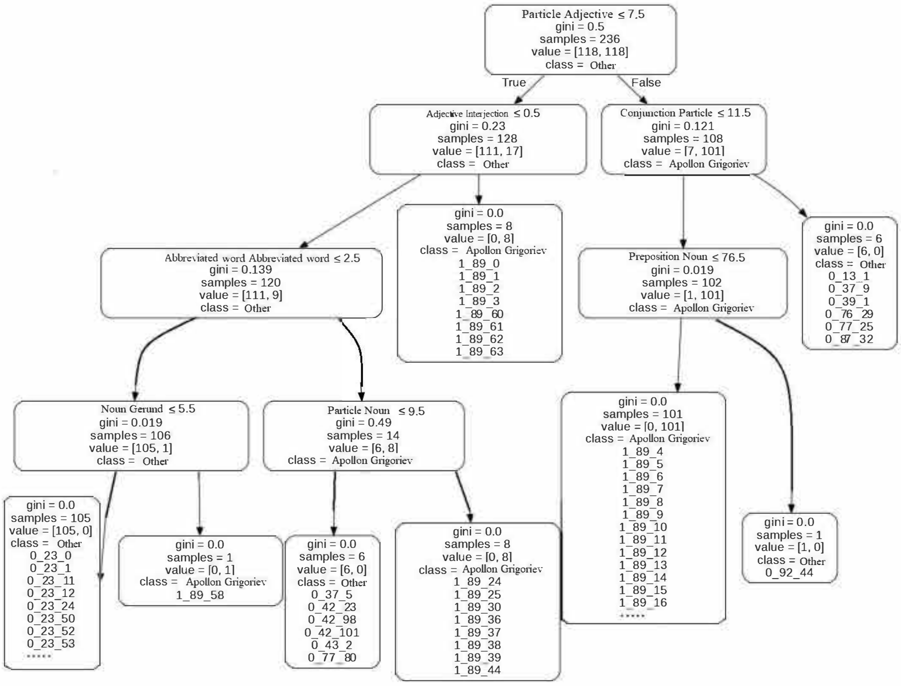

Fig. 1. A fragment of one of the trees

The influence of the universally accepted methods for processing unbalanced data "UpSampling", "UnderSampling", "SMOTE" on the accuracy of classification of works by Apollon Grigoriev was analyzed.

The available data set was divided into test (42 - Apollon Grigoriev, 310 - Other) and training samples. The training sample was subjected to the techniques listed above to confront class imbalance. Then the accuracy ("Accuracy", "roc-auc" curve) was calculated on a test sample, which was the same for all three techniques. The results of the experiment are shown in Table 3.

This analysis showed approximately the same accuracy of all three methods. UpSampling looks worse. The advantage of UnderSampling is that it is easier to explain. Therefore, the authors decided to focus on it.

Table 3. Experimental results

Depth	Accuracy (test)	roc-auc (test)	Accuracy (training)	roc-auc (training)
UnderSampling				
1	0,838068182	0,87718894	0,927711	0,927710843
2	0,920454545	0,923963134	0,975904	0,975903614
3	0,934659091	0,942319508	0,993976	0,993975904
4	0,946022727	0,948771121	1	1
UpSampling				
1	0,838068182	0,87718894	0,90938	0,909379968
2	0,90625	0,874731183	0,961844	0,961844197
3	0,931818182	0,889247312	0,980922	0,980922099
4	0,940340909	0,894086022	0,992846	0,992845787
5	0,963068182	0,906989247	0,99682	0,99682035
6	0,96875	0,910215054	0,999205	0,999205087
7	0,965909091	0,898310292	1	1
SMOTE				
1	0,838068182	0,87718894	0,90938	0,909379968
2	0,931818182	0,930414747	0,964229	0,964228935
3	0,980113636	0,957834101	0,983307	0,983306836
4	0,988636364	0,983256528	0,996025	0,996025437
5	0,980113636	0,957834101	0,99841	0,998410175
6	0,985795455	0,971351767	1	1

3 Analysis of "Poems by A. S. Khomyakov"

When discussing the affiliation of certain articles to certain authors, it should be noted, that in some cases there is no unequivocal evidence relating this article to a particular author. In particular, one of the controversial and still unresolved issues is the article "Poems by A. S. Khomyakov" a discussion about whose authorship in the literary criticism continues over the past twenty years.

The work of "Poems by A. S. Khomyakov" has long been attributed to Apollon Grigoriev. However, recently it has been considered the copyright text of F. M. Dostoevsky [14]. It was interesting to check where our classifier will take it. The text will be attributed to the author that most of the text fragments belong to. Fig. 2 shows one of the resulting decision trees. If we take the classification on the first node, then 6 of the 7 decision trees classify it as "Other", i.e. as not the text of Apollon Grigoriev. Only on one tree, there was an equality (5 fragments "for belonging" and 5 "against"). During the split on the second level 3 "for belonging", 3 "against" and in one rejection of the classification. Our study confirms the earlier conclusion [14] that there is no reason to consider the article "Poems by A. S. Khomyakov" as belonging to Apollon Grigoriev.

The combination of parts of the speech "Particle" + "Adjective" that is so often encountered in two texts precisely belonging to Apollon Grigoriev (in transliteration from Russian "Lermontov i ego napravlenie. Statya vtoraya" and

"Oppoziciya zastoya. Cherty iz istorii mrakobesiya"), almost does not appear in the text of the controversial article "Poems by A. S. Khomyakov". The author repeatedly uses this combination in the two indicated articles, then in the desired article it occurs only 10 times (the text consists of 2031 words), in six cases of which it is a "ne" particle, and in three cases - a "dazhe" particle; over large parts of text, such combinations of parts of speech could not be found (while in other articles belonging to A. Grigoriev, such combination is found more often and more diverse in terms of emerging types of particles - not only "ne" and "dazhe", but also "tolko", "to", "vse-taki", "zhe" followed by the adjective. Of course, this observation alone is not enough to doubt A. Grigoryev's text attribution, however, the application of methods based on decision trees can help with comprehensive analysis of texts in general, and the article "Poems by A. S. Khomyakov" in the context of the issue of the attribution of journalistic texts.

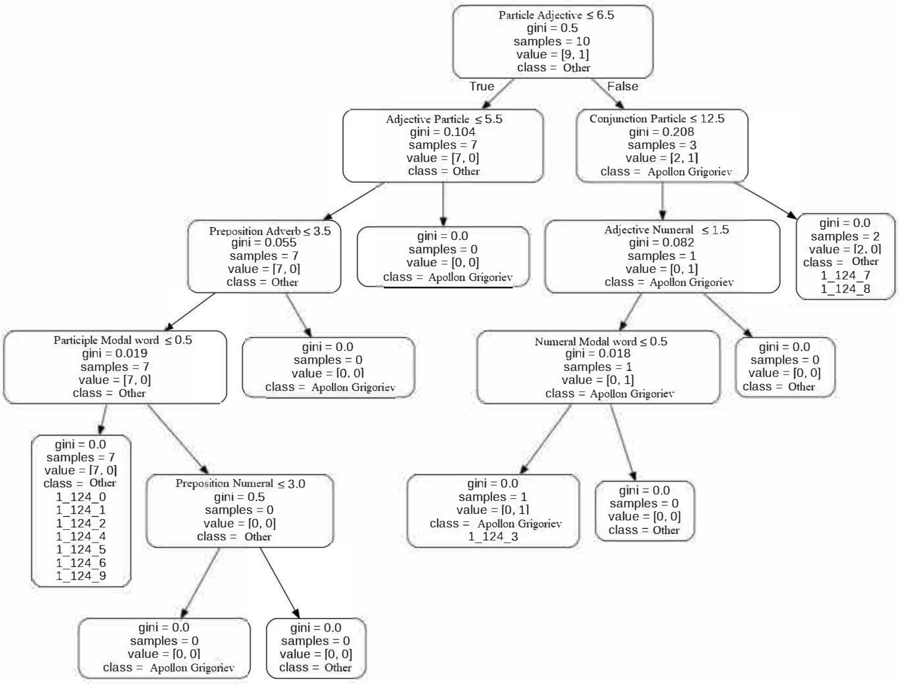

Fig. 2. One of the trees of the classifier in the analysis of the work "Poems by A. S. Khomyakov".

4 SMALT Information System

Specialized software is required for research in the field of text attribution. As an example, we note several software tools that are described in more detail in [1]:

- "Stileanalizator" (graphematic and statistical analysis, work with marked texts);
- "Avtoroved" (graphematic, morphological and statistical analysis);
- "Atributor" (statistical analysis);
- "Lingvoanalizator" (graphematical and statistical analysis).

The SMALT information system developed at Petrozavodsk State University [7], [9], [12] is designed for the collective work of various specialists with texts. The information system can be divided into three sections (see Fig. 3): import of new texts, verification of texts by philologists and the use of various analysis methods both on a single text and for a group of texts.

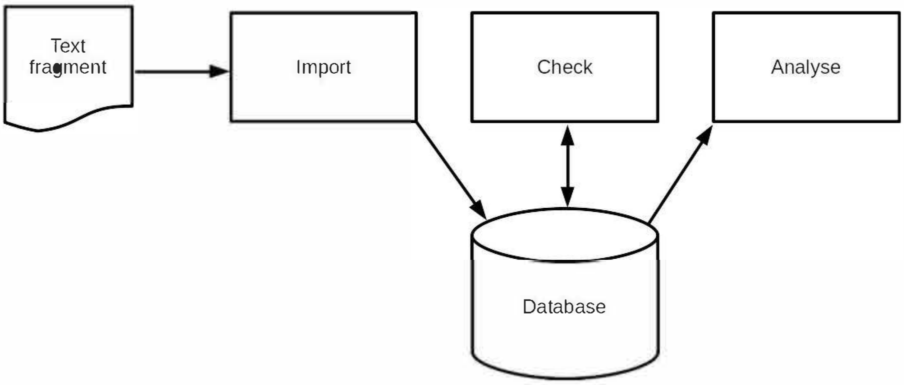

Fig. 3. The architecture scheme of the SMALT software system.

As part of the text import process, the text is divided into sections, paragraphs, sentences and words, as well as matching each word with its morphological analysis. If the task of text separation is typical, then the task of comparing the morphological analysis is rather complicated. The problem is both in the wide variety of spelling of the word (using pre-revolutionary graphics, a more flexible dictionary allowing different spelling of the word), and in the need to take into account the context of the use of the word. At different times, algorithms for finding the first possible variant, a frequently used variant and an algorithm based on n-grams were used to select the semantic analysis of the word. The latter has a great prospect due to the small number of subsequent corrections.

As part of the text verification process, philologists perform correction of text analysis (for example, combining or separating words), correction of morphological analysis of a word, or creation of a new analysis. Using the web interface allows several specialists to work on the text at the same time.

During the analysis process, the SMALT information system provides researchers with access to the accumulated database in various sections. For example, one of the popular statistical characteristics is Kjetsaa metrics [15]. The SMALT information system calculates the characteristics of both a single work and a group of texts. Another objective of the analysis is to identify the causes of the results. For example, to identify the reasons for the separation of text fragments between different nodes of the decision tree. The SMALT information system allows you to access the source data of the required fragment for subsequent linguistic analysis.

5 Conclusion

When solving the problem of determining the author's style of Apollon Grigoriev, the problem of sampling imbalance often arises, i.e. when the number of objects of one class significantly exceeds the number of objects of another class (in this case, the objects are the texts of the analyzed authors). As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*). The idea of this method is to train several models on random subsamples of the original sample (using *Bootstrap*) with further averaging. The authors believe that it meets the meaning of the task better than *Boosting*. Analyzing decision trees built using Python 3.6 (libraries: scikit-learn-tree implementation, pandas-data reading), we can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev.

The obtained knowledge was used to study the authorship of the article "Poems by A. S. Khomyakov", a discussion about whose authorship in the literary criticism continues over the past twenty years. If we take the classification on the first node, then 6 of the 7 decision trees classify it as "Other", i.e. as not the text of Apollon Grigoriev.

The obtained results were presented for further consideration to the specialists of the Department of Russian Language and the Department of Classic Philology, Russian Literature and Journalism (Petrozavodsk State University).

Acknowledgements. This work was supported by the Russian Foundation for Basic Research, project no. 18-012-90026.

References

1. Batura, T. V.: Formal methods for determining the authorship of texts. Novosibirsk State University Bulletin. Series "Information Technology". Novosibirsk **10**(4), 81–94 (2012)
2. Bühlmann, P.: Bagging, Boosting and Ensemble Methods. In: Gentle J., Härdle W., Mori Y. (eds) Handbook of Computational Statistics. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-21551-3_33

3. Calle-Martin, J., Miranda-Garcia, A.: Stylometry and Authorship Attribution: Introduction to the Special Issue. English Studies **93**(3), 251–258 (2012) https://doi.org/10.1080/0013838X.2012.668788
4. Gurova, E. I.: Methods of Authorship Attribution in Contemporary National Philology. The New Philological Bulletin **3**(38), 29–44 (2016).
5. Farringdon, J. M.: Analyzing for Authorship / J. M. Farringdon with contributions by Morton A. Q., Farringdon M. G., Baker M. D. Cardiff, University of Wales Press (1996).
6. Kjetsaa, G.: Attributed to Dostoevsky: The Problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S. (1986)
7. Kotov, A. A., Mineeva, Z. I., Rogov, A. A., Sedov, A. V., Sidorov, Y. V.: Linguistic Corpuses. Petrozavodsk: PetrSU Publ. (2014)
8. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0
9. Rogov, A., Kulakov, K., Moskin, N.: Software support in solving the problem of text attribution. Software engineering **10**(5), 234–240 (2019) https://doi.org/10.17587/prin.10.234-240
10. Rogov, A., Sedov, A., Sidorov, Y., Surovceva, T.: Mathematical methods for text attribution. Petrozavodsk, PetrSU Publ. (2014)
11. Romanov, A. S.: Methodology and software complex for identifying the author of an unknown text. Tomsk (2010)
12. Sidorov, Y. V.: Mathematical and informational support of literary text processing methods based on formal grammatical parameters. Petrozavodsk (2002)
13. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology **60**(3), 538–556 (2009) https://doi.org/10.1002/asi.21001
14. Zakharov, V.: Question about Khomyakov. In: Zakharov, V. The name of the author is Dostoevsky. Essay on creativity. Moscow, Indrik, 231–247 (2013)
15. Zakharov, V.N., Rogov, A.A., Sidorov, Y. V.: The problem of Dostoevsky grammatical constants search and anonymous and pseudonymous articles, published in "Time" and "Epoch" magazines (1861-1865) attribution. Works and Materials of "Russian Language Historical Destiny and the Present" International Congress. Moscow, MSU, 404–405 (2001)