Glottometrics 50, 2021, 76-89.

Automatic Identification of Authors' Stylistics and Gender on the Basis of the Corpus of Russian Fiction Using Extended Set-theoretic Model with Collocation Extraction

Alexandr Osochkin[1]
Xenia Piotrowska[2]
Vladimir Fomin[3]

Abstract

We present a novel quantitative approach for classification of authors' stylistics and gender differences based on extraction of word collocation. The proposed algorithm attenuates previously described issues of text processing using the vector models. We demonstrate the approach by analyzing a corpus of Russian prose. We discuss different approaches for classification and identification of the author's style implemented by currently-available software solutions and libraries of morphological analysis, methods of parameterization, indexing of texts, artificial intelligence algorithms and knowledge extraction. Our results demonstrate the efficiency and relative advantage of regression decision tree methods in identifying informative frequency indexes in a way that lends itself to their logical interpretation. We develop a toolkit for conducting comparative experiments to assess the effectiveness of classification of natural language text data, using vector, set-theoretic and the author's set-theoretic with collocation extraction models of text representation. Comparing the ability of different methods to identify the style and gender differences of authors of fiction works, we find that the proposed approach incorporating collocation information alleviates some of the previously identified deficiencies and yields overall improvements in the classification accuracy.

Keywords: *Natural language processing, frequency and morphological analysis, text-mining, gender linguistics, collocation extraction, set-theoretic model, vector text analysis.*

1. Introduction

Recent studies on the use of deep machine learning in the field of natural language processing (NLP) and text-mining (Kang, et al., 2020; Moschitt, 2004) have shown that statistical methods can be more effective (Grekhov, 2012) when used in combination with linguistic (morphological and parsing) analysis (Khalezova, et al., 2020). This concept has given rise to a separate direction in linguistics, which studies language based on statistical regularities, including the use of linguistic and semantic analysis, expanding the statistical approach to text analysis through the use of latent semantic connections between text elements (Maheshan, et al., 2018; McCann et al., 2017; Yang, et al., 2019). Increasing availability of computational resources and advanced algorithm implementations has now enabled individual researchers to process large volumes of data, and employ sophisticated computational methods for their analysis. One of the most promising avenues for improving NLP technologies is through incorporation of parsing-based quantitative methods (for example, the relationship of compositional construction and word formation, the length of compounds and the length of their components).

[1] Herzen State Pedagogical University of Russia, Moika Emb., 48, Saint-Petersburg, 191186, Russian Federation osa585848@bk.ru, http://orcid.org/0000-0001-9449-5603.
[2] Herzen State Pedagogical University of Russia, Moika Emb., 48, Saint-Petersburg, 191186, Russian Federation, krp62@mail.ru, http://orcid.org/57207357482.
[3] Herzen State Pedagogical University of Russia, Moika Emb., 48, Saint-Petersburg, 191186, Russian Federation, v_v_fomin@mail.ru, http://orcid.org/0000-0001-7040-5386.

The modern quantitative approach to text analysis arose from the development of many different models of text representation that were focused on solving highly specialized problems (Devlin, et al., 2018; Belinkov & Bisk, 2018; Belinkov & Glass, 2019; Iyyer et al., 2018). Modern NLP data analysis and processing packages rely on complex linguistic algorithms for text analysis (Piotrowska, 2014). In 2019 Google introduced Bidirectional Encoder Representations from Transformers (BERT), which has shown to be highly efficient in solving a wide range of tasks (Macro, et al., 2020), and formed the basis of NLP digital services. Recent studies, however, have demonstrated that in some settings BERT can pose a number of notable disadvantages.

Microsoft Azure Machine Learning, based on the BERT model, is part of the Cortana Intelligence Suite that enables predictive analytics and interaction with data using natural language and speech. One of the promising BERT applications is the improvement search systems based on the classification and indexing of texts on sites and repositories.

The paper by T. Macro (Macro, et al., 2020) received an award for identifying critical flaws in modern text processing models at the "Association for Computational Linguistics" (ACL) in 2020. This critical survey examined performance of advanced applications of BERT in "Google AI", "Microsoft Azure Text Analysis", "Amazon Comprehend", "Facebook RoBERTa AI", etc. The survey noted shortcomings in the ability to capture the grammatical and lexical cohesion structure of the text, its integrity, and incorporation of term collocations in texts. These limitations of the modern text representation models suggest that further research is needed to improve text representation models, procedures for generating and extracting significant digital indicators, as well as in the development of artificial intelligence algorithms.

The current study aims to evaluate the effectiveness of technology in identifying and classifying the author's style and gender differences in literary works using a quantitative approach based on collocation algorithms and regression decision trees.

2. Data analysis models

Most importantly, a quantitative approach to text analysis requires a formalized representation of textual data. There are several notable paradigms of text representation that rely on various mathematical models, including the vector model, probability word distribution, and the set-theoretic model (Wang & Zhu, 2019; Martin & Jurafsky, 2019).

A specific mathematical model for text representation enables extraction of quantitative characteristics from text data (Kashcheyeva, 2013). Specific quantitative representations include, frequency-based models (Beel et al., 2017), frequency-morphological (Osochkin, et al., 2018), vector (Salton G., Allan J. & Buckley, 1994), topic vector (Devlin et al., 2018; Belinkov & Bisk, 2018; Belinkov & Glass, 2019; Iyyer, et al., 2018), and set-theoretic models (Allahyari, et al., 2017; Harish, 2012; Marcus, 1967). Despite the large number of approaches for text conversion, all models can be divided into two types: vector and set-theoretic.

A number of advanced computational models are utilized in the computer text processing industry. Here we will consider and compare the vector, and the set-theoretic models, together with our extension of the set-theoretic model incorporating term collocation.

2.1. Vector model of text representation

The vector model became popular at the beginning of the 20th century, and nowadays it remains relatively unchanged despite the appearance of alternative models (Wang & Zhu, 2019; Popescu, et al., 2010). A vector text model represents each word or sentence in a text as a vector that captures the underlying meaning (Popescu et al., 2010). The vector model is often referred to as a topic vector model, because the basis of text class division is rooted in a semantics of the words, which in aggregate represent the subject field. Vector text representation can use

different text elements for analysis: words, sentences, paragraphs, particles of speech, etc., though sentences are the most commonly used text elements.

Topic vector text representation supposes that text contains chapters that have a common subject and include paragraphs. Paragraphs, in turn, contain sentences.

$$G = \{G_1, G_2, \dots, G_n\}; Vg = \{Vg_1, Vg_2, \dots, Vg_n\}, (1)$$

where *G* represent multiple topic chapters in the text, G_i represents an *i*-th text chapter, $i = 1 \dots n$, *Vg* –multiple vectors of topic chapters, Vg_i topic vector of i-th chapter.

Therefore:

$$A_i = \{A_{i1}, A_{i2}, \dots, A_{\Im}\}; Va_i = \{Va_{i1}, Va_{i2}, .., Va_{\Im}\}, (2)$$

where A_i are multiple paragraphs of *i*-chapter, A_{ij} are paragraphs of *i*-th chapter, $j = 1 \dots m$; Va_i are multiple vectors of paragraph topics. Va_{ij} are topic vector of the *j*-paragraphs of the *i*-th chapter. The mathematic model of sentence interpretation is represented as:

$$P_{ij} = \{P_{ij1}, P_{ij2}, \dots, P_{ijk}\}; Vp_{ij} = \{Vp_{ij1}, Vp_{ij2}, \dots, Vp_{ijk}\}, (3)$$

where P_{ij} are multiple sentences of *i*-th chapter of *j*-th paragraph, P_{ijh} are *h*-th sentence of *i*-th chapter of *j*-th paragraph, $h = 1 \dots k$; Vp_{ij} are multiple vectors of sentences topics of *i*-th chapter of *j*-th paragraph; Vp_{ijh} are the topic vector of *h*-th sentence of *i*-th chapter of *j*-th paragraph.

The vector models of text representation were initially able to overcome key disadvantages of frequency and theoretical models of data representation, including the homonym problem and the consideration of the semantics of sentences. Further development of the vector model of data representation, however, could not solve a number of outstanding challenges, including time-consuming vector calculation needed for analysis of large texts. Therefore, the vector model is mainly applied to processing of small texts.

Aside from the computational requirements, the significant disadvantage of vector models lies in the lack of consideration for the language specificity of the word order, position of the subject and predicate, characteristics of parts of speech, forms and other text features.

We choose the "Word2Vec" library as the main tool for studying the vector model of text representation, because it:

- supports more than 40 languages, including Russian,
- makes use of an embedded model of replacing associative words and homonyms (Bag of Words),
- does not require supervised training data.

2.2. Set theoretic model of text representation

Set-theoretic models assume that a text is composed of distinct terms (words, n-grams, sentences), possessing common characteristics and unique traits. The main concept of such models is the reflection of different text characteristics in relative indicators, to which mathematical methods for identification of common and unique characteristics of each sample analyzed text are applied. Set-theoretic models commonly rely on analysis of frequency and measures of metric proximity (e.g. Dice, Ochiai, Jaccard, Simpson etc.) (Marcus, 1967; Zakharov & Khochlova, 2010; Kolesnikova, 2016; Belyaeva, et al., 2019).

In our experiments on classification, the Jacquard similarity index was used as a metric for evaluating the similarity of words in texts (Jadhao, et al., 2016). This index is the easiest to calculate and is widely employed in linguistic analysis. Its values are equivalent in particular cases to other similarity measures: Sokal-Sneath and Serensen distance measures.

The values of the Jacquard coefficient vary from 0 to 1. The Jacquard coefficient measures the similarity between the sets of words used by two texts, and is defined as the size of the intersection (i.e words used in both texts) divided by the size of the union of the word

sets (i.e. total number of unique words used in both texts). To compare the proximity of two texts A and B, the Jaccard similarity index can be calculated using the formula:

$$K(A,B) = \frac{|A \cup B|}{|A \cap B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}. \quad (5)$$

In addition to words, such similarity indexes can be calculated for n-grams, sentences, etc. Detailed information on how indicators are calculated on the basis of Jaccard similarity index can be found in Moulton & Jiang (2018). We used the freely-licensed Python library "Jaccard-index", which is able to calculate the similarity index between texts to implement calculations of the Jaccard similarity index. The library is both fast and is well-supported by the text and image analysis communities. Words were used as the main unit of analysis.

2.3. Set-theoretic model with collocation extraction

The limitations identified in the work of T. Macro (Macro, et al., 2020) suggested that to improve the classification accuracy of author's style and gender identification (Mikros, 2013; Vincze, 2015), it is important to identify not only the topic aspects of the text, but also the morphological features of the words as well as their collocations.

In this context, we developed the FaM software, which was described in details in (Osochkin, et al. 2018). It uses an algorithm for text representation as a frequency-morphological set of indicators, considering collocations, and can improve the accuracy of classification in NLP applications.

The mathematical model of text representation with collocation extraction models texts as interconnected sequences of terms. It is based on the hypothesis that taking into account sustainable links in phrases and the relationship between text elements will create a more accurate model of text representation.

In order to take into account collocation, FaM calculates a number of custom indicators based on the usage frequency word sequences (n-grams) in the text. It is expected that some of the words in the sentence will not have a semantic connection with other members, such as: prepositions, parenthesis, etc. To accommodate such cases, FaM incorporates an algorithm based on morphological libraries, which identifies and omits the functional words and words that were not in a semantic connection with the sentence members when calculating n-gram sequences. A normalized text is represented as an array of objects, where each word is described as an object with properties: part of speech and morphological characteristics. For each sequence of objects and for each combination of their morphological characteristics, a separate frequency indicator is then calculated. This indicator is defined as a number of times a given object sequence occurrence occurs in a normalized text, divided by the total number of objects.

Thus, the set of n-gram indicators is determined by the natural language in which the text is written and by the length of the sequence (i.e. by n). The total set of indicators of bigrams ($n=2$) extracted for the Russian language can reach more than 200.

A key factor in improving the efficiency of classification is the conversion of text into a set of numerical indicators using frequency-morphological analysis. In FaM, morphological analysis is performed by a hybrid algorithm that uses two morphological analysis modules – Natural language processing (AOT) [4] and "Solarix Engine"[5].

The morphological module AOT is based on the multi-level representation of data in a natural language, and was first used in the French-Russian automated translation system (FRAM). The module contains a Russian morphological dictionary: about 161,000 words with various forms. It also incorporates syntactic and semantic analysis of the text.

4 Official website of the library "AOT" URL: http://www.aot.ru.
5 Website of the "Solarix Engine" library URL: http://www.solarix.ru/for_developers/api/grammar-engine-api.shtml.

Solarix Engine is a morphological analysis module that includes a dictionary of 1,800.000 words and 218,000 thesaurus articles, containing information about possible subordination and associative relationships between words, suitable for machine learning. The main advantage of this module is the support of different languages: English, French, German, etc.

A custom algorithm embedded in FaM enables the use these two libraries simultaneously, allowing one to obtain aggregated information about the analyzed word, its semantic relationship with other words in the sentence, and carry out morphological, syntactic, and frequency analysis. To identify semantic connections, the algorithm carries out syntactic analysis which identifies parts of speech and functional words in a sentence, and builds a syntactic tree. In the subsequent stages, the algorithm searches for words that are syntactically related to the subject or predicate in the sentence, and checks for semantic connections. The semantic relationship is evaluated by synthesizing a new sentence without the analyzed word, building a new syntactic tree, and analyzing the resulting node changes. If no context changes were observed for the tree nodes associated with the deleted word, the word was omitted from the normalized text.

2.4. Normalization and relevance of indicators

Almost all text analysis packages perform pre-processing to normalize the data. Text pre-processing enables more accurate and reliable extraction of the features present in the text. In this regard, the lemmatization procedure is a key pre-processing step that can significantly reduce the size of the vector space, by removing inflectional endings and collapsing words into their basal forms. This word variant reduction is also beneficial for estimation of the vector indexes, as it reduces the dimensionality of the vector space. The NLTK4Russian library was used to carry out text lemmatization [6].

Normalization of the data is carried out using the TF-IDF technology, the Scikit-learn library [7].

TF_{ij} indexes are defined as the frequency of word's use in the analysed text, relative to the total number of words in the text:

$$TF_{ij} = \frac{f_{ij}}{f_{i1}+f_{i2}+\cdots+f_{in}}, i = 1,\acute{m}, (6)$$

where TF_{ij} is the index for the j-th word in the i-th text, f_{ij} is the frequency of use of the j-th word in the i-th text.

The TF-IDF method (Roul, et al. 2017) calculates the value of the j-th term IDF_{ij} in the i-th text as the product of the frequency of term usage in the TF_{ij} document and the normalized inverse frequency of term content in the documents.

$$IDF_{ij} = TF_{ij} * log\left(\frac{|D|}{|Df_i|}\right), i = 1,\acute{m}, j = 1,\acute{n}, (7)$$

where D is the total number of documents in the collection. Df_i is the number of documents in which the term f_j occurs. If the term is not present in any of the documents Dfi is taken to be equal to 1.

This approach allows one to determine the importance of the term in the entire collection of the analyzed documents. Terms with high uniqueness, which are less common in other documents, and often occur in the analyzed document, have the highest IDF value.

6 Website of NLTK4 Russian developer: Department of mathematical linguistics SPbSU. URL: http://mathling.phil.spbu.ru/node/160.

7 Official website of the developer "SciKit-learn" URL: https://scikit-learn.org/stable/index.html.

2.5. Artificial intelligence algorithms

For tasks of parametric analysis, regression, classification, identification and knowledge extraction, NLP uses an extensive toolkit of artificial intelligence algorithms (neural networks, genetic algorithms, metric algorithms, reference vectors, decision trees, etc.). Earlier studies of classification methods (Osochkin et al., 2018; Fomin & Osochkin, 2016) have shown that regression decision tree algorithms were effective in identifying the style and gender of the author of literary works. The advantage was due to their ability to attain higher classification accuracy when using small texts (less than 80,000 objects), compared to neural networks and the support vector machines. A significant advantage of all decision tree methods is their ability to represent the results as a hierarchical set of logical rules "if-then", which allows for meaningful identification, interpretation, verification of the classification results, as well as quantitative estimation of significance of each indicator. A variety of algorithms for constructing decision trees exist (Random Forest, ID3, C4.5, C5.0, CRT, CHAID, etc.), allows application of full potential of statistical analysis in the framework of a quantitative approach to natural language text processing.

In this paper, several algorithms were used for building decision trees in the IBM SPSS data analysis package. When identifying the author's gender, the CRT algorithm was used, as it is most suitable for binary classification. When classifying texts by the author's style, the CHAID algorithm was used, as it is the most suitable for classification of a large number of clusters.

3. Research materials

Two sets of texts in Russian were collected and analyzed in this study. The first corpus (*Corpus 1)* of texts contains Russian and Soviet literary prose of the 19th-20th centuries. Novels and stories were divided into chapters, containing several texts and each author is represented by 30 texts, as detailed in Table 1. The column "Average quantity of text symbols" shows the average number of characters contained in each text (without spaces).

Table 1.

Corpus of Russian fiction

№	Class	Number of texts	Average quantity of text symbols
1	V.I. Belov	30	67,024
2	A.P. Beliaev	30	54,029
3	M.A. Bulgakov	30	89,525
4	D.A. Granin	30	48,078
5	F.M. Dostoyevsky	30	92,031
6	I.A. Efremov	30	53,089
7	A.I. Kuprin	30	56,380
8	A.N. Ostrovsky	30	62,022
9	Strugatsky brothers	30	72,097
10	L.N. Tolstoy	30	110,705
11	A.A. Fadeev	30	55,092
12	A.P. Chekhov	30	56,092
13	M.A. Sholokhov	30	105,032

The second corpus (*Corpus 2)* was compiled to evaluate identification of gender based on the author's style. It consists of 120 works of fiction by Russian writers of the XXI century, for example, a series of novels and fantasy by O.M. Sergeeva, A. Sergeeva, Surzhevskaya Marina Eff IR, K.G. Nazimov, M. Kamensky etc. The literary works included in the corpus of texts were taken from various Internet sites dedicated to fiction and scientific literature. Table 2 shows the characteristics of the Corpus 2.

Table 2.
Author's gender text corpus

№	Class	Number of texts	Average quantity of text symbols
1	M	60	487 792
2	F	60	589 126

4. Results and Discussion

4.1 Author's style classification experiment

The experiment was conducted to identify the author's style based on fiction prose in order to evaluate the accuracy of the proposed data extraction method. The objective was to classify the corpus of texts presented in Table 1 according to authors' style identification.

The classification is based on the data extracted using the text data processing models described in the previous sections. The "exhaustive CHAID" algorithm using the Gini coefficient was chosen as the main algorithm for building the decision tree. This algorithm was chosen due to its ability to handle large number of classes (more than 10 clusters at the same time), high accuracy (Osochkin et al., 2020; Piotrowska, 2012), and a lower complexity of the resulting decision tree due to its use of non-binary tree-splitting algorithm.

The ratio of the training and test samples was taken to be 50%. The maximum tree depth was limited to 10, to avoid internal nodes with low number of texts. The Pierson Chi-square test was used to check the hypothesis of finding common characteristics. Since all indicators are relative, the node split significance criteria was set to 0.005.

The results of the experiments[8] on the author's style identification based on the text Corpus 1 using different mathematical models are shown in Figure 1.

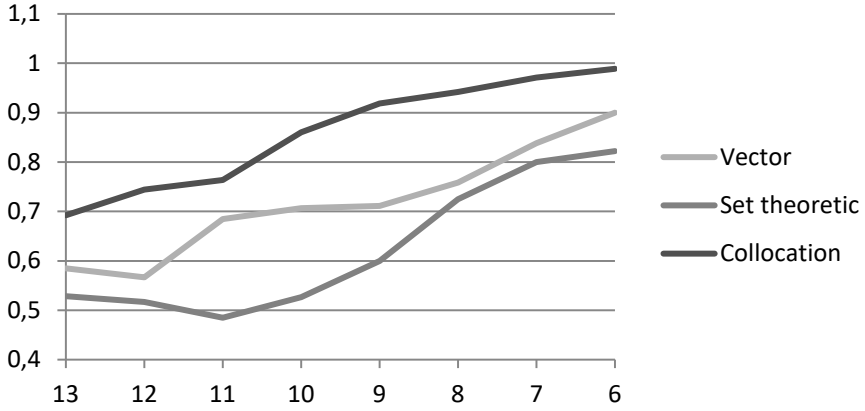

Figure 1. Chart of the author's style identification accuracy.
(*x*-axis shows the number of clusters and *y*-axis shows the average classification accuracy)

Figure 1 shows the dependence of the classification accuracy on the number of clusters used. It is seen that the trend of classification accuracy increases when related authors (authors who had similar stylistic features) are removed from the classification, as the use of similar stylistic forms led to misclassification of texts among such authors.

[8] Classification was performed using the IBM SPSS software package.

Table 3.
Author's style identification accuracy

Number of clusters	Vector model of text representation	Set-theoretic model	Set-theoretic model with collocation extraction
13	58.46%	52.82%	69.23%
12	56.67%	51.67%	79.65%
11	68.48%	48.48%	78.18%
10	70.67%	52.67%	87.72%
9	71.11%	60.00%	95.74%
8	75.83%	72.50%	95.00%
7	83.81%	80.00%	97.14%
6	90.00%	82.22%	98.88%

The data of Table 3 shows the identification accuracy of the author's style based on the parameters of the mathematical models and the number of utilized clusters. The results show that when classifying texts using 13 classes, the largest number of errors in the classification was encountered in the works of V.I. Belov and A.A. Fadeev. The total share of correctly identified works of A.A. Fadeev was 38.89%, when using the collocation method. An erroneous attribution of V.I. Belov is observed, since when classifying indicators extracted from a text using vector and set-theoretical text representation, the greatest percentage of false identifications of the author's style is seen in the works of V.I. Belov.

To increase the accuracy of classification, Fadeev's texts were removed from the corpus, since they were often identified as the works of V.I. Belov. The results of the removal of these texts slightly increased the overall accuracy of classification with the collocation method, but in other text representation methods the accuracy decreased. Furthermore, V.I. Belov has remained as the most mis-identified author.

Also, removing of Belov's texts from the text corpus had a positive effect on the general accuracy of the set-theoretic model with collocation extraction and vector representation methods. As the general accuracy of the set-theoretic method decreased, the largest proportion of errors in the classification with 12 clusters was made in the identification of V.I. Belov and A.R. Belyaev. Therefore, their works were removed from the subsequent classification based on the 11 and 10 clusters.

In the subsequent experiments, the authors' works classified with the least accuracy were removed from the corpus. The general accuracy reached 95.74% by using the set-theoretic text method with collocation extraction, when classifying by 9 clusters. That is 22.5% more accurate than in the set-theoretic text representation model and 19.17% more accurate than in the vector model.

In order to improve further classification accuracy, such authors as A.I. Kuprin, M.A. Bulgakov, Strugatsky Brothers, A.N. Ostrovsky, were removed. Each author's removal increased the accuracy of the general classification.

Table 4 shows detailed classification results using a mathematical model of text representation based on a set-theoretic text representation with collocation extraction.

Table 4.
Classification of Russian prose by 13 authors

	Belov	**Beliaev**	**Bulgakov**	**Granin**	**Dostoyevsky**	**Efremov**	**Kuprin**	**Ostrovsky**	**Strugatsky**	**Tolstoy**	**Fadeev**	**Chekhov**	**Sholokhov**	**Accuracy (%)**
Belov	15	4	2	0	0	1	0	0	0	0	1	0	0	**65.22**
Beliaev	2	7	0	0	0	0	0	0	0	0	0	0	0	**77.78**
Bulgakov	1	0	19	0	2	0	0	0	0	0	3	0	1	**73.08**
Granin	3	1	0	18	0	0	0	0	0	0	0	0	0	**81.82**
Dostoyevsky	1	1	0	1	9	0	1	0	0	0	0	0	2	**60.00**
Efremov	5	0	0	0	0	6	0	0	0	0	0	0	0	**54.55**
Kuprin	0	3	0	0	0	0	13	0	0	0	2	0	0	**72.22**
Ostrovsky	1	0	0	0	0	0	1	8	0	0	1	0	0	**72.73**
Strugatsky	0	0	0	0	0	4	0	0	10	0	0	0	0	**71.43**
Tolstoy	0	0	0	0	0	0	0	0	0	10	0	1	2	**76.92**
Fadeev	4	5	0	0	0	2	0	0	0	0	7	0	0	**38.89**
Chekhov	0	0	0	0	0	0	2	0	0	0	0	7	0	**77.78**
Sholokhov	0	0	0	0	0	0	0	0	0	0	0	0	6	**100.00**
Share of author's material in the test sample (%)	16.41	10.77	10.77	9.74	5.64	6.67	8.72	4.10	5.13	5.13	7.18	4.10	5.64	**69.23**

Our results (Table 3-4) show that the data classification using the set-theoretic model, with collocation increased the accuracy of author identification by 24.63%, and the average increase was 15.81%, which indicates the effectiveness of the method. The vector text representation model showed on average 25.15% lower accuracy compared to the set-theoretic text representation with collocation extraction.

Table 5 shows the indicators and their values that were used by the exhaustive CHIAD decision tree construction method.

Table 5.
The most important nine indicators of the author's identification

№	**Indicator**	**Weight(%)**
1	*Noun in accusative form + verb 1-st person*	6.26
2	*Noun in accusative form + verb 2-nd person*	5.58
3	*The use of Latin symbols*	5.53
4	*Adjective + Adjective*	4.99
5	*Adverb + Adverb*	4.87
6	*Numerals per sentence*	4.87
7	*Personal pronouns per sentence*	4.51
8	*Adjective + unanimated noun*	4.21
9	*Punctuation marks per sentence*	4.09

The main attributes used to identify the author were related not only to the frequency of individual parts of speech, but to their sequences. For example,

– the frequency of using a pair of nouns in the accusative form with the verb in the 1-st and the 2-nd personal;

– the use of Latin symbols allows one to distinguish most of the authors by their time periods at the first levels of decision trees;

– the authors of the Soviet period do not use Latin characters, which makes it possible to uniquely identify the works of Strugatsky, Belyaev, etc.

4.2 Author's gender identification experiment

It was shown in (Macro, et al., 2020) that different digital services based on the Bert language model[9] made mistakes when identifying the author's work by gender.

We conducted an experiment to identify the author's gender. Specifically, we classified the Corpus 2 according to the author's gender. A binary algorithm for building a CRT decision tree was chosen, due to its efficiency in carrying out binary classifications. The results are presented in the Table 6.

Table 6.
Classification by author's gender

Vector model	**Books quantity**	**Gender**	**F**	**M**	**Accuracy**
	37	F	25	11	69.44%
	23	M	12	12	50.00%
	Total 60	Share (%)	61.67	38.33	61.67%
Set-theoretic model	**Books quantity**	**Gender**	**F**	**M**	**Accuracy**
	36	F	19	12	61.29%
	24	M	17	12	41.38%
	Total 60	Share (%)	60.00	40.00	51.67%
Set-theoretic model with collocation extraction	**Books quantity**	**Gender**	**F**	**M**	**Accuracy**
	31	F	29	2	93.50%
	29	M	2	27	93.10%
	Total 60	Share (%)	51.70	48.30	93.30%

As can be seen from the classification results, the best accuracy was shown by the collocation method, with the total accuracy of 93.3%. The main characteristics that were used to identify the text were: the frequency of particles usages, the frequency of constructions such as *"a noun + verb in the 2-nd person"*, the average length of wordd, adverbs and punctuation marks per sentence, etc.

The main features for identifying the author's gender were the number of adjectives used, the frequency of adverbial verbs, and the number of n-grams: "*a noun in accusative + verb in the 1-st person*, *a noun in accusative + verb in the 2-nd person"*.

At the first level of binary classification, it was possible to divide the samples almost in half, thanks to the feature of particles from the total number of words. It was found that the female writers used particles in their works much more often than the male writers. In cases where particles did not accurately identify the cluster, it was found that the authors could be identified using the average word length in the text, with the male writers using on average longer words. A distinctive feature of the female authors was a more frequent use of the n-gram

[9] It is mentioned modifications of the Bert language model such as Google BERT and Facebook RoBERTa AI.

"*nouns in the instrumental case + verbs in the 1-st person"*. At the last level of the classification tree, punctuation marks were used; the male writers employed fewer figures of speech, direct speech, and other constructs where punctuation marks are used.

Table 7.
Indicators value

№	Indicator	Weight(%)
1	*Percentage of the total number of particles*	27.67
2	*Nouns in the instrumental case + verb in the 1 - st person*	20.63
3	*Average words length*	20.27
4	*Percentage of participles*	11.79
5	*Punctuation marks per sentence*	11.32
6	*Vowel letters per word*	8.28

As it can be seen from the Table 7, the extraction of indicators related to the use of parts of speech and their features from the text has significantly increased the accuracy of classification when identifying the author's style in different literary works. The second most important indicator identified by the regression trees was the template "*nouns in the creative case + a verb in the 1-st person*", as this sequence was more often used by the female authors.

When identifying the author's gender, the template that took into account the morphological form of the bigram: "*nouns in the instrumental case + verb in the 1-st person*", which had an indicator value of 20.633 and was often used in the algorithm of regression decision trees to determine the author's gender.

The conclusions, which are based the author's gender identification evaluations, confirm that the text representation model based on the set-theoretic model with collocation is more effective compared to other models.

6. Conclusion

The use of a quantitative approach with collocation extraction allows one to increase the accuracy of the authorial style identification. The experimental results of the style and gender identification accuracy confirmed the effectiveness of the proposed modification of the set-theoretic model of text processing of Russian prose. Algorithms for frequency-morphological extraction of numerical indicators and the formation of text indexes that reflect the frequency of individual parts of speech and n-gram parts of speech use, can be successfully used to identify the style. Our experiments have confirmed an increase in the total classification accuracy using collocation compared to the vector model of text representation.

Using the set-theoretic model with collocation extraction allows one to eliminate some of the disadvantages of the BERT data representation model that were pointed out earlier, and in conjunction with the methods of regression decision trees, the potential of text mining can be expanded. We also plan to conduct further experiments to analyze the accuracy of identifying the emotional colors of messages.

Acknowledgements

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (project No. FSZN-2020-0027).

References

Allahyari, M. et al. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 13–17, 2017, Halifax, Nova Scotia, Halifax, Canada, CoRR abs/1707.02919/.*

Beel, J. et al. (2017). TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections. *Conference Preliminary Results Papers*, 1–8.

Belinkov, Y. & Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. *International Conference on Learning Representations*. URL: https://arxiv.org/abs/1711.02173

Belinkov Y. & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*. 7, 49–72.

Belyaeva, L.N. et al. (2019). *Setevyye lingvisticheskiye tekhnologii. Kollektivnaya monografiya* (Network linguistic technologies. Collective monograph), 111. Saint-Petersburg: Herzen State University of Russia.

Devlin, J., et al. (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *Google AI Language*. URL: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Fomin, V.V. & Osochkin, A.A. (2016). Text classification in the creative category with application of the frequency-morphological analysis algorithms and regression trees. *Current issues and prospects for the development of mathematical and natural Sciences. Collection of scientific papers on the results of the international scientific and practical conference. May 11, 2016 Omsk*, 64–66.

Grekhov, A.V. (2012). Kvantitativnyy metod: poisk latentnoy informatsii (Quantitative method: searching for latent information). *Vestnik Nizhegorodskogo universiteta im. Lobachevskogo.* 1 (3), 94–100.

Harish, B. (2012). Text Document Classification: An Approach Based on Indexing. *International Journal of Data Mining & Knowledge Management Process*, 1, 43–66. DOI: 10.5121/ijdkp.2012.2104

Iyyer M. et al. (2018). Adversarial example generation with syntactically controlled paraphrase networks. *Proceedings of NAACL-HLT*, 1875–1885. https://www.aclweb.org/anthology/N18-1170

Jadhao, A. et al. (2016). Text Categorization using Jaccard Coefficient for Text Messages. *International Journal of Science and Research (IJSR)*, 5, 2046–2050.

Kang, Y. et al. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics,* 7 (12), 1–34. DOI: 10.1080/23270012.2020.1756939

Kashcheyeva, A.V. (2013). Kvantitativnyye i kachestvennyye metody issledovaniya v prikladnoy lingvistike (Quantitative and qualitative methods of research in applied linguistics). *Sotsial'no-ekonomicheskiye yavleniya i protsessy*, 3 (49), 18.

Khalezova, N., et al. (2020). Cross-sectional Study of Clinical and Psycholinguistic Characteristics of Mental Disorders in HIV Infection. *R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019).* Proceedings of the III International Conference on Language Engineering and Applied Linguistics. CEUR-WS, 2552, 161–178 URL: http://ceur-ws.org/Vol-2552/Paper14.pdf

Kolesnikova, O. (2016). Survey of Word Co-occurrence Measures for Collocation Detection. *Comp. y Sist.* [online]. 2016, 20 (3), 327–344. URL: https://doi.org/10.13053/cys-20-3-2456

Macro, T., et al. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList», *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 49024912 URL: https://www.aclweb.org/anthology/2020.acl-main.442

Maheshan, M., et al. (2018). Indexing-Based Classification: An Approach Toward Classifying Text Documents Information Systems. *Design and Intelligent Applications*, 1, 894902. DOI: 10.1007/978-981-10-7512-488

Marcus, S. (1967). *Algebraic Linguistics; Analytical Models*. Academic Press: New York.

Martin, D. & Jurafsky, D. (2019). *Speech and Language Processing. An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

McCann, B., et al. (2017). Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems* 30 (NIPS 2017), 6294–6305.

Mikros, G. (2013). Systematic stylometric differences in men and women authors: a corpus-based study. Köhler, R.; Altmann, G. (eds.): *Issues in quantitative linguistics.* Lüdenscheid: RAM, 206–223.

Moschitt, A. (2004). Complex Linguistic Features for Text Classification: a comprehensive study. *Lecture Notes in Computer Science. 26 European Conference on IR Research, Sunderland, UK,* 181–196.

Moulton, R. & Jiang, Y. (2018). Maximally Consistent Sampling and the Jaccard Index of Probability Distributions. *International Conference on Data Mining, Workshop on High Dimensional Data Mining 2018*, 347–356. URL: https://arxiv.org/abs/1809.04052

Osochkin, A.A., et al. (2018). Eksperimenty text-minig po klassifikacii tekstov v ramkah zadach personalizacii obrazovatel'noj sredy (Text-minig experiments on the classification of texts in the framework of the problems of personalization of the educational environment). *Informatizaciya obrazovaniya i nauki*, 2 (38), 38–50.

Osochkin, A.A., et al. (2020). Comparative Research of Index Frequency - Morphological Methods of Automatic Text Summarisation. *NESinMIS-2020. Proceedings of the XV International Conference "New Educational Strategies in Modern Information Space", Saint-Petersburg, Russia, March 25, 2020. Vol. 2401,* 73–86. URL: http://ceur-ws.org/Vol-2630/paper_8.pdf

Piotrowska, X.R. (2012). Kvantitativnyy psikholingvisticheskiy analiz khudozhestvennogo tvorchestva (Quantitative psycholinguistic analysis of artistic creativity). *Nauchnoye mneniye* (*Scientific opinion*), 6-7, 16–20.

Piotrowska, X.R. (2014). Tekst mayning: perspektivy razvitiya (A Survey of Text mining). *Izvestiya RGPU im. A.I. Gertsena*, 168, 128–134.

Popescu, I.-I. et al. (2010). Vectors and codes of text. Lüdenscheid: RAM.

Roul R. et al. (2017). Modified TF-IDF Term Weighting Strategies for Text Categorization. *Proceedings of 14th IEEE India Council International Conference (INDICON),* 16.

Salton, G., Allan, J. & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2), 97–108.

Yongchang, W. et al. (2019). Research on improved text classification method based on combined weighted model. *National Natural Science Foundation of China*, 7(11), 783–796.

Vincze, V. (2015). The relationship of dependency relations and parts of speech in Hungarian. *Journal of Quantitative Linguistics,* 22(1), 168–177.

Wang, Y. & Zhu, L. (2020). Research on improved text classification method based on combined weighted model. *Concurrency and Computation: Practice and Experience*, 32 (6), 783–796.

Yang, Zh., et al. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of Advances in Neural Information Processing Systems 32* URL: https://arxiv.org/abs/1906.08237

Zakharov, V.P., Khokhlova, M.V. (2010). Analiz effektivnosti statisticheskikh metodov kollokatsiy v tekstakh na russkom yazyke (A Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts). *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii. Proceedings of the Annual International Dialogue Conference. Bekasovo. 26–30 May 2010, 9 (16).* Moscow: Russian State University of Humanities, 137–14