MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Comparison of Style Features for the Authorship Verification of Literary Texts

K. V. Lagutina[1]

DOI: 10.18255/1818-1015-2021-3-250-259

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in English

Received May 4, 2021
After revision August 20, 2021
Accepted August 25, 2021

The article compares character-level, word-level, and rhythm features for the authorship verification of literary texts of the 19th-21st centuries. Text corpora contains fragments of novels, each fragment has a size of about 50 000 characters. There are 40 fragments for each author. 20 authors who wrote in English, Russian, French, and 8 Spanish-language authors are considered.

The authors of this paper use existing algorithms for calculation of low-level features, popular in the computer linguistics, and rhythm features, common for the literary texts. Low-level features include n-grams of words, frequencies of letters and punctuation marks, average word and sentence lengths, etc. Rhythm features are based on lexico-grammatical figures: anaphora, epiphora, symploce, aposiopesis, epanalepsis, anadiplosis, diacope, epizeuxis, chiasmus, polysyndeton, repetitive exclamatory and interrogative sentences. These features include the frequency of occurrence of particular rhythm figures per 100 sentences, the number of unique words in the aspects of rhythm, the percentage of nouns, adjectives, adverbs and verbs in the aspects of rhythm. Authorship verification is considered as a binary classification problem: whether the text belongs to a particular author or not. AdaBoost and a neural network with an LSTM layer are considered as classification algorithms. The experiments demonstrate the effectiveness of rhythm features in verification of particular authors, and superiority of feature types combinations over single feature types on average. The best value for precision, recall, and F-measure for the AdaBoost classifier exceeds 90% when all three types of features are combined.

Keywords: stylometry; natural language processing; style features; rhythm features; authorship verification

INFORMATION ABOUT THE AUTHORS

Ksenia Vladimirovna Lagutina correspondence author	orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru postgraduate student.

Funding: The reported study was funded by RFBR, project number 20-37-90045.

For citation: K. V. Lagutina, “Comparison of Style Features for the Authorship Verification of Literary Texts”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 250-259, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Сравнение стилистических характеристик для верификации авторов художественных текстов

К. В. Лагутина[1]

DOI: 10.18255/1818-1015-2021-3-250-259

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на английском языке

Получена 4 мая 2021 г.
После доработки 20 августа 2021 г.
Принята к публикации 25 августа 2021 г.

В статье сравниваются характеристики уровней символов, слов и ритма для верификации авторства художественных текстов 19-21-го веков. Корпуса текстов содержат фрагменты романов, каждый фрагмент имеет размер около 50 000 знаков. Для каждого автора приводится 40 фрагментов. Рассматриваются по 20 авторов, писавших на английском, русском, французском языках, и 8 испаноязычных авторов.
Авторы статьи используют существующие алгоритмы для вычисления популярных в современной компьютерной лингвистике низкоуровневых характеристик и распространённых в художественной литературе ритмических характеристик. Низкоуровневые характеристики включают в себя n-граммы слов, частоты встречаемости букв и знаков пунктуации, среднюю длину слова и предложения и т. д. Ритмические характеристики основаны на лексико-грамматических средствах: анафоре, эпифоре, симплоке, апозиопезе, эпаналепсисе, анадиплозисе, диакопе, эпизевксисе, хиазме, многосоюзие, повторяющихся восклицательных и вопросительных предложениях. Данные характеристики включают в себя частоты появления отдельных ритмических средств на 100 предложений, количество уникальных слов в аспектах ритма, доли существительных, прилагательных, наречий и глаголов в аспектах ритма. Верификация авторов рассматривается как задача бинарной классификации: принадлежит текст конкретному автору или нет. В качестве алгоритмов классификации рассматриваются AdaBoost и нейросеть со слоем LSTM. Эксперименты демонстрируют эффективность ритмических характеристик при верификации конкретных авторов и превосходство комбинаций типов характеристик над отдельными типами характеристик в среднем. Лучшее значение точности, полноты и F-меры для классификатора AdaBoost превышает 90%, когда комбинируются все три типа характеристик.

Ключевые слова: стилометрия; обработка естественного языка; стилистические характеристики; ритмические характеристики; верификация авторов

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ксения Владимировна Лагутина — автор для корреспонденции | orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
аспирант.

Финансирование: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-37-90045.

Для цитирования: K. V. Lagutina, "Comparison of Style Features for the Authorship Verification of Literary Texts", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 250-259, 2021.

Introduction

The authorship verification is the task of determination whether the text belongs to a given author or not. It is based on the assumption that the author has the individual set of style markers that can distinguish the author from others, but occurs in each of his/her texts [1].

In the state-of-the-art of the authorship verification and close text classification tasks there is no set of style features that would be versatile for different texts. Some feature types like character-level, word-level, and syntactic features appear in many investigations, but are often combined with more complex linguistic features [2, 3]. Researchers admit that the influence of different types of features on the quality of text classification remains underexplored [4].

Rhythm features are the subtype of the linguistic features that most often describe the style of literary texts [5]. They can be applied for authorship verification [6], but are rarely compared with other feature types [2].

The goal of this paper is comparing how different feature types affect the quality of the authorship verification of literary texts. We analyse rhythm features and popular low-level features based on statistics of text elements. The comparison is performed on the corpora of English, Russian, French, and Spanish literary texts.

1. State-of-the-art

The task of authorship verification is usually performed for the texts from the Internet: news articles, emails, reviews, etc. [2, 7].

In many cases the researchers modeled texts using only standard low-level features and experimented with classification. Halvani et al. [8] used stylometric features based on n-grams. The verification was realized by the determination of the proximity of the numerical feature vectors of the texts. Experiments were conducted in five European languages: Dutch, English, Greek, Spanish, and German. The F-measure varied from 67.37 % for Greek up to 83.33 % for Spanish. The method also showed good results at the PAN-2020 competition [9].

Potha and Stamatatos [7] introduced an intrinsic profile-based verification method that apply latent semantic indexing for topic modeling and low-level features: word and character n-grams. Then the algorithm calculated the text model that represents all texts of the same author as a common vector. Then it identified the authors by searching for test texts the closest vector from the authors' train ones. The researchers compared in experiments corpora of prose, newspaper articles, reviews in four languages: Dutch, English, Greek, and Spanish. The method achieved more than 80 % of the AUC.

Boenninghoff et al. [10] proposed a new neural network topology to identify whether two documents with unknown authors were written by the same author. This approach showed the best results of the precision, recall, and F-measure 84 % for short multi-genre social media posts.

Adamovic et al. [11] searched a wide range of word and character-based language-independent text stylistic features. Then they applied the SVM-RFE feature selection method to remove redundant and irrelevant characteristics. Authorship verification of articles in four languages: English, Greek, Spanish, and German showed a high result over 90 % of the accuracy.

To improve the quality of authorship verification and take into account domain peculiarities and the authors' idiolect, the researchers frequently applied linguistic features.

Al-Khatib and Al-qaoud [12] verified native and non-native speakers of online opinion articles. The feature set included statistical and linguistic features: number of unique words, complexity, Gunning-Fog readability index, character space, letter space, average syllables per word, sentence count, average sentence length, and the Flesch-Kincaid Readability. The accuracy varied for text corpora from 47 % to 77 %.

Lagutina et al. [6] investigated application of rhythm features to the authorship verification of the artistic prose. They found the features based on repetitions of words and sentences (anaphora, epiphora, aposiopesis, etc.) and verified authors of English, Russian, French, and Spanish prose. The F-measure achieved from 60 % to 95 % for different authors and about 80 % in average.

The literary texts are usually analysed not in the authorship verification but in the close task of the authorship attribution. For example, Stanisz et al. [13] created adjacency networks with words frequently appearing in texts, and their co-occurrences as vertices and edges' weights. Then the authors computed various graph characteristics: clustering coefficients of vertices, an average shortest path length, an assortativity coefficient, and modularity. The experiments showed the accuracy of 85–90 % for English and Polish books.

The analysis of the state-of-the-art papers shows the lack of comparison of different feature types with linguistic ones, especially for artistic texts. The authors usually rely on standard statistical features based on words and characters and try to extend them by relatively small number of syntactic, topical, or other linguistic features. Deep linguistic features remains under-researched, most probably, because of their complexity in search. Although such features are directly identify the author's style [5] and can be the most interpretable ones.

2. Style features

We compare three types of features: character-level, word-level, and rhythm-level ones. The first two feature types are the popular effective features from the state-of-the art. The rhythm features describe the specific style marks of the authors that frequently appear in literary texts.

Before feature calculation we search in plain texts the following elements:

- Top-40 unigrams and top-40 bigrams of words among the text corpora. They will be used for computing frequencies of occurrences for n-grams.
- Lexico-grammatical rhythm figures. For each text we found the lists of the following figures: anaphora, epiphora, symploce, anadiplosis, diacope, epizeuxis, epanalepsis, chiasmus, polysyndeton, repeating exclamatory sentences, repeating interrogative sentences, and aposiopesis. Their definitions and search algorithms are taken from the works of Lagutina et al. [6, 14]. The quality of figures search achieves 80–95 % of precision.

We compute the following style features:

- Character-level features:
  - Average sentence length in characters including punctuation marks and spaces.
  - Frequencies of occurrences of each letter among all letters. The uppercase letters are previously reduced to lowercase ones.
  - Frequencies of occurrences of each punctuation mark (.!?:, etc.) among all punctuation.
- Word-level features:
  - Average sentence length in words.
  - Average word length in characters.
  - Frequencies of occurrences of unigrams and bigrams among top-40 n-grams.
- Rhythm features:
  - The density of the figure — the number of occurrences of the rhythm figure (anaphora, epiphora, etc.) divided by the number of sentences.
  - The fraction of unique words—words that appear only once in rhythm figures.
  - The fraction of words of a particular part of speech (noun, verb, adverb, and adjective) in rhythm figures.

All features are calculated separately for each text. Character and word-level features represent the base statistics of the text style. Rhythm features represent the density and linguistic structure of the text rhythm. So the text is modeled as the vector of statistical and linguistic features.

3. Authorship verification

3.1. Design of authorship verification

After feature extraction we get the matrix where rows are texts of particular authors, columns are feature types. We verify each author separately using the whole matrix for the author's language. His/her texts are labeled as belonging or not belonging to him/her. Then the binary classification is performed.

Two classifiers are compared: AdaBoost and Bidirectional LSTM. They have already show their quality in solution of state-of-the-art text classification tasks [15].

The AdaBoost classifier combines the results of 50 Decision Tree classifiers. The Bidirectional LSTM neural network contains the Bidirectional LSTM layer with 64 units and a dense output layer with the sigmoid activation function. The loss function is categorical cross-entropy, the optimization algorithm is Adam, the number of epochs is 100.

In order to estimate the stability of classifiers, we apply the five-fold cross-validation technique: 80 % of texts are the training samples, 20 % are the test ones. The estimation is performed with three standard measures: precision, recall, and F-measure [16], and also their standard deviations.

The code for the feature selection and authorship verification is published at https://github.com/text-processing/prose-rhythm-detector. It is written in Python programming language and uses Stanza 1.1.1 NLP library for text representation and determination of parts of speech. For the verification it uses Scikit-Learn 0.23.2 and Keras 2.4.3.

3.2. Text corpora

We compare literary texts of four languages: English, Russian, French, and Spanish. The corpora were created manually collecting famous works of famous authors written in their native language.

In order to make texts equal in size, we extracted 1–4 fragments with the size about 50 000 characters including spaces from each prose text. In such a way each author is presented by 40 text fragments. English, Russian, and French corpora contain texts of 20 famous authors of 19th–21st centuries, 800 texts per corpora. The Spanish corpus has texts of 8 authors of 19-th–20th centuries, 320 texts in total.

4. Experiments

During experiments we compare features of three types: 36–43 character-level features (the letters differs for corpora in different languages), 82 word-level features, and 17 rhythm features.

Comparing two classifiers, we discover that AdaBoost outperforms the neural network by 10–15 % of precision, recall, and F-measure. Most probably, it happens due to the fact that the training sample has the insufficient size for better performance of the LSTM network. So the tables in this section contains classification quality for the AdaBoost algorithm.

Table 1 describes authorship verification quality for all feature types and their combinations. Ch means character-level features, W — word-level ones, Rh — rhythm ones, + marks the combination of two feature types, All — the combination of three feature types. Precision, recall, and F-measure are calculated as the averages for all authors. Bold marks the lines with best quality and best F-measures.

From Table 1 we can see that rhythm features provide the good classification quality. It is lower by 3–11 % of F-measure in the most cases, but has quite high values of 78–87 %. Besides, the number of rhythm features is several times less than character- and word-level ones, so the relatively small number of specific style parameters allow to achieve significant authorship verification quality.

Any combination of feature types improve quality by 2–14 %, but the combination of all types is slightly higher than of the two types.

Authors of Russian, French, and Spanish texts in most cases are verified better than English. In English and French texts the best feature type is character-level, In Russian and Spanish texts it is word-level.

Table 1. Mean measure values of the authorship verification

Language	Feature type	Precision	Recall	F-measure
English	Ch	87.8	80.7	84.1
English	W	85.8	78.2	81.8
English	Rh	82.0	74.2	77.9
English	Ch + W	92.2	84.0	**88.0**
English	Ch + Rh	90.8	80.9	85.6
English	W + Rh	88.8	81.7	85.1
English	All	**94.7**	**85.4**	**89.8**
Russian	Ch	91.2	81.4	86.0
Russian	W	92.0	81.9	86.7
Russian	Rh	84.7	76.7	80.5
Russian	Ch + W	96.9	86.7	**91.5**
Russian	Ch + Rh	94.3	85.4	89.6
Russian	W + Rh	92.2	82.6	87.1
Russian	All	**96.9**	**87.4**	**91.9**
French	Ch	93.7	86.5	90.0
French	W	91.8	80.1	85.6
French	Rh	83.5	75.9	79.5
French	Ch + W	95.4	89.2	**92.2**
French	Ch + Rh	96.2	86.6	**91.2**
French	W + Rh	93.3	83.0	87.9
French	All	**97.5**	**90.0**	**93.6**
Spanish	Ch	89.9	85.0	87.4
Spanish	W	92.3	87.9	90.1
Spanish	Rh	88.5	86.3	87.4
Spanish	Ch + W	92.5	87.8	90.1
Spanish	Ch + Rh	94.5	88.8	**91.6**
Spanish	W + Rh	93.7	88.6	91.1
Spanish	All	**94.1**	**90.0**	**92.0**

Table 2. Verification of English authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
W. Scott	Ch	95.4	8.4	88.0	6.9	88.5	5.7
W. Scott	W	95.9	6.6	91.7	9.0	92.1	5.9
W. Scott	Rh	89.6	5.6	88.8	5.6	89.1	5.5
W. Scott	Ch + W	**99.5**	0.3	92.6	11.6	91.9	7.6
W. Scott	Ch + Rh	98.1	2.2	76.9	16.6	89.1	**2.7**
W. Scott	W + Rh	98.1	2.2	**94.2**	7.7	88.7	12.6
W. Scott	All	97.7	3.3	93.5	6.1	**95.2**	4.6
Z. Smith	Ch	94.4	10.0	**86.8**	16.6	82.6	10.5
Z. Smith	W	48.4	0.5	57.2	10.1	59.1	20.4
Z. Smith	Rh	62.0	6.7	62.5	7.8	63.4	**5.4**
Z. Smith	Ch + W	89.0	19.8	83.2	21.0	**87.0**	11.8
Z. Smith	Ch + Rh	90.2	11.4	85.0	14.6	82.7	10.7
Z. Smith	W + Rh	53.5	10.2	67.7	19.1	52.0	5.5
Z. Smith	All	**99.3**	0.6	77.3	13.5	66.4	16.4
N. Gaiman	Ch	91.5	7.7	**77.5**	11.4	64.5	11.2
N. Gaiman	W	64.2	9.6	66.8	12.2	67.5	15.5
N. Gaiman	Rh	81.5	6.7	74.6	10.3	78.7	8.0
N. Gaiman	Ch + W	**94.5**	5.6	68.7	12.4	81.3	**6.5**
N. Gaiman	Ch + Rh	92.5	7.5	69.5	13.8	**82.2**	8.6
N. Gaiman	W + Rh	94.3	8.5	69.1	12.6	74.1	7.3
N. Gaiman	All	90.0	7.5	75.1	12.6	80.8	8.7

Table 3. Verification of Russian authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
M. Bulgakov	Ch	84.2	19.9	66.7	10.5	66.4	15.4
M. Bulgakov	W	76.0	22.4	68.2	11.2	83.8	18.7
M. Bulgakov	Rh	73.7	22.6	68.3	13.0	68.5	17.6
M. Bulgakov	Ch + W	**99.3**	0.5	76.7	12.2	**82.3**	11.9
M. Bulgakov	Ch + Rh	89.0	19.5	74.0	13.1	81.9	**9.6**
M. Bulgakov	W + Rh	75.7	22.0	60.0	8.8	75.2	13.2
M. Bulgakov	All	89.2	19.9	**78.2**	19.3	78.2	18.4
N. Leskov	Ch	**99.3**	0.6	70.7	12.2	69.1	11.5
N. Leskov	W	89.3	19.9	72.8	17.3	75.2	21.8
N. Leskov	Rh	85.0	8.9	78.7	8.4	80.7	11.7
N. Leskov	Ch + W	94.5	10.0	77.5	12.2	87.7	6.7
N. Leskov	Ch + Rh	88.4	13.1	80.1	16.5	87.8	10.8
N. Leskov	W + Rh	82.4	21.0	82.0	16.5	91.4	**5.6**
N. Leskov	All	89.5	20.7	**85.3**	18.1	**96.6**	6.8
A. Prokhanov	Ch	98.0	3.2	98.8	2.2	89.0	19.7
A. Prokhanov	W	96.9	3.4	94.1	4.9	96.0	6.7
A. Prokhanov	Rh	98.1	3.2	**98.9**	1.7	96.3	3.6
A. Prokhanov	Ch + W	98.2	3.3	98.5	2.8	97.9	2.9
A. Prokhanov	Ch + Rh	**99.9**	0.2	98.6	2.9	96.7	3.2
A. Prokhanov	W + Rh	96.1	4.5	95.7	3.7	97.1	2.5
A. Prokhanov	All	96.9	5.6	93.1	8.6	**98.4**	**2.0**

Table 4. Verification of French authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
S. Colette	Ch	95.2	1.4	94.1	2.2	96.3	2.9
S. Colette	W	94.1	3.3	88.3	3.5	92.6	3.8
S. Colette	Rh	92.3	3.7	88.1	3.6	90.2	4.6
S. Colette	Ch + W	97.0	1.8	96.1	2.3	95.6	2.0
S. Colette	Ch + Rh	**99.6**	0.2	95.1	3.9	98.1	2.2
S. Colette	W + Rh	96.0	1.8	94.2	5.2	95.0	3.6
S. Colette	All	**99.6**	0.3	**97.7**	3.4	**98.6**	**1.2**
V. Hugo	Ch	94.9	4.7	81.7	7.2	81.4	6.7
V. Hugo	W	92.4	6.0	73.3	6.7	80.4	**3.9**
V. Hugo	Rh	73.1	8.5	66.1	3.8	70.3	8.4
V. Hugo	Ch + W	**99.0**	0.7	80.3	6.2	83.3	8.1
V. Hugo	Ch + Rh	94.9	5.6	76.6	10.7	**88.6**	4.6
V. Hugo	W + Rh	85.1	14.0	75.9	6.2	82.4	6.8
V. Hugo	All	93.9	5.8	**85.0**	5.8	83.9	11.9
A. Exupery	Ch	96.1	6.4	80.3	16.7	67.6	**10.7**
A. Exupery	W	89.2	20.1	63.8	19.7	65.2	20.6
A. Exupery	Rh	99.3	0.5	69.2	11.7	70.5	12.0
A. Exupery	Ch + W	84.3	20.1	**78.5**	16.6	82.4	18.6
A. Exupery	Ch + Rh	89.6	20.0	73.6	17.0	81.7	12.6
A. Exupery	W + Rh	99.3	0.5	58.3	10.6	65.5	15.9
A. Exupery	All	**99.6**	0.3	75.3	8.5	**91.4**	15.0

Table 5. Verification of Spanish authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
V. Ibáñez	Ch	92.8	2.0	88.5	4.9	90.0	8.0
V. Ibáñez	W	98.2	1.5	94.6	3.7	97.1	2.2
V. Ibáñez	Rh	96.7	3.3	95.1	3.9	93.0	3.2
V. Ibáñez	Ch + W	95.1	4.0	95.6	4.2	95.4	2.2
V. Ibáñez	Ch + Rh	96.9	2.1	94.8	1.3	96.3	3.0
V. Ibáñez	W + Rh	99.3	0.6	**99.2**	1.2	**99.2**	**1.0**
V. Ibáñez	All	**99.4**	0.8	97.4	1.7	98.5	1.8
J. Dicenta	Ch	79.3	24.5	79.8	24.4	71.1	19.7
J. Dicenta	W	79.1	24.4	**85.0**	20.0	**88.3**	**11.7**
J. Dicenta	Rh	82.4	21.3	62.9	19.8	61.1	14.6
J. Dicenta	Ch + W	78.8	24.7	70.0	19.4	79.5	19.9
J. Dicenta	Ch + Rh	89.2	20.0	65.0	20.0	78.2	23.8
J. Dicenta	W + Rh	**89.4**	20.1	70.0	24.5	74.4	20.9
J. Dicenta	All	79.4	24.9	60.0	20.0	71.1	19.7

Tables 2–4 illustrate the typical cases of the authorship verification. Columns "Std dev" contain standard deviations of the measures in the left.

Almost all authors have very high precision of verification 78–99 %. Recall is varied significantly more: 60–98 %.

Several authors are verified with high quality 88–96 % of the F-measure by any feature type: W. Scott, A. Prokhanov, S. Colette, V. Ibáñez. The combination of feature types improves classification even more up to 96–99 %. We can say that the chosen features describe the style of such authors quite effectively.

Several authors are verified with lower quality 66–88 % of the F-measure: Z. Smith, M. Bulgakov, V. Hugo, J. Dicenta. They also have very high standard deviations of 10–20 %. We can point out the feature types that provide quite good precision and recall (usually there are word-level features or combinations with them). But we can conclude that the proposed feature set does not describe common details in the style of such authors.

For some authors the F-measure grows significantly for combination of features. For example, texts of N. Leskov are verified with 69–80 % of F-measure, combinations of two feature types provides the F-measure of 87–91 %, and the combination of all features allows to achieve the best value of 96 %. Besides, the N. Leskov's style is better described by rhythm features than by others, because rhythm features provide higher F-measure of 80 % against 69 % and 75 %. Texts of A. Exupery show the same tendencies.

Several authors are verified significantly better by rhythm features than by statistical ones, for example, N. Gaiman, N. Leskov, A. Exupery.

Thus, all feature types can provide good verification quality. The specific linguistic features — rhythm features — achieve in many cases high precision, recall, and F-measure with small standard deviations. So they are as useful and stable style markers as standard statistical features: character and word level ones.

Verification of particular authors shows that the many authors have the same style in different fragments. They can be successfully separated from others using only one feature type or the combination of standard and rhythm features. Nevertheless, texts of several authors are verified with very high standard deviations, so there are needed other linguistic features to verify reliably their text fragments.

Conclusion

We applied three types of style features: character, word, rhythm-level features, and their combinations to the authorship verification of literary texts in English, Russian, French, and Spanish. Experiments revealed the same tendencies for all four languages. In average combinations of features provide higher classification quality than the single feature type. Moreover, rhythm features are almost as good style markers as popular low-level features.

The more detailed analysis of authorship verification allowed to discover the fact that many authors write text fragments in the same style, so they can be successfully verified by the single feature type or by the combination. But several authors are verified with significantly lower quality than others. The future investigations can be devoted to the error analysis of classification of their texts and search of the larger set of linguistic style markers that help to verify more authors.

References

[1] E. Stamatatos, "A survey of modern authorship attribution methods", *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[2] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, and I. Paramonov, "A survey on stylometric text features", in *Proceedings of the 25th Conference of Open Innovations Association (FRUCT)*, IEEE, 2019, pp. 184–195.

[3] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications", *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–36, 2018.

[4] C.-G. Lim, Y.-S. Jeong, and H.-J. Choi, “Survey of Temporal Information Extraction.”, *Journal of Information Processing Systems*, vol. 15, no. 4, pp. 931–956, 2019.

[5] E. Boychuk, I. Paramonov, N. Kozhemyakin, and N. Kasatkina, “Automated approach for rhythm analysis of French literary texts”, in *Proceedings of 15th Conference of Open Innovations Association FRUCT*, IEEE, 2014, pp. 15–23.

[6] K. Lagutina, N. Lagutina, E. Boychuk, V. Larionov, and I. Paramonov, “Authorship Verification of Literary Texts with Rhythm Features”, in *Proceedings of the 28th Conference of Open Innovations Association FRUCT*, 2021, pp. 240–251. DOI: 10.23919/FRUCT50888.2021.9347649.

[7] N. Potha and E. Stamatatos, “Intrinsic author verification using topic modeling”, in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ACM, 2018, pp. 1–7.

[8] O. Halvani and L. Graner, “Rethinking the evaluation methodology of authorship verification methods”, in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2018, pp. 40–51.

[9] O. Halvani, L. Graner, and R. Regev, “TAVeer: an interpretable topic-agnostic authorship verification method”, in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020, pp. 1–10.

[10] B. Boenninghoff, R. M. Nickel, S. Zeiler, and D. Kolossa, “Similarity learning for authorship verification in social media”, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2457–2461.

[11] S. Adamovic, V. Miskovic, M. Milosavljevic, M. Sarac, and M. Veinovic, “Automated language-independent authorship verification (for Indo-European languages)”, *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 858–871, 2019.

[12] M. A. Al-Khatib and J. K. Al-qaoud, “Authorship verification of opinion articles in online newspapers using the idiolect of author: a comparative study”, *Information, Communication & Society*, pp. 1–19, 2020.

[13] T. Stanisz, J. Kwapień, and S. Drożdż, “Linguistic data mining with complex networks: a stylometric-oriented approach”, *Information Sciences*, vol. 482, pp. 301–320, 2019.

[14] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, “Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries”, in *Proceedings of the 26th Conference of Open Innovations Association FRUCT*, IEEE, 2020, pp. 247–255.

[15] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey”, *Information*, vol. 10, no. 4, 150 (1–68), 2019.

[16] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks”, *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.