The Influence of Different Stylometric Features on the Classification of Prose by Centuries

Ksenia Lagutina, Nadezhda Lagutina
P.G. Demidov
Yaroslavl State University
Yaroslavl, Russia
ksenia.lagutina@fruct.org, lagutinans@gmail.com

Elena Boychuk
Yaroslavl State Pedagogical
University named after K.D.Ushinsky
Yaroslavl, Russia
elena-boychouk@rambler.ru

Ilya Paramonov
P.G. Demidov
Yaroslavl State University
Yaroslavl, Russia
ilya.paramonov@fruct.org

***Abstract*—In this paper the authors compare by classification quality different types of stylometric features: low-level features that include character-based and word-based ones, and high-level rhythm features. The authors classified texts into centuries with each feature type separately and their combinations applying four classifiers: Random Forest and AdaBoost meta-algorithms, a LSTM neural network, and a GRU neural network. The experiments with three text corpora in English, Russian, and French languages showed that combining rhythm features and low-level features significantly improved quality of classification by centuries. Besides, classification results allowed to compare the styles of writing in different languages from a point of view of structure of sentences.**

I. Introduction

The determination of time period to which a document belongs is important both for historical documents and journalistic texts, as well as for the fiction. Linguists consider the division of literature into periods as an aspect of the study of socio-cultural situations, historical processes, and the author's style [1]. The explicit determination of the period when a text was created or published is often impossible, whereas many texts do not have metadata or direct references of the publication date. In this case another information should be used, for example, the features of a language that the text is written in, that is to say, its style [2].

The changes in the style of a literary text over time characterize both the individual author's style [3] and the language in general [4]. The mathematical model for automatic determining whether a text belongs to a certain time period is usually based on very simple characteristics: n-grams of words [5], [6], lengths of sentences, quantitative characteristics of parts of speech [7]. At the same time, researchers point to the need to add more complex stylometric features to the model, such as rhythmic and grammatical ones [8].

Modern software libraries and frameworks for text processing make it easy to calculate stylometric features based on characters and words. In contrast, extraction of the features based on the structure of phrases and sentences requires the substantial effort and additional research.

Besides, complex stylometric features are used much less often, not only because of difficulties in their calculation. The reason is that the small number of works systematizes the influence of different types of features on the quality of text classification tasks [2].

Thus, we set two tasks for this research: (i) the automatic classification of fiction of the 19–21st centuries by periods of their publication using rhythm features and (ii) the comparison of classification quality of three types of stylometric features: character-based, word-based, and rhythm-based. Such classification can provide the explanation of changing and evolving writing styles.

The paper is structured as follows. Section II describes modern research in the field of text classification by time periods and the detection of temporal information in texts. Section III defines stylometric features that we use in our investigations. In Section IV we give the main structure of our experiments with classification. In Section V we experiment with several text corpora in three languages and compare classification quality of different feature types. Section VI reveals the peculiarities of the feature types and propose the interpretation of the results. Conclusion summarizes the paper.

II. State-of-the-art

Most researchers who solve the problem of the text classification by time use the word-based features.

For example, Zhao et al. [7] note that determining time of a document is an important step in information search and is necessary for solving various problems, including clustering documents, creating a timeline, and adapting a search system for temporal queries. They present methods for time stamping individual parts of web documents. For each part of a document, 44 features are defined: statistics of terms, sentence length, meanings and a number of dates, numbers of verbs in different forms. Document parts are divided into 5 categories by time of writing: no later than 30 days, from a month to a year, from a year to three years, from three to six years and more than six years. The F-measure varies from 44 % to 72 % for different document corpora. The authors underline that the quality of a text corpus is one of the important factors of the classification quality.

Jatowt et al. [5] propose the text dating system. The authors estimate the time that it takes to create a document by associating a word-time pair. The words are extracted from articles that are known as related to the particular period of time. If a document under study contains many words related to a certain period of time T, then it is considered as having a strong connection with the period T. The text features are based on calculating frequencies of n-grams ($n = 1, 2, 3, 4, 5$). Unfortunately, the numerical assessment of quality is not given in the article. The authors consider their system as an interactive online tool meant to facilitate the process of determining an age of a document, as well as to support the understanding of historical documents. They plan to conduct the research with experts and consider various genres of documents.

Khan et al. [9], [10] search for the news related to a certain event. Their text corpus consists of 3 500 news documents for 35 queries. The search is based on the selection of words and phrases that express the time of described events. Each temporal expression has a particular weight. The method for determining suitable documents is based on the ranking documents according to the degree of the correspondence of time of the requested event. The results reach the accuracy of 35–77 % for different time periods. This work is largely based on extracting temporal expressions, analyzing the vocabulary of different times, and the assumption that the more parts of a document belongs to one time period, the more likely the document itself also belongs to this period.

Thus, in the considered works, the basis for the documents dating is the use of lexicon features of a particular time. Many works are devoted directly to identifying words corresponding to the particular time periods.

Spitz et al. [11] correlate time periods and the terms using a weighted bipartite graph that is constructed on the base of the sentence structure. Lin et al. [12] reveal the semantic evolution of words over time, compute word frequencies for texts, and use them to classify documents by time periods. The accuracy is from 39 % for 6-year periods to 49.3 % for 20-year ones. Fukumoto and Suzuki [13] propose a similar method for selecting features for document classification. They distinguish two classes of words: time-independent terms and time-dependent terms. The frequencies of these terms form a feature vector for each text. In experiments the authors use Japanese newspaper articles from 1991 to 2012. The corpus consists of 2 883 623 documents, divided into 16 categories. The F-measure is 68.8 %. This is one of the few works devoted to a non-English language.

The text features, calculated on the base of the frequency of the occurrence of individual words, sometimes phrases are very popular, although the task of determining the correspondence of time and term is complex and ambiguous. However, the art historians, when distinguishing literary periods, often pay attention to more complex elements of the language: sentence structure, context, and order of words, syntactic and rhetorical figures of speech [1], [14].

One of the studies on the automatic classification of texts by time that use a large number of various text features, is conducted during the Semeval 2015 competition [15]. The organizers note that the language changes over time, even for relatively short periods, and suggest the participants to solve the problem of the automatic determining of the time periods of news from newspapers published between 1700 and 2010. From 7 teams, only 4 had a problem solution. The best results (the accuracy from 76.7 % to 86.8 %) are obtained if the large set of text features is used: meta-properties of the document length, stylistic, grammatical, lexical features, and even the search for a direct mention of the document date [16]. However, the analysis of the influence of these different features on the quality of classification has not been carried out.

Stanjer and Zampieri [17] classify the Portuguese historical texts to different centuries basing on the changes in writing style. Four criteria are used for the classification: average sentence length, average word length, lexical density, and lexical richness. Detection of the changes of these features allows to determine time of the historical text creation. The analysis of diachronic changes in these four features show that the texts written in the 17th and 18th centuries have similar feature values, and it differs significantly from texts written in the 19th and 20th centuries. The F-measure for classification for the 4 centuries is 52 %, while for the classification into two classes (17th–18th centuries and 19th–20th centuries) it is 92 %. The text corpus consists of 87 texts.

An interesting study at the intersection of classical and computer linguistics was conducted by [8]. The authors examine the stylistic differences between English poetry and prose of two periods: 1870–1920 and 1970–2019. Text features include features of the grammatical structure of sentences, meter, arrangement of stressed and unstressed syllables, rhythmic patterns. As a result of experiments, the authors conclude that poetry of 1970–2019 is more similar to prose of its period than poetry of 1870–1920 with prose of the same period. Although the changes in stylistic features of prose of two periods are minimal, but they are significant in poetry.

From this overview, we can conclude that complex stylometric features are rarely used for the automatic dating of documents, although they have a significant potential for improving the quality of the text classification. Using grammatical and rhythm features can give not only the high quality solutions to the text processing problems, but also an additional material for the interpretation of results by linguists. Another important point is that the vast majority of the experiments are performed with the text corpora in English. The use of dating methods for documents in other languages is not investigated enough.

III. Stylometric features used in research

Stylometric features can be classified into two types: low-level features like numbers of words, characters, n-grams, etc., and high-level or linguistic features. They are subdivided into character-based and word-based depending on the language unit they describe. For this research we chose the following low-level features that are popular in state-of-the-art [18]:

- Character-based features:
  - Average sentence length in characters including both letters and punctuation marks.

- Frequencies of letters. For each letter we count the number of occurrences and divide it by the number of all letters. The text is previously reduced to lowercase.
- Frequencies of punctuation marks: .!?:, etc. For each character we count the number of occurrences and divide it by the number of all marks.

• Word-based features:
  - Average sentence length in words.
  - Average word length in letters.
  - Frequencies of top-40 n-grams for $n = 1, 2, 3$. For each unigram, bigram, or trigram we calculate the number of occurrences in a text corpus, then we choose the most frequent 40 unigrams, bigrams, and trigrams. For each text we also compute their numbers of occurrences and divide them by the total number of occurrences of these 120 n-grams in the text.

These features are compared with a specific subtype of linguistic features: rhythm-based ones. We chose rhythm features that are based on 8 rhythm figures of speech: anaphora, anadiplosis, diacope, epanalepsis, epiphora, epizeuxis, polysyndeton, and symploce. Their definitions are described in our article [19]. These figures frequently appear in the texts and differ for each century and language.

We have calculated the rhythm features that describe rhythm figures of speech as integral units and as the structure consisting of different parts of speech.

• Rhythm-based features:
  - The number of occurrences of each rhythm figure divided by the number of sentences.
  - The fraction of unique words—words that repeat only once among all words that appear in rhythm figures.
  - The fraction of words of a particular part of speech: noun, verb, adverb, and adjective—among all words that appear in rhythm figures.

Thus, we have three types of features that describe the style of a text in different ways. None of the features have absolute values, therefore, they do not depend on the size of a text and can be used for the comparison of prose texts with different length.

IV. Design of experiments with classification

A. Structure of experiments

Stylometric features form a model of a text style that can be used for the text classification. We classify texts into three classes by their publication date: the 19th, 20th, and 21st century. Texts published in the 19th century has the similar style, it is true for the 21st century, and 20th century is more heterogeneous, but different from others as we discovered in our previous work [19].

The main stages of classification experiments are presented in Fig. 1. All the stages are performed fully automatically.

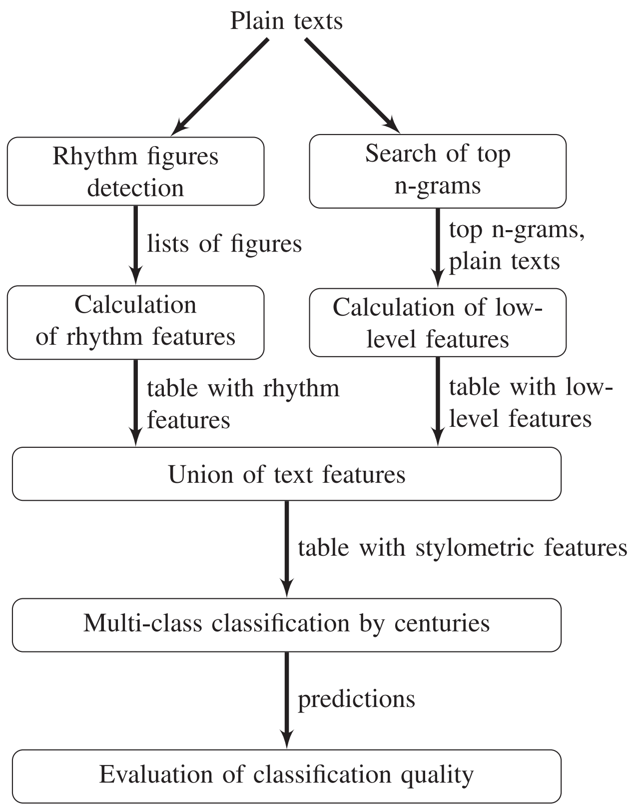

Fig. 1. Structure of experiments with classification

Firstly we extract rhythm figures from the plain texts using algorithms from our previous research [19] with accuracy of 80–95 % and calculate rhythm features from Section III separately for each text.

In parallel, we compute character- and word-based features for plain texts. We find top n-grams by their appearance in a text corpus, then for each text we compute independently the fractions of their use and other low-level features from Section III.

After feature calculation we unite results for each text. In such a way we represent every text as a vector of numerical stylometric features that are the same for all texts.

Finally we perform multi-class classification of texts by centuries with two machine learning algorithms and two neural networks and evaluate the predictions of the classifiers.

B. Classification and evaluation of results

The vectors of stylometric features are given as inputs to four supervised classifiers:

- AdaBoost classifier—a supervised machine learning meta-algorithm that combines the results of 50 Decision Tree classifiers adjusting incorrectly classified texts;
- Random Forest classifier—a supervised machine learning meta-algorithm that averages the results of 50 Decision Tree classifiers;
- Bidirectional LSTM—a recurrent neural network with a Bidirectional Long Short Term Memory (LSTM) layer

TABLE I. TOTAL WORD NUMBERS IN TEXT CORPORA

Language	Min	Median	Mean	Max
Russian	13 417	66 815	78 493.12	428 600
English	10 580	95 528	99 083.35	323 578
French	12 647	63 521	80 121.88	504 737

with 64 units and a dense output layer that uses the softmax activation function;

- GRU—a recurrent neural network with a Gated Recurrent Unit (GRU) layer with 4 units and a dense output layer that uses the softmax activation function.

These four algorithms often demonstrate the high quality of the text classification [20], [21], so we have chosen them for our experiments.

All the algorithms are trained on a half of a text corpora. The text corpora sizes are not large, that is why we test classifiers on the significant fraction of samples. For neural networks training we apply categorical cross-entropy as a loss function and Adam as an optimization algorithm.

The results of the test phase of multi-class classification were evaluated with four common measures: accuracy, macro-average precision, recall, and F-score. Accuracy is a number of right predictions for all classes divided by the number of all predictions. The macro-average precision and recall are average values of precision and recall values for particular classes. The macro-average F-score is a harmonic mean of the macro-average precision and recall [22].

The algorithms for the feature extraction, text classification, and results evaluation are implemented in the ProseRhythmDetector tool, which is available on the Internet at https://github.com/text-processing/prose-rhythm-detector. It is written in Python programming language and uses StanfordNLP 0.2.0 and TextBlob 0.15.2 NLP libraries for text representation, calculation of n-grams, and determination of parts of speech. For the classification it applies Scikit-Learn 0.21.3 and Keras 2.3.1.

V. EXPERIMENTS

We conducted the experiments with three text corpora in English, Russian, and French languages.

English and Russian corpora have 240 prose fiction texts of about 90 famous authors. The French corpus contains 150 prose fiction texts of about 40 famous authors. We took from 1 to 5 texts of each author. All texts are marked by a publication date from 1815 to 2019 for British literature, from 1832 to 2019 for Russian literature, and from 1826 to 2019 for French literature. Main statistical features: minimal, median, mean, and maximal numbers of words in texts are presented in Table I.

The texts vary in length significantly, that is why we apply for their classification only the features that do not depend on text size.

Before the experiments with time periods we classified the texts by three languages to estimate how languages differ by rhythm. The result quality reached 100 % of the accuracy and F-measure for neural network classifiers and every type of stylometric features, and 98–99 % for Random Forest and AdaBoost. Therefore, the texts in different languages are radically different in style and using only stylometric features can determine the text language practically precisely.

Then we classified the texts for each language separately by centuries. We compared four classifiers: Random Forest, AdaBoost, the neural network with the Bidirectional LSTM layer, and the neural network with the GRU layer. To each classifier we gave input vectors with all types of stylometric features: character-, word-, and rhythm-based.

The quality of classifiers is compared in Table II. We can see that for all languages neural networks outperform Random Forest and AdaBoost meta-classifiers. The accuracy and F-measure for meta-classifiers is less than 80 %, while neural networks provide from 82 to 89 % in the best cases.

The highest classification results for English literature are presented in Tables III and IV. They correspond to the classification by centuries and half-centuries for the best classifiers: LSTM and GRU networks. In each table the first column denotes the used neural network classifier, the second one—types of features or their combination (pair of feature types or all three types of features). The last column contains accuracy, macro precision, recall, and F-measure.

For the classification of English-language texts by centuries we can see the following tendency: combinations of several types of stylometric features improve the classification quality. Each type of features provides not more than 74.1–74.4 % of all metrics, whereas the use of two features allows to reach 75.9–86.0 %. The combinations with rhythm reach better quality than union of low-level character- and word-based features. And the greatest results are shown by combination of all features: 84.2–89.5 % of the accuracy and 84.0–89.6 % of the F-measure. Besides, the LSTM network outperforms the GRU network: LSTM results for union of all features and word- with rhythm-based are higher than all the GRU measures.

If we classify texts by 5 half-centuries: the first half of the 19th century, the second half of the 19th century, the first half of the 20th century, the second half of the 20th century, and the 21st century, the tendencies remain the same. The more feature types we combine, the better results we get. But in absolute values accuracy, precision, recall, and F-measure are less than for the classification by centuries. They reach only 71.0–78.9 %. Most probably, this decrease is caused by fewer training examples for half-centuries.

The corresponding Tables for Russian prose are V and VI. Again, we can see that a single feature type provides quality not more than 76.2 % of the accuracy or F-measure. The combinations of two feature types reach 78.0–82.0 %, and the best results of 88.1–88.7 % of all measures are shown by all feature types. Among all feature types, word-based one provides better quality than others by 2–9 %. Besides, com-

TABLE II. CLASSIFICATION OF PROSE BY CENTURY WITH ALL FEATURES

Classifier	Language	Accuracy	Precision	Recall	F-measure
AdaBoost	English	75.4	78.4	75.2	76.8
RandomForest	English	73.7	75.6	79.8	77.7
LSTM	English	**89.5**	**89.8**	**89.5**	**89.6**
GRU	English	84.2	84.2	83.8	84.0
AdaBoost	Russian	67.8	78.9	68.0	73.0
RandomForest	Russian	79.7	80.4	79.3	79.8
LSTM	Russian	83.1	83.7	83.9	83.8
GRU	Russian	**88.1**	**88.1**	**88.7**	**88.4**
AdaBoost	French	68.4	78.0	70.5	74.0
RandomForest	French	71.1	82.5	73.8	77.9
LSTM	French	**84.2**	**82.6**	**82.6**	**82.6**
GRU	French	76.3	77.4	73.3	75.3

TABLE III. CLASSIFICATION OF ENGLISH-LANGUAGE PROSE BY CENTURY

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character	74.1	75.4	73.5	74.4
LSTM	Word	70.7	69.2	69.2	69.2
LSTM	Rhythm	70.0	70.5	70.9	70.7
LSTM	Character + Word	75.9	76.0	75.0	75.5
LSTM	Character + Rhythm	82.5	83.2	82.2	82.7
LSTM	Word + Rhythm	**86.0**	**85.9**	**85.7**	**85.8**
LSTM	All	**89.5**	**89.8**	**89.5**	**89.6**
GRU	Character	69.0	68.0	68.0	68.0
GRU	Word	74.1	72.2	72.6	72.4
GRU	Rhythm	68.3	71.1	71.6	71.3
GRU	Character + Word	77.6	77.3	76.9	77.1
GRU	Character + Rhythm	82.5	82.9	82.3	82.6
GRU	Word + Rhythm	78.9	78.7	78.8	78.7
GRU	All	**84.2**	**84.2**	**83.8**	**84.0**

TABLE IV. CLASSIFICATION OF ENGLISH-LANGUAGE PROSE BY HALF-CENTURY

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character + Word	67.2	63.2	54.6	58.6
LSTM	Character + Rhythm	77.2	66.8	68.6	67.7
LSTM	Word + Rhythm	70.2	63.1	59.5	61.3
LSTM	All	**78.9**	**73.1**	**71.0**	**72.0**
GRU	Character + Word	72.4	67.5	61.6	64.4
GRU	Character + Rhythm	70.2	48.0	50.4	49.2
GRU	Word + Rhythm	71.9	72.2	63.4	67.5
GRU	All	77.2	61.8	62.6	62.2

bination of character- and word-based features gives slightly better results than other pairs by 1–4 %. The GRU network outperforms the LSTM by 5 %.

The results for half-centuries are also significantly lower than for centuries. The best measure values reach only 72.1 % that is less by 17 % than the same cases for the classification by centuries.

The quality of French text classification is presented in Tables VII and VIII. The best classifier for French is the LSTM network, it outperforms GRU by 9–12 % for all measures. The combinations of features provide better results than the use of only one type. The best combination differs from English and

TABLE V. CLASSIFICATION OF RUSSIAN-LANGUAGE PROSE BY CENTURY

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character	72.9	73.1	73.9	73.5
LSTM	Word	76.2	75.4	76.0	75.7
LSTM	Rhythm	65.0	65.4	67.3	66.3
LSTM	Character + Word	78.0	79.4	79.2	79.3
LSTM	Character + Rhythm	78.0	77.9	79.1	78.5
LSTM	Word + Rhythm	78.0	77.6	78.2	77.9
LSTM	All	**83.1**	**83.7**	**83.9**	**83.8**
GRU	Character	66.1	63.3	65.9	64.6
GRU	Word	74.6	74.3	74.3	74.3
GRU	Rhythm	68.3	69.1	70.6	69.8
GRU	Character + Word	**81.4**	**81.4**	**82.0**	**81.7**
GRU	Character + Rhythm	79.7	79.3	79.3	79.3
GRU	Word + Rhythm	78.0	77.7	77.7	77.7
GRU	All	**88.1**	**88.1**	**88.7**	**88.4**

TABLE VI. CLASSIFICATION OF RUSSIAN-LANGUAGE PROSE BY HALF-CENTURY

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character + Word	69.5	68.9	67.6	68.3
LSTM	Character + Rhythm	69.5	60.8	54.9	57.7
LSTM	Word + Rhythm	67.8	64.0	65.6	64.8
LSTM	All	**71.2**	**72.1**	**70.6**	**71.3**
GRU	Character + Word	69.5	50.9	55.7	53.2
GRU	Character + Rhythm	61.0	58.8	59.4	59.1
GRU	Word + Rhythm	66.1	60.8	57.3	59.0
GRU	All	71.2	66.0	65.0	65.5

TABLE VII. CLASSIFICATION OF FRENCH-LANGUAGE PROSE BY CENTURY

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character	78.9	75.5	75.2	75.4
LSTM	Word	76.3	78.4	74.9	76.6
LSTM	Rhythm	65.8	61.2	60.3	60.7
LSTM	Character + Word	78.9	75.5	75.2	75.4
LSTM	Character + Rhythm	**86.8**	**86.9**	**85.6**	**86.3**
LSTM	Word + Rhythm	78.9	77.1	76.4	76.7
LSTM	All	**84.2**	**82.6**	**82.6**	**82.6**
GRU	Character	68.4	65.6	65.9	65.7
GRU	Word	60.5	65.7	56.1	60.5
GRU	Rhythm	68.4	69.1	64.0	66.4
GRU	Character + Word	68.4	74.0	67.3	70.5
GRU	Character + Rhythm	71.1	71.8	73.1	72.5
GRU	Word + Rhythm	71.1	69.6	67.5	68.6
GRU	All	76.3	77.4	73.3	75.3

Russian: it is the union of character- and rhythm-based features that allows to reach 85.6–86.8 %. The results for all features are very close: 82.6–84.2 %.

The quality of the classification by half-centuries is also lower: up to 68.4 % of the accuracy. Again, the combination of character- and rhythm-based features outperforms other feature types for the French language.

To sum up, for all three languages we can see the same

TABLE VIII. CLASSIFICATION OF FRENCH-LANGUAGE PROSE BY HALF-CENTURY

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character + Word	65.8	55.8	54.9	55.4
LSTM	Character + Rhythm	**68.4**	**63.8**	**66.1**	**65.0**
LSTM	Word + Rhythm	65.8	60.0	54.9	57.3
LSTM	All	65.8	55.0	56.0	55.5
GRU	Character + Word	57.9	38.4	47.8	42.6
GRU	Character + Rhythm	63.2	58.8	62.1	60.4
GRU	Word + Rhythm	63.2	43.0	50.4	46.4
GRU	All	60.5	62.9	50.5	56.0

TABLE IX. ERRORS IN CLASSIFICATION WITH LSTM AND ALL FEATURES

Language	Publication year	Author and name	Predicted century
English	1899	Maugham - Orientations	20
English	1931	Lovecraft - At the Mountains of Madness	19
English	1962	Lessing - The Golden Notebook	21
English	2007	McEwan Ian - On Chesil Beach	20
English	1997	Ian McEwan - Enduring Love	21
English	1973	Murdoch - The Black Prince	21
Russian	1911	Tolstoj - Otec Sergij	19
Russian	2010	Danilov - Gorizontal'noe polozhenie	20
Russian	1901	Stanyukovich - Duel' v okeane	19
Russian	1886	Stanyukovich - Beglec	20
Russian	2016	Pelevin - Lampa Mafusaila	20
Russian	2004	Aksenov - Vol'ter'yancy i vol'ter'yanki	20
Russian	2005	Mikushevich - Voskresen'e v tret'em Rime	20
Russian	2004	Pelevin - Svyashchennaya kniga oborotnya	20
Russian	1921	Sologub - Zaklinatel'nica zmej	19
French	2009	Musso - Que serais je sans toi	20
French	1998	Pancol - Encore une danse	21
French	2019	Levy - Le voleur d'ombres	20
French	1999	Cusset - Jouir	21
French	2000	Beigbeder - 99 francs	21
French	2007	Pennac - Chagrin d'école	20

tendencies for the classification by centuries. The more feature types we use, the greater classification quality we reach. The only exception is the French language where the combination of character- and rhythm-based features is slightly better than others. Besides, among single feature types character- and word-based features perform better than rhythm-based ones. But rhythm-based features themselves achieve quite good results of the accuracy 65–70 %.

Observing the quality of classification by half-centuries for all languages, we can conclude that the decrease of time periods reduces significantly the classification quality. Also the tendencies of efficiency of different feature types repeat the same tendencies for classification by centuries.

VI. DISCUSSION

Considering the influence of pairs of feature types on the classification quality, we can conclude that for English and French, the pairs with rhythm-based features give the best results. For the Russian language, on the contrary, a greater result is achieved by a combination of character- and word-based features. However, the other two pairs do not differ significantly by their measure values. Perhaps, it happens due to the specifics of word-formation process in English and French, as analytical languages, where a more strict syntactic structure is observed. The combination of such structure and smaller variety of forms of pronouns, and more active use of functional words that link the words in a sentence, lead to the increasing frequency of rhythm figures, for example, pronominal anaphora, that is the repetition of pronouns in the subject. To enhance the quality of the research it is necessary to continue the experiments with other centuries. For this it is important to involve more experts for more detailed analysis of the results.

The quality of the classification by half-centuries is less than by centuries. This is due to the increasing number of classes. In addition, the number of training samples for each class become smaller, since the text corpora is used, and it affects quality of all classifiers.

Separately, we would like to draw attention to the errors of the classifiers. Table IX shows the erroneously classified texts by the LSTM neural network using all feature types. A significant part of misclassified texts is at the turn of a century. The style of these works can objectively be more similar to the previous or next century.

The information about errors in the classification gives a direction for an additional research on the style of prose texts and their authors. These studies will help to understand the influence of stylometric features on the classification results, including computer linguistics methods.

VII. Conclusion

We compared three types of text stylometric features: character-, word-, and rhythm based ones, classifying prose texts by centuries and half-centuries. We processed three corpora in different languages: 240 English texts, 240 Russian texts, and 150 French texts, extracted all features for each text, and classified texts by 19th–21st centuries or their half-centuries using different combinations of feature types. The experiments showed that the use of all feature types provided significantly better results (up to 84.2–89.9 % of the accuracy) than use of one type and in most cases of two types. Rhythm-based features also became good markers for distinguishing time periods: their classification quality reached 65.0–70.0 % of the accuracy. The comparison of the classification quality of feature types and their combinations allowed to analyze the difference in style of texts in different languages.

The implementation of feature extraction algorithms and classification scripts are available as parts of the ProseRhythmDetector tool: https://github.com/text-processing/prose-rhythm-detector.

The experiment results reveal directions for the future investigations. The analysis of errors of classifiers can reveal the texts and the authors whose style differs from the contemporaries. Besides, the influence of rhythm features should be studied for other natural language processing tasks.

Acknowledgment

The reported study was funded by RFBR according to the research project No. 19-07-00243.

References

[1] J. L. Rowlett, "Ralph cohen on literary periods: Afterword as foreword," *New Literary History*, vol. 50, no. 1, pp. 129–139, 2019.

[2] 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.

[3] Y. Liu and T. Xiao, "A stylistic analysis for gu long's kung fu novels," *Journal of Quantitative Linguistics*, vol. 27, no. 1, pp. 32–61, 2020.

[4] R. Rubino, S. Degaetano-Ortlieb, E. Teich, and J. van Genabith, "Modeling diachronic change in scientific writing with information density," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 750–761.

[5] A. Jatowt and R. Campos, "Interactive system for reasoning about document age," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2471–2474.

[6] X. Han and G. Toner, "Dating texts by multi-class classification with sliding time intervals," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2017, pp. 1–6.

[7] Y. Zhao and C. Hauff, "Sub-document timestamping: A study on the content creation dynamics of web documents," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2016, pp. 203–214.

[8] A. Gopidi and A. Alam, "Computational analysis of the historical changes in poetry and prose," in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 2019, pp. 14–22.

[9] S. U. R. Khan, M. A. Islam, M. Aleem, M. A. Iqbal, and U. Ahmed, "Section-based focus time estimation of news articles," *IEEE Access*, vol. 6, pp. 75 452–75 460, 2018.

[10] S. U. R. Khan, M. A. ISLAM, M. Aleem, and M. A. Iqbal, "Temporal specificity-based text classification for information retrieval," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 6, pp. 2915–2926, 2018.

[11] A. Spitz, J. Strötgen, T. Bögel, and M. Gertz, "Terms in time and times in context: A graph-based term-time ranking model," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1375–1380.

[12] Z. Lin, X. Wan, and Z. Guo, "Learning diachronic word embeddings with iterative stable information alignment," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2019, pp. 749–760.

[13] F. Fukumoto and Y. Suzuki, "Temporal-based feature selection and transfer learning for text categorization," in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, vol. 1. IEEE, 2015, pp. 17–26.

[14] R. Cohen and J. L. Rowlett, "On the presuppositions of literary periods," *New Literary History*, vol. 50, no. 1, pp. 113–127, 2019.

[15] O. Popescu and C. Strapparava, "Semeval 2015, task 7: Diachronic text evaluation," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 870–878.

[16] V. Niculae, M. Zampieri, L. P. Dinu, and A. M. Ciobanu, "Temporal text ranking and automatic dating of texts," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, 2014, pp. 17–21.

[17] S. Štajner and M. Zampieri, "Stylistic changes for temporal text classification," in *International Conference on Text, Speech and Dialogue*. Springer, 2013, pp. 519–526.

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

[19] 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.

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

[21] J. Nowak, A. Taspinar, and R. Scherer, "LSTM recurrent neural networks for short text and sentiment classification," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2017, pp. 553–562.

[22] 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.