A Survey on Stylometric Text Features

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

Elena Boychuk, Inna Vorontsova,
Elena Shliakhtina, Olga Belyaeva
Yaroslavl State Pedagogical
University named after K.D.Ushinsky
Yaroslavl, Russia
elena-boychouk@rambler.ru, arinna1@yandex.ru,
elenav_yar@mail.ru, olbelyaeva@yandex.ru

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

Abstract—**Ways of individual style expression in a natural language include amongst other things stylometric features. These can be automatically detected with the use of computational linguistics methods. In this survey we systematize the recent studies devoted to extraction and application of stylometric features in solving natural language processing tasks: authorship attribution, authorship verification, style change detection, authorship profiling, and text classification by genre and sentiment. For that purpose we define stylometric feature categories that provide for the most effective solutions, discuss reasons for their successful application, touch upon the limitations of approaches based on their application, and make suggestions for future research.**

I. Introduction

Stylometry is a branch of computational linguistics that studies quantitative assessment of linguistic features in the natural language texts. Stylometry is closely related to the terms of author's individual style and idiolect that imply a system of language features used by the author. Language experts state that idiolect presents a complex of language features such as terms, figures of speech and syntax, with a particular emphasis on sentence structure, its size and type according to its purpose, etc. [1] It is the distinguishing features of an author's idiolect that can be statistically detected thanks to stylometry.

Stylometry methods are applied to a number of tasks of natural language processing (NLP) including authorship attribution, authorship verification, authorship profiling, style change detection, and classification of written texts. They are based on the assumption that it is possible to reveal text features, which definitely verify the authorship and, moreover, that text subject together with the functional and author's style make up the originality of the text [2].

The choice of text stylometric features is the most important study phase. The researchers single out about a thousand features at different levels of analysis: lexical (including the levels of characters and letters), syntactical, semantic, structural, and subject-specific levels [3], [4]. This is indicative of text complexity and multidimensionality that need relevant evaluation of the text units, selected for quantitative analysis, and their capacity to present the originality of an author's style.

Today there is no consensus on an optimal set of stylometric features. Their choice is mainly random and often depends on the applied classifier. Experts point out that the selection of features is one of the greatest issues in stylometry [3], [5].

Papers dedicated to feature type correlation are quite scarce. Still, they have established a connection between syntactic features of the text and its length, between semantic features, subject and authorship [6]. The research has so far focused on the quantitative measures of quality assessment, whereas little attention has been given to interpreting the results of computational stylometry method usage. If it were possible to explain a classifier solution that would help to understand why a text belongs to a particular author and refers to a particular genre or subject, it could improve the efficiency of solving tasks at hand [7].

A possible reason for the above issues is lack of integrity of stylometry methods and approaches that are used by experts in different research fields. Computer experts seldom take into account the language findings in theory of linguistic persona, text linguistics, stylistics. For their part, language experts do not take full advantage of the potential of quantitative methods that are applied in modern information theory. They use only simple calculations while working with the facts concerning relative predominance of one or another text feature. For this reason, we have set a task of systematizing the information about text stylometry features used for the study of authorship attribution, authorship verification, authorship profiling, style change detection, classification by text genre as well as analyzing the results of their implementation from a linguistic point of view.

The paper is structured as follows. In Section II we describe the main stylometric features. Section III is devoted to authorship attribution of texts from different categories: literature, journalistics, and Internet. In Section IV we observe the use of stylometric features for authorship verification. Section V describes style change detection with application of stylometry. In Section VI we analyze how researchers use stylometric features to construct authors' profiles. Section VII is devoted to text classification by genre or sentiment. In Section VIII we discuss advantages, disadvantages, and limitations of state-of-the-art approaches. Conclusion summarizes the paper.

II. Linguistic features

Here we discuss the main stylometric characteristics of a text considered when tackling linguistic and philological tasks. Parameters pertaining to the character and word level of a text, often referred to as text lexical features, are among most carefully scrutinized.

ISSN 2305-7254

At the character level, the text is presented as a sequence of characters, whereas the features themselves present the simplest document structure. N-gram defined as a contiguous sequence of *n* items from a given sample of text is a regular characteristic at the character level. The *n* value is often a varied parameter when fulfilling the task. Different *n* values are included in the common set. The optimum *n* is sometimes chosen with a view to other features depending on a language. Character n-grams are easily extracted. However, the number of different n-grams in a text can be very large, which affects the dimensionality of document representation for computer processing and increases the algorithm complexity. For this reason further optimization of such features is required. This is complemented by measuring the frequency of characters, lower- and upper-case letters, figures, and spaces.

At the word level, the text is often seen as a bag-of-words regardless of the word order, grammar or context. In such case word frequency, word character length, average word length, word n-grams and vocabulary richness are measured. This also requires detection of word boundaries and their comparison. Considerable preprocessing may be necessary for the effective detection of word-based features. It provides for text normalization and noise (e.g., spelling mistakes) elimination.

Syntactic features are based on sentence structure. Punctuation mark frequency, sentence length, average sentence length, and functional word frequency are among the simplest and most common. More complex characteristics include syntactic tree features.

Research in this field has explored a few other linguistic features, but they are scarce. This is most probably explained by the complexity of such feature measurement as well as by the specific nature of the study area. Semantic features reflecting word, phrase, or sentence meaning may serve an illustration. They are difficult to formalize and detect, as a consequence. On the other hand, structural features of a document, which are easy to reveal demonstrate their heavy dependence on the specificity of the task. This is starkly illustrated by their being employed to study texts of different genres: a research paper, an email, or a blog whose structures differ fundamentally.

It is important to point out that features considered in stylometry are not exactly the same as those applied by literary scholars and linguists to the research of an author's individual style.

In terms of philology, an individual (writing) style is a complex concept reflecting one's sociohistorical nature, ethnic, psychological, moral, and ethical peculiarities. A lot of researchers suggest a two-step analysis of an individual style that implies the study of the linguistic and literary (hermeneutic) aspects of a text, declaring this a proper synthetic approach that modern science upholds.

The first step involves the study of an author's idiolect, i.e., frequency and distribution in a text of various linguistic units. The second step is concerned with the author's individual writing style which is treated as a set of particular features of expression by way of the idiolectic features. Linguistic analysis of a text is the initial stage of its philological analysis. Thus, from the standpoint of linguistics, the basic idiolectic features are as follows:

- at phonetic level—distinctiveness of intonation and melodics, a particular number of syllables, vowels, and consonants repeated to enhance the power of expression, euphony, use of phrases providing for rhythm and harmony;
- at lexical level—phrases and set expressions, typical recurrent sentence parts (discourse patterns), favored terms, sayings and quotes, loan, dialectic and industry words, synonyms, antonyms, paronyms, neologisms, words denoting specific concepts;
- at syntax level—dominating sentence types (declarative, interrogative, exclamatory), one- or two-member sentences, complete and elliptical sentences, types of syntactic cohesion, syntactic parallelism, chiasm, sentence length.

Idiolectic features of artistic expression consist of tropes and stylistic figures typical of a given author.

The research is furthered by the scrutiny of the author's individual writing style itself and investigates into the author's personal background, the ideas and concepts developed in the text, the genre features, the text composition, image structure, intertextuality etc. This stage of text analysis is difficult to automate, which makes it a plausible reason for leaving it beyond the scope of stylometric research.

Apparently, word frequency and sentence length measurement alone is insufficient for an integral characteristic of an author's individual writing style. To make it complete, one needs to adopt a complex approach that would assist in effective merging the parameters of text analysis employed by philology and computational linguistics.

III. Style analysis for authorship attribution

Authorship attribution is the definition of the author of a given text. This task can be split into two subtasks: closed-set attribution, if the author is necessarily one from the given set, and open-set attribution, provided that the set of authors is not limited. In any case, the basis for solving the problem is a corpus of texts, whose authorship is known prior to classification [2]. In the case we dispose of a set of documents for each author in question, we can determine the features of an individual style and classify texts whose authorship is unknown. The attribution problem is challenging in many subject areas, frequently resolving itself into the definition of literary text, journalistic article, and Internet text authorship.

A. Authorship attribution of literary texts

The first attempts to use quantitative characteristics to determine the style of literary text authors were made at the end of the XIX century. Currently, a large number of studies is devoted to automatic determination of authorship of prose and poetic works.

One of the most successful approaches is application of adjacency networks that allows to achieve quite high results. Amancio [8] applied an adjacency network of words with graph characteristics as text features: degrees, accessibility, betweenness, assortativity, clustering coefficients, and an average shortest path length. Besides, the author preprocessed texts removing stopwords, but took such words into account by counting their frequency and intermittence. These text features as well as frequencies of character bigrams were added to the list of text features. Experiments showed that the proposed method outperformed methods based on simpler adjacency networks by 30–40 %. Unfortunately, the author did not provide absolute values of an accuracy, so that it is hard to compare the performance of his algorithm with other approaches.

Stanisz et al. [9] also used adjacency networks, but with words frequently appearing in texts, and their co-occurrences as vertices and edges' weights. Besides, the researchers computed various graph characteristics: clustering coefficients of vertices, an average shortest path length, an assortativity coefficient, and modularity. The classification of texts with the computation of all the features described provided an accuracy of 85–90 % for English and Polish books.

Segarra et al. [10] constructed adjacency networks for functional words in texts. Looking at them as text features, they considered functional words and their similarity whether their location was adjacent or isolated. Thus, the authors modeled a text as a graph (or a word adjacency network) with functional words as vertices and values of similarity measures as edges' weights. Similar graphs were considered as texts of the same author. Experiments adopting the method showed an accuracy exceeding 90 %, for corpora with a small number of authors, ranging two or three, and/or quite big text length: 25 000 words. The fewer words and the more authors the algorithm received, the lower accuracy it ensured: 35 % for 10 authors with 1 000 words. Therefore, the algorithm cannot be considered robust.

The use of functional words and other categories of words is a popular approach for authorship attribution. Just to name a few, Boukhaled and Ganascia [11] analyzed the efficiency of using sequential rules of functional words as style markers. Classic French literature (40 novels) being their case study, they investigated into the frequencies of functional words as the most reliable indicator of authorship. The algorithm segmented each text into a set of sentences based on punctuation marks, then extracted sequences of functional words. Each text was represented by a vector of normalized frequencies of functional word occurrence. In the end the authors used the SVM (support-vector machine) classifier. The method achieved a nearly perfect attribution performance: the best F-measure was about 95 %.

Ramezani et al. [12] evaluated the effect of 29 textual features exercised on the accuracy of author identification on Persian corpora in 30 different scenarios. Several classification algorithms were used on the corpora with 2, 5, 10, 20, and 40 different authors and a comparison was made. The author studied character and word n-grams; character, word, and parts of speech frequency; word and sentence length. The evaluation results showed that the information about the words and verbs used were the most reliable criteria for authorship accuracy tasks. Besides, NLP based features were more reliable than bad of words based features.

Ferracane et al. [13] introduced a new method to embed discourse features in a Convolutional Neural Network text classifier. The researchers have tested several featurization methods in order to define the conditions under which discourse features contribute to non-trivial performance gains, and have analyzed discourse embeddings. This paper explores an effective method to (1) featurize the discourse information, and (2) integrate discourse features into the character-bigram CNN text classifier. In order to carry out the research they chose to use the entity-grid model as it captures coreference chains which are critical to improving the performance on this task. Two approaches were taken to represent salient entities: grammatical relations (GR) and RST (Rhetorical Structure Theory) discourse relations. The discourse embedding proved to be a superior featurization technique as it increased the F1 score by a noticeable amount. This study revealed the overall superiority of RST features over GR features in larger and more difficult datasets: average accuracy and F-measure up to 99 % for 50 authors and 250 novels.

Other linguistic features that also characterize an author's style are rhythm features. Dumalus and Fernandez [14] explored the writer's rhythm as a possible style marker using a simple Naive Bayesian Classifier and a collection of 587 texts of 51 authors from the Gutenberg634 corpora. Each text in the input data was processed by a program to create a file representing the text as a sequence of stresses and pauses. For the purpose of constructing the lexical stresses of the words in the texts, the CMU Pronouncing Dictionary was used to create a lexical stress string for each respective word. These stress strings are classified as unstressed, with primary stress, and secondary stress. The classification demonstrated the achieved accuracies of roughly 50 % for most authors and about 90 % for Shakespeare.

Plecháč et al. [15] used rhythm features to determine the authorships of poetic texts. These features included the frequencies of the stressed syllables at particular metrical positions and the frequencies of particular sounds. The authors conducted experiments with four corpora of poetic texts: Czech, German, Spanish, and English. They used versification rhythm features with different classifiers: Burrows' Delta, Argamon's Quadratic Delta, Smith-Aldridge's Cosine Delta, and SVM. The method provided the best precision of 84–99 % when proposed versification features were combined with single words and trigrams frequencies. The authors indicated reasons why rhythm analysis is useful in determining the author's style. Effective application of post popular stylometric features: words and n-grams, requires large amounts of data that can be found too rarely in practice, while rhythm features can be estimated in small corpora. The most powerful stylometric analysis is a combined analysis of lexicon and versification. In their statements about measuring rhythm and determining authorship, the authors relied on the opinions of expert linguists.

Hou and Huang [16] proposed an innovative and robust approach to stylometric analysis without annotation and leveraging lexical and sub-lexical information. In particular, they proposed to leverage the phonological information of tones and rimes in Mandarin Chinese automatically extracted from unannotated texts. The texts from different authors were represented by tones, tone motifs, and word length motifs as well as rimes and rime motifs. Tones and rimes and their bigrams were taken from different sentence positions. The method is also based on the use of SVM and Random forest classifiers. The accuracy of the method exceeded 85–90 %. From the experiment results Hou and Huang concluded that the combination of bigrams of rimes, word-final rimes, and segment-final rimes can discriminate texts from different authors efficiently.

Several authors investigated authors' styles basing only on statistical features. Zenkov [17] suggested a method of statistical analysis of texts applying the frequency distribution of the first significant digits in numerals in Russian texts. Benford's law was found to hold approximately for these frequencies with a marked predominance of the digit 1. The author concluded that Benfords law held approximately for coherent texts. Deviations from Benford's law were statistically significant author features that allow, under certain conditions (the most important of which is a sufficient length), to distinguish between the texts with a different authorship. The actual frequency of occurrence was higher than the probability according to Benford's law for significant digits 1, 2, 3; for the subsequent digits the situation was reversed.

In the research of Jamak et al. [18] the data collected by counting words and characters in around a thousand paragraphs of each sample book (6 books in Bosnian all together) underwent the principal component analysis performed using neural networks. The achieved results showed that every author leaves a unique signature in written text that can be discovered by analyzing counts of short words per paragraph. In their article the authors have demonstrated that based on analyzing counts of short words per paragraph authorship could be traced using the principal component analysis.

In both works the authors conducted statistical experiments to show that their methods were able to distinguish different author's styles. But they did not establish classification experiments that could allow to compare efficiency of their methods with others.

The comparative study of different attribution methods were organized during the PAN-2018 competition [19]. PAN-2018 was a scientific event from the series devoted to various tasks on text forensics and stylometry. For this contest the authors took fanfiction texts written by non-professional authors in five languages: English, French, Italian, Polish, and Spanish. Kestemont et al. set the task of cross-domain authorship attribution, where texts of known and unknown authors belong to different domains. Most of participants used n-grams of characters and words. Other types of applied features were complexity measures, word and sentence lengths, and lexical richness functions. As a classifier they used SVM, neural networks, and ensembles of different algorithms. As a result, simple approaches based on n-grams of characters/words were much more effective than more complex methods based on in-depth study and linguistic analysis of texts. In average the highest results were obtained for English and Spanish languages, while the Polish texts turned out to be the most difficult to analyze. Besides, experiments revealed that the number of candidate authors was inversely proportional to the attribution accuracy, especially when more than 10 authors were included in the dataset, while an increase of the number of texts in the training set improved the recognition accuracy. Such effects can be observed in the most works in this area.

For PAN-2018 corpora authorship attribution was positioned as cross-domain. However, the text dataset consisted of texts of a specific genre written by non-professional authors. Perhaps this was the reason why the best results were showed by relatively simple methods.

Llorens and Delany [20] also solved the problem of cross-language authorship attribution. They performed classification using the Random Forest method. As a feature vector for a text, the authors proposed a set of language-independent features that estimated the vocabulary of fragments with equal lengths from randomly selected texts. The features of random selection were chosen experimentally. Better results of experiments were achieved for larger numbers of samples: about 80–90 % of accuracy. Although syntactic features proved to be effective in classification, the authors did not find them absolutely sufficient to make successful author identification and proposed to combine them with others, for example, syntactic.

Summarily, in the field of literary texts, authorship attribution methods achieved the highest results on text corpora that are either relatively small or belong to the specific genre. Nevertheless, we can point out different stylometric features that demonstrate the great efficiency: not only lexical, syntactic, and rhythm ones, but also simple features based on n-grams.

B. *Authorship attribution of articles*

The determination of authors of journalistic texts is a task very close to authorship attribution of literary texts. It is used in journalism and forensics to determine plagiarism or authorship of anonymous articles. These texts belongs to publicistic genres, so they significantly differs from literature by authors' style.

Most of the works apply features based on n-grams. Stuart et al. [21] focused on the question of identifying or confirming the authorship of a text based on the known body of work. They studied 100 randomly-selected author texts of formal writing style (essays, articles) in English. Among other features they analyzed letter trigrams, letter bigrams, words, functional words, POS bigrams, POS tags, letters, POS trigrams, prepositions, word length, pronouns, conjunctions, etc. The authors chose five features as the best: top letter bigrams, top letter trigrams, functional words, part-of-speech bigrams and trigrams, because they allowed to achieve about 95 % of accuracy.

Sari et al. [22] utilize continuous representations for authorship attribution. The model presented in the paper learnt continuous representations for n-gram features via a neural network jointly with the classification layer. Experimental results demonstrated that the proposed model classified the articles on the state-of-the-art level: 70–75 % accuracy. According to the authors, character models were superior to word models. In particular, they found that models that employ character level n-grams appear to be more suitable for datasets with a large number of authors, while a steep decrease in the accuracy of word models occurred when the number of authors increased. The drop in accuracy of the character n-gram model was less pronounced. Character models also achieved a better result on longer datasets, which consisted of fewer authors. Combining word and character n-grams only produced a very small improvement on the dataset. The experimental results provided evidence that continuous representations are suitable for a stylistic (as opposed to topical) text classification task, such as authorship attribution.

Character n-grams have been identified as the most successful feature in both single-domain and cross-domain authorship attribution, but the reasons for their discriminative value were not fully understood. Sapkota et al. [23] identified subgroups of character n-grams that corresponded to linguistic aspects commonly claimed to be covered by these features: morpho-syntax, thematic content, and style. They evaluated the efficiency of each of these groups in two authorship attribution settings: a single domain setting and a cross-domain setting where multiple topics were presented demonstrating that character n-grams that captured information about affixes and punctuation account for almost all of the power of character n-grams as features. Algorithms with these features achieved 78 % of accuracy for articles from the CCAT 10 corpus, but only 57 % for Guardian articles. The authors concluded that applying n-grams according to their linguistic aspect can also be beneficial for other classification tasks, for example, native language identification, document similarity and plagiarism detection.

In other works lexical features are analyzed. Gómez-Adorno et al. [24] classified the corpus C10 that is a subset of the Reuters Corpus Volume 1 of news articles. They constructed the integrated syntactic graphs for texts and extracted textual features from them: numbers of words, POS tags, dependency tags, combinations and permutations of vowels, suffixes, and synonym expansion. Then they applied non-supervised classification approach that consisted in calculation of a similarity measure for texts represented as vectors of syntactic graph features. The best accuracy 68 % was achieved with the combinations of all these features. It is quite low comparing with state-of-the-art in authorship attribution.

Stamatatos [25] presented a novel method that enhanced authorship attribution efficiency by introducing a text distortion step before extracting stylometric measures. The proposed method attempted to mask topic-specific information that was not related to the personal style of authors. Based on experiments on two main tasks in authorship attribution, closed-set attribution and authorship verification, the authors demonstrated that the proposed approach could enhance existing methods especially under cross-topic conditions, where the training and test corpora did not match in topic. The proposed algorithms transformed texts into a form where topic information was compressed while textual structure related to personal style was maintained. These algorithms were language-independent, did not require complicated resources, and could easily be combined with existing authorship attribution methods. But they revealed the significant limitation when tested on cross-topic and cross-genre corpora. The algorithms with character n-grams showed quite good accuracy about 80 % for cross-topic texts, but only 50–60 % for cross-genre ones. Experimental results demonstrated a considerable gain in efficiency when using the proposed models under the realistic cross-genre conditions. On the other hand, when the corpora were too topic-specific where the texts by a given author were consistently on certain subjects different than the ones of the other candidate authors, the distortion methods seemed not to be helpful.

Sundararajan and Woodard [26] conducted extensive research into the role of syntax and lexical words (nouns, verbs, adjectives, and adverbs) in representing style. Purely syntactic language model has been used to study the significance of sentence structures in both single-domain and cross-domain attribution, i.e. cross-topic and cross-genre attribution. Apart from syntactic models, the researchers have studied the role of word choice. In order to do it, they performed attribution by masking all words or specific topic words corresponding to nouns, verbs, adjectives, and adverbs. They highlighted the considerable influence of common nouns and proper nouns in attribution, thereby stressing the topic interference. A syntactic language model was obtained by constructing the probabilistic context-free grammar for each author using the constituency parse trees of sentences in their training posts. The experiments with Guardian articles showed that the method achieved 67–70 % or less of F-measure and accuracy. The authors note that authorship attribution approaches in literature focus mostly on single-domain attribution where content and style are highly entangled. Further analysis shows that syntax may be useful with cross-genre attribution while cross-topic attribution and single-domain attribution may benefit from both syntax and lexical information.

Researching the attribution of journalistic articles shows similar results to attribution of literary texts: the highest scores for small corpora and the decline of quality for corpora with bigger numbers of texts and authors, and with different domains/topics or genres.

C. Authorship attribution of short texts and e-mails

Due to the development of the Internet with its specific content, there appeared the task of determining authorship of short texts, such as emails, comments, blogs, reviews, etc.

The researchers frequently use sets of features from different categories. Cristani et al. [27] presented the results of the experiment performed over a corpus of dyadic chat conversations (77 individuals in total) collected with Skype in Italian. The conversations were modeled as sequences of turns, where "turn" means a stream of symbols and words (possibly including "return" characters) typed consecutively by one subject without being interrupted by the interlocutor. The authors presented the analysis of linguistic features united in several groups: lexical, syntactic, structural, content-specific, idiosyncratic. In lexical group on the word level they explored a total number of words, a number of short words, characters in words, characters per word, frequency of stop words. On the character level they studied a total number of characters, a number of uppercase, lowercase, digit characters, frequency of letters, and special characters. On the digit n-grams level the letter digit n-grams were counted, on the level of word-length distribution the histograms and the average word length were determined and the vocabulary richness was estimated by the study of hapax legomena and dislegomena. In syntactic group the authors determined the frequency of functional words, the occurrence of punctuation marks (!, ?, :), emoticons, and acronyms. The structural group was presented by message level such as greetings, farewell, signature; the content-specific group characterized word n-grams (bags of word, agreement (ok, yeah, wow), discourse markers, onomatopee (ohh), stop words, abbreviations, gender age-based words, slang words). Finally, they observe the idiosyncratic group that includes misspelled words. The authors underline that conversational features increase the matching probability of around 10 %; conversational features alone give higher performance of standard stylometric features, calculated over the whole set of turns, and not over each one of them. The experiment showed that the Authorship Attribution performance, measured with the area under the cumulative match characteristic curve, was 89.5 %.

Seroussi et al. [28] classified judgments, emails, reviews, and blogs. They proposed a set of text features that included not only simple statistical ones: words, stopwords, and authors occurrences and their numbers, the vocabulary size, but also features that described topics of texts. There were priors of the Dirichlet and beta distributions for words and authors in texts by topics. Topics were derived from texts using LDA or AT algorithms, or their combinations. The classification with the SVM method showed an accuracy more than 90 % for small datasets with judgments, but 50–60 % for emails. For large datasets results were lower for datasets with many authors and different topics: 40–45 %, and about 90 % for a specific corpus with a relatively small number of authors and topics. These results showed that the proposed algorithm highly depends on numbers of authors, topics, and genres of texts.

Sharma et al. [29] focused on the study of short online texts, retrieved from WhatsApp messaging application and studying the distinctive features of a macaronic language (Hinglish), using supervised learning methods and then comparing the models. Such features as word n-gram and character n-gram were classified by Nave Bayes, SVM, Conditional Tree, and Random Forest algorithms. The results showed that SVM attained a test accuracy of up to 95 % while similarly, Nave Bayes attained an accuracy of up to 94.5 % for the corpus. Conditional Tree & Random Forest failed to perform as well as expected. Word unigram and character 3-grams features were more likely to distinguish authors accurately than other features.

N-grams and another low-level features showed their efficiency in several research. Johnson and Wright [30] solved the task of forensic authorship attribution of emails. They assumed that every native speaker had an individual ideolect, and that ideolect could be identified through search for n-grams. Then the authors computed the Jaccard's similarity coefficient for emails comparing n-grams that appeared in texts. The more shared n-grams emails had, the higher probability was that they belonged to the same author. The accuracy of such a classifier reached 80–90 %, but only for the author whose emails made up a fraction in the sample no less than 10 %. The authors noted that polite words like "please", "thank you", and n-grams with them became the most helpful features for identification of the author of emails.

Sari et al. [22] classified with n-grams not only articles (see III-B), but also reviews. Their results for these corpora were better than for articles: 94 % accuracy that is close to the state-of-the-art level for short texts.

Ruder et al. [31] used convolutional neural networks (CNN) for large-scale authorship attribution of emails, reviews, blogs, comments, and tweets. They split texts into characters and represented them as a concatenation of their embeddings. Such feature vectors were processed by the multi-channel CNN. The approach showed the great results up to 85–95 % accuracy for emails, reviews, and tweets, but classified comments and blogs significantly lower: less than 60 % for 10 authors, and less than 50 % for 50 authors.

The algorithms demonstrate their dependency on corpora: the best scores for one category of texts do not repeat for others. The highest results are shown by n-grams and their combination with lexical features.

IV. Style analysis for authorship verification

Authorship verification, on the one hand, can be considered as a subtask of authorship attribution when it is necessary to solve a binary problem: whether the text belongs to a given author or not. On the other hand, it is often considered as an independent task due to its use in humanitarian and forensic research: resolving copyright disputes, identifying several pseudonyms of the same user on social networks, verifying the author [32].

Three research groups solved both verification and attribution task. Two approaches showed lower results. Seroussi et al. [28] (see algorithm description in Subsection III-C) solved the task reviewer identification using topic models and simple statistics. They achieved good results of an accuracy more than 70 % (15 out of 19 documents) only for a small dataset.

Stamatatos [25] (see III-B) used the text distortion preprocessing procedure, character n-grams, and word n-grams. Verification of PAN-2014 and PAN-2015 texts showed up to 75-80 % accuracy.

Gómez-Adorno et al. [24] (see III-B) could perform verification better. They verified prose using textual features from syntactic graphs. Experiments showed the best accuracy 83 % in the case when the authors constructed a common syntactic graph for all texts of each known author.

Other researchers achieved good results applying n-grams. Brocardo et al. [33] proposed a supervised learning technique combined with n-gram analysis for authorship verification in short texts taken from Enron e-mail corpus. They realized experimental evaluation of these texts involving 87 authors that yielded very promising results consisting of an equal error rate (EER) of 14.35 % for message blocks of 500 characters. The average number of words per e-mail was 200. The emails were plain texts and covered various topics ranging from business communications to technical reports and personal chats. The experiment demonstrated better results compared to the accuracy obtained using similar techniques in the literature.

Potha and Stamatatos [32] developed an intrinsic profile-based verification method that used latent semantic indexing. This approach was positioned as language independent. It applied popular low-level text features: word unigrams and character n-grams. Besides, the authors applied topic modeling to reduce dimensionality and create the better text model that represented all texts of the same author as a common vector (the approach implemented the profile-based paradigm). Then the method compared texts from the test set with such authors' vectors and marked the text as belonged to the author with the closest vector. The authors experimented with corpora of prose, newspaper articles, reviews, and other genres from PAN shared tasks of 2014 and 2015 in four languages: Dutch, English, Greek, and Spanish. The method achieved more than 80 % for the AUC (the area under the ROC curve) measure and outperformed algorithms from scientific contests PAN-2014 and PAN-2015.

Li et al. [34] made an attempt to apply domain-specific features. They developed algorithms and explored various classifiers to determine the authenticity of short messages on social networks (an average of 20.6 words) from Facebook. The authors studied the possibility of using standard machine learning methods to verify whether the specified user is the author of this message. They used 233 features including 227 stylometric features and six novel social network-specific features like character-based ones: numbers of alphabets, uppercase characters, special characters; word-based ones: the total number of words, average word length, the number of words with 1 char, etc.; syntactic ones: numbers of punctuation marks and functional words, the total number of sentences and many others. Such features are popular in state-of-the-art works. The set of social network-specific features included emoticons, abbreviations, starting a sentence without an uppercase letter, ending a sentence without a punctuation mark, and not mentioning "I" or "We" in the post. Such features are frequently used in the texts of authors in social networks. Experimental results showed an average accuracy of 79.6 % for 30 users and 9259 posts. This quality was achieved with stylometric features. Sentence-based features showed the worst performance of 53.6 % accuracy. Social network-specific features did not improve classification. The fact that the sentence-based features did not affect the classification quality can be explained by the peculiarity of short social network posts because they rarely consist of a large number of sentences. It is more interesting that special social network symbols influence the solution of this specific task much less than lexical and syntactic features. The possible reason is that the author's style depends primarily on lexical and syntactic features.

Summarily, the best features for verification are syntactic and lexical ones including n-grams. The combination of n-grams with more complex lexical features seems the most promising for future research.

V. STYLE ANALYSIS FOR STYLE CHANGE DETECTION

Style change detection task is, in a broad sense, the fact of changing the style of different documents or fragments of one document. This problem is solved in the case of determining the number of co-authors of the document, studying the change in the style of the writer over time [19].

From the point of view of text processing, the style of the text characterizes the purpose for which the text is written: artistic, journalistic, scientific, business, and conversational. Determining the style of a document is important for analyzing its structure, extracting knowledge, annotating, and machine translation [35].

One of the goals of the PAN-2018 [19] was style change detection, where single-author and multi-author English texts were to be distinguished. Five participants competed in solving this problem. Participants used grammatical structure and its features: document was "represented as a consecutive order of parse tree features, sentence-level: stop words, most/least frequent words or word pairs, and punctuation frequencies, statistical text features including number of sentences, text length, and frequencies of unique words, punctuations, or letters, character n-grams, word 1-6-grams" [19]. The best result of 89.3 % accuracy was obtained using several different groups of features: character-based, word-based, and sentence-based ones; and popular machine learning classifiers: SVM, Random forest, TF-IDF-based gradient boosting model, and logistic regression meta-classifier.

The task of determining a style change is very complex. The organizers of the competition reduced the classification to binary: they asked the participants to identify texts written by one or several authors. It seems indicative to us that even in a simplified version, this problem is better solved using lexical features.

Gómez-Adorno et al. [36] used stylometry-based approach in order to identify changes in the writing style of seven native English speaking authors of novels. Three stages of writing were defined for every author, each stage comprises three novels with a maximum of two years of difference between each publication. The researchers first ranked the novels chronologically, then defined three writing stages (initial, middle and final). As a part of further tests they conformed testing sets with three novels (1 per stage), and training sets with six novels (the remaining 2 stage). There were in total nine novels which constituted twenty-seven pairs of training-testing sets for each author. Various stylometric-based features such as lexical usage, punctuation and phraseology analysis, were used in the research. The authors built vector space models divided into three categories: phraseology analysis, punctuation analysis and lexical usage analysis. The evaluated machine-learning algorithms were SVM and Logistic Regression. This research revealed that the probability of assigning the correct class (writing stage) to a document at random is 33 %. The highest classification accuracy is obtained in average when using the combination of all stylometric features. Although, when the evaluation is performed individually, i.e. using only one type stylometric feature, the features within the punctuation analysis category yielded the best performance. The models built on punctuation-based features correctly classified the writing stage of a work above 77 % of the times for two authors. The obtained results show that that the writing stage of a literary work can be identified with high accuracy (more than 70 %), for four out of seven evaluated authors using different stylometric-based features.

The cycle of works of Kern, Rexha et al. [37], [38], [39], [40], [41] was devoted to the task of linking fragments of text with their real authors. Researchers suggest adding the algorithm of author attribution to the preprocessing of scientific publications, as a more detailed analysis of scientific authorship. The advantage of this approach is that scientific search engines can implement author search which allows researchers to specifically search for text passages written by a particular author. This more accurate attribution of the author allows to create profiles of researchers. These profiles will reflect the more detailed contribution of the author in various fields of science.

The authors performed text segmentation to identify potential copyright changes in the main text of a scientific article [38]. They applied stylometric characteristics to capture stylistic changes in the text, following the hypothesis that different authors can be identified by different writing styles in the document. The set of stylometric characteristics included fractions of characters from several categories (letter, uppercase. etc.) in the paragraph, the number of different words in the vocabulary, numbers of words occurring once and twice, vocabulary richness measures, average lengths of words in characters, of sentences in characters and words. A study of the distribution of various stylometric characteristics showed the difference between articles written by a different number of authors.

The next task was to predict the number of authors [39]. The best results (F-Measure 82.7 %) were achieved by the Random Forest classifier for a one-author article class. As the number of authors increased, the quality of classification decreased. The authors believed that this result may be due to two different aspects. The first aspect was related to the size of the contribution of each author. The smaller the amount of text that the author wrote, the more difficult it was to distinguish it from the contribution of other authors. The second aspect was related to the actual contribution to the text: the more authors there were, the greater the probability was that some of them did not participate in the writing of the article.

Two works [40], [41] set the goal to better understand how people evaluate the authors writing style, and which content-independent, stylometric characteristics they prefer to use to identify the author. The results showed that the task of distinguishing between different styles became difficult even for people. The authors “statistically compared the decisions against content features and content-agnostic features” [41], but could not explain why human annotators provided very different results. They gave detailed descriptions of the authors identification process and noted that such descriptions would be valuable in trying to develop algorithms in areas such as plagiarism detection or forensic analysis. These works show the important role of lexical and syntactic features in determining the style of the author, and also indicate a promising area of research to identify stylometric characteristics with the help of experts.

The research shows that high quality of style change detection requires analysis and comparison of various textual features including statistical, lexical, and syntactic ones. The future investigation can add to this set rhythm features that now are understudied.

VI. Style analysis for author profiling

The task of determining an authors profile, based on the written text, means identification of explicit and hidden information about gender, age, nationality, educational level, psychological characteristics of the person. This analysis is important for marketers, sociologists, psychologists, and forensic scientists because it extracts information about people directly from raw texts.

Bergsma et al. [42] explored the style of scientific works. The authors determined whether an article is written by: (1) a native or non-native speaker, (2) by a man or a woman, and (3) in the style of a conference report or an article. The authors distinguished three types of text features. The first type considered the text as a “bag of words”, not taking into account grammatical and syntactic features. The second type consisted of traditional stylometric aspects: punctuation marks, stop words, Latin abbreviations, as well as stylistic meta-features: mean-words-per-sentence, mean-word-length, etc. The third type included the syntactic and grammatical features of the language like features of sentence parsing trees and many others. For example, use of rules for making grammatical construction helped to distinguish style features of native speakers and non-native speakers. The chosen features formed vectors for the linear, L2-regularized SVM classifier. The best results in all experiments were achieved due to using all three types of features. The best F-measure was obtained when solving the problem of determining the style of the native and non-native speakers (91.6 %). The gender was determined worse than other profile characteristics: 48.2 % of F-measure. The F-measure in the classification by articles and reports was 66.7 %. This is one of the few works that includes options for parsing sentences. The authors showed that the use of such features improved the classification quality of the author’s style of scientific works.

Ashraf et al. [43] presented a stylometry-based approach for detection of author traits (gender and age) for cross-genre author profiles. There were used different types of stylistic features including 7 lexical features (average sentence length in characters, average sentence length in words, average word length, percentage of question sentences, total number of words, total unique words and words ratio of length), 16 syntactic features (number of adjectives, nouns, foreign words, and other word categories; POS unigram, bigram, and trigram density), 26 character-based features and 6 vocabulary richness. The system was trained using all the 56 features and different machine learning algorithms were explored including Random Forest, J48 and LADTree. Using the proposed approach, promising results were obtained on the training dataset (98.3 % for age, 78.7 % for gender and 78.0 % for both (jointly identifying age and gender)). On the test data set, the proposed approach obtained accuracy of 37.1 % for age, 57.6 % for gender and 25.6 % for both.

Melka and Místeckỳ [44] studied H. Beam Pipers classical story Omnilingual and focused on measuring vocabulary richness. In order to capture stylistic features of the novelette, a number of quantitative indicators were drawn in. The study is concentrated on vocabulary-richness indexes, a complex assessment of activity (Busemanns coeffcient, the chi-square testing classification), and a sketch of the Belza chain analysis. Frequencies of distinct words in individual sections were counted and evaluated. The degree of vocabulary dispersion in a text and relative frequency of a given word were measured. The statistical indicators show that H. Beam Piper has on the whole an estimable level of vocabulary richness.

In their exploratory study, De Bruyne et al. [45] investigated whether texts of Dutch-speaking adolescents with Autism spectrum disorder (ASD), aged 12–18, were automatically distinguishable from texts written by typically developing peers. First, they revealed the fact that specific characteristics could be found in the writing style of adolescents with ASD, and secondly, they examined the possibility to use these features in an automated classification task. They looked for both surface features (word and character n-grams, and simple linguistic metrics) and for deep linguistic features (syntactic, semantic, and discourse features). The differences between the ASD group and control group were tested for statistical significance and the authors showed that mainly syntactic features were different among the groups, possibly indicating a less dynamic writing style for adolescents with ASD. For the classification task, a Logistic Regression classifier was used. The best combination in the deep feature approach originally reached an F-score of just 62.14 %, which could not be boosted by automatic feature selection. However, by taking into account the information from the statistical analysis and merely using the features that were significant or trending, the authors could equal the surface-feature performance and again reached an F-score of 72.15 %. This suggested that a carefully composed set of deep features was as informative as surface-feature word and character n-grams. Moreover, combining surface and deep features resulted in a slight increased in F-score to 72.33 %.

The results of profiling algorithms are lower than for attribution or verification because profiling requires extracting specific information from texts. Presumably, the quality can be increased by use of more various linguistic features, for example, semantic and rhythm ones.

VII. Style analysis for classification by genre or sentiment

The task of text classification includes many subtasks that can be solved applying stylometric features. The most popular is classification by genre where there can be different sets of genre categories, for example, prose and verse, speeches and essays, prose of several types, etc. Besides, stylometric features are investigated in sentiment classification.

Gianitsos et al. [46] described their approach to applying stylometric classification of Ancient Greek texts by genre. The authors suggested using some language-specific stylometric features to classify texts, such as prose and verse. The feature set consisted of 23 features: function or non-content ords like pronouns and syntactical markers; rhetorical functions like questions and uses of superlative adjectives and adverbs. Three criteria were used: amenability to exact or approximate calculation without the use of syntactic parsing, substantial applicability to the corpus, and diversity of function. The authors performed 400 trials of stratified 5-fold cross-validation for a corpus of classical Greek literature. As a result of their research the authors managed to classify Ancient Greek literary texts as prose or verse with accuracy and F1 score more than 97 %.

Balint et al. [47], [48] classified texts by other genre categories: speeches, essays, and newspaper articles. They found in the texts a set of various rhythm features: lexical, grammatical, phonetic, metrical, and organizational. Since there were too many features, the authors identified the most predictive ones using the Discriminant Function Analysis. So they chose eight features: numbers of syllables per word, words deemed frequent; normalized numbers of sentence anaphora, punctuation unit anaphora, and commas; the percentage of falling word-length patterns, frequent words at the end of sentences and at the beginning of punctuation units. The classification quality achieved about 81 % of accuracy.

Amancio [8], whose method of authorship attribution we discussed above (see III-A), uses an adjacency network of words, stopwords, and bigrams to classify texts by style into informative or imaginative prose. The method outperforms similar methods based on simpler adjacency networks by 20-30 %.

The research of Anchiêta et al. [49] is a comparative study of different text features used for sentiment classification. The baseline features were TF-IDF and Delta TF-IDF, they were compared with stylometric ones: different word statistics like total number of words, characters, and their types; syntactic features like punctuation marks and parts of speech; content-specific features like adjective synonyms. The authors classified 2000 reviews about Smartphones using three algorithms: SVM, Nave Bayes and J48. This study revealed that better result was with the SVM classifier: 82,75 % of accuracy with stylometry.

The best quality of text classification by genre or sentiment is achieved using only large sets of various stylometric features: not only n-grams and word-level statistics, but also syntactic features and more complex lexical and rhythm markers.

VIII. Discussion

We compared all the described works in Table I. The first column contains marks for different tasks according to this paper' sections: authorship attribution (AA), authorship verification (AV), style change detection (SC), authorship profiling (AP), and classification by genre and sentiment (C). The works in the second column appear according to their order in the sections.

For the table we chose the most popular stylistic features and combined them into the categories (character-level, word-level, syntactic, semantic, topical, and rhythmic) and extracted subcategories. The character-level features include n-grams and character frequencies, which are common for natural language processing approaches, character types like letters, digits, etc., and the specific feature "character embeddings" that is mentioned only once in the Ruder's work [31]. The word-level features contain word frequencies, n-grams, usage of stop words and word parts, detection of errors, and measures of vocabulary like vocabulary-richness indexes. The subcategory "token-based" features unites a wide variety of word features that are separately less frequent than others, e.g., word length, number of occurrences of particular word types, etc.

These two categories consist of features that can be computed very simply, quickly, and fully automatically by NLP algorithms. It is one of the main reasons why they are the most common in authors style analysis as we see from the table.

The next four categories include linguistic features that are especially interesting for our research. Syntactic features are subdivided into the sentence length, structure, punctuation, and usage of different parts-of-speech (functional words, nouns, adjectives, etc.). As semantic features we picked out synonyms and RST-relations. Topical features are split into text-level and word-level ones depending on what amount of data belongs to the particular topic: different words or a text as a whole. Rhythm features include repetition of parts of words, words, and word groups, detection of rhyme and stresses, and features describing syllables. Most of these linguistic features are extracted from texts not as accurate as statistical ones, so they are used less frequently.

TABLE I. MOST POPULAR FEATURES FOR STYLOMETRIC ANALYSIS

Task	Work	Character-level				Word-level							Syntactic				Semant.	Topical		Rhythmic			
		em-bed.	char freq.	n-grams	char types	word freq.	n-grams	token-based	voca-bulary	er-rors	stop-words	word parts	parts-of-speech	sent. length	struc-ture	punc-tuation	rela-tions	text-level	word-level	repe-tition	rhyme	syl-lable	stress
AA	Amancio [8]			+		+		+			+		+										
AA	Stanisz et al. [9]					+		+															
AA	Segarra et al. [10]												+		+								
AA	Boukhaled and Ganascia [11]												+		+								
AA	Ramezani et al. [12]		+	+	+	+	+	+					+	+									
AA	Ferracane et al. [13]														+		+		+				
AA	Dumalus and Fernandez [14]																						+
AA	Plecháč et al. [15]		+			+															+	+	+
AA	Hou and Huang [16]			+	+		+	+													+		
AA	Zenkov [17]				+																		
AA	Jamak et al. [18]		+			+		+															
AA	Kestemont et al. [19]			+			+																
AA	Llorens and Delany [20]					+	+	+															
AA	Stuart et al. [21]						+	+					+										
AA	Sari et al. [22]			+			+																
AA	Sapkota et al. [23]			+			+				+	+	+										
AA	Gómez-Adorno et al. [24]							+				+	+		+		+			+			
AA	Stamatatos [25]			+		+	+	+								+							
AA	Sundararajan and Woodard [26]												+		+				+				
AA	Cristani et al. [27]					+	+	+	+	+	+					+			+				
AA	Seroussi et al. [28]							+	+		+							+					
AA	Sharma et al. [29]			+		+	+	+		+									+				
AA	Johnson and Wright [30]						+																
AA	Ruder et al. [31]	+																					
AV	Seroussi et al. [28]							+	+		+							+					
AV	Stamatatos [25]			+		+	+	+								+							
AV	Gómez-Adorno et al. [24]							+				+	+		+		+			+			
AV	Brocardo et al. [33]						+																
AV	Potha and Stamatatos [32]			+																			
AV	Li et al. [34]			+		+		+			+		+										
SC	Kestemont et al. [19]			+		+		+			+		+	+		+							
SC	Gómez-Adorno et al. [36]							+	+		+			+		+							
SC	Kern et al. [37], [41]		+	+	+	+	+	+	+				+	+									
AP	Bergsma et al. [42]							+	+		+		+	+	+								
AP	Ashraf et al. [43]		+	+	+	+	+	+	+			+	+	+					+				
AP	Melka and Místecký [44]					+			+														
AP	De Bruyne et al. [45]			+			+	+					+	+									
C	Gianitsos et al. [46]				+		+	+				+	+	+	+	+							
C	Balint et al. [47], [48]						+	+						+	+					+		+	
C	Amancio [8]			+		+		+			+		+										
C	Anchiêta et al. [49]					+		+	+				+				+						

Thus, the number of stylistic features used in computer linguistics is very large and heterogeneous. However, researchers pay insufficient attention to systematization of these features, study of their influence on the quality of solving tasks and justification of feature choice. Most authors experimentally compare algorithmic approaches like [19]. Much less often, researchers set the task of studying the influence of various parameters on the quality of text classification by the author's style [15]. Almost none of the researchers consider the reasons why features or feature groups are relevant and efficient.

Most studies are devoted to authorship attribution and verification. The best results are obtained using lower-level features, such as n-grams and character and word frequencies, with which statistical characteristics of accuracy and F-measure reach 99 %. The reason for the successful use of n-gram characters is that they provide a compromise between the rarity of occurrence and information content. They can be used for every type of texts, and are convenient to compute statistical features. At the same time, they combine information about punctuation, morphology (n-grams of characters can represent both morphemes and word roots), vocabulary (the length of functional words is often small) and even context (when extracting n-grams at the sentence level, rather than at the word level). In addition, they are tolerant of spelling changes and errors. Besides, from a practical point of view, the models based on symbolic n-grams are very easy to build, and they are language-independent. Daelemans claims that the formation of symbol n-grams is carried out by a person at the subconscious level and is not amenable to control or imitation [7]. That can explain why the character n-gram approaches achieve one of the best accuracy scores for the state-of-the-art.

Comparing studies with the highest quality scores (about 90 % and higher) of algorithms with different feature categories, we can conclude that these results are most often achieved under one or more of the following conditions:

- a relatively small text corpus (not more than 200–250 texts), and the texts are quite voluminous in size;
- texts belong to a small number of authors, usually 10 or less;
- a large number of texts of a given author is analyzed, then one of the best classification results is obtained for this author;

- researchers successfully selected stylistic features according to which the classifier makes decisions, and the features may differ for texts with different topics and genres.

The problem of successful selection of stylistic features means the following. When we study a model that can distinguish two or more particular authors, there is no guarantee that this model generalizes the authors style and will be able to differ these authors from others [7]. Authorship attribution and verification among a large number of candidates and uneven training samples remain a challenge.

In our opinion, one of the reasons for this situation is that a set of stylistic features is selected on the basis of the "mathematical" approach: firstly, due to the fact that they are convenient to calculate, and secondly, they are traditionally used in computer linguistics. Coefficients and features of classifiers and text models that give the maximum result for a particular task, possibly even for a specific sample of authors and texts, are randomly selected for experiments.

Computational stylometry is essentially a way of studying the author's idiostyle. One of the main goals of this task is to describe and explain the cause-effect relationships between the psychological and sociological characteristics of the authors, on the one hand, and the style of their writing on the other. Experts obtain a significant part of the information about this from semantic and purely linguistic features (aspects of rhythm, synonyms). Some researchers notice a discrepancy between the feature used by experts in computer and classical linguistics and invite to discuss and conduct additional research on this issue [3], [41].

In addition, researchers most often take into account only some features of the idiolect or the linguistic specificity of an author's style, which consist, as a rule, in reflecting quantitative indicators of rather low-level text features, such as the number of words, syllables, sentence size, etc. However, the idiostyle is expressed in features that are rather complicated for the search and related to the personality of the author. The added complexity is that "there is no taxonomy or checklist of the elements of individual style, since anything can be an element of individual style if it is consistently used in such a way as to contribute to the expression of the personality of the author" [50].

Thus, the implementation of a comprehensive analysis of an authors individual style is a rather difficult task. The analysis of author's language specificity is only one of its many stages. Automating the search for these formal features is the first step towards a comprehensive understanding of an individual author's style.

IX. Conclusion

In this survey we have systematized stylometric features that describe an author's style of natural language texts. We have analyzed application of these features in solving text processing tasks: authorship attribution, authorship verification, style change detection, authorship profiling, and text classification by genre and sentiment. We have chosen approaches that use popular state-of-the-art features in order to investigate helpfulness of different feature types in determining of an authors individual style. Comparing approaches with the greatest results, we have chosen the best stylistic features, explained their popularity and efficiency, and discussed their drawbacks and limitations.

From our survey we can conclude that researchers in the field of stylometry primarily operate with features reflecting quantitative indicators of quite low-level text features. However, an author's style is often expressed in aspects that are quite difficult to search.

Problems that prevent the use of the specificity of the author's style in tasks of attribution, verification and others, are lack of information about the correlation of stylometric features with each other and of features with domains. Besides, the studies usually do not explain why chosen feature sets characterizes the style of a particular author. These tasks can be solved by studying peculiarities of stylistic features in the idiostyle of different authors, manifestations of stylistic features in different genres, and creation of text corpora with the markup of an author's style that can be used for deeper research.

Acknowledgment

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

References

[1] D. J. Bergman and C. C. Bergman, "Elements of stylish teaching: Lessons from strunk and white," *Phi Delta Kappan*, vol. 91, no. 4, pp. 28–31, 2010.

[2] 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.

[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, p. 86, 2018.

[4] J. Rudman, "The state of authorship attribution studies: Some problems and solutions," *Computers and the Humanities*, vol. 31, no. 4, pp. 351–365, 1997.

[5] P. Juola, "Future trends in authorship attribution," in *Advances in Digital Forensics III*. Springer, 2007, pp. 119–132.

[6] U. Sapkota, T. Solorio, M. Montes-y Gómez, and P. Rosso, "The use of orthogonal similarity relations in the prediction of authorship," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2013, pp. 463–475.

[7] W. Daelemans, "Explanation in computational stylometry," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2013, pp. 451–462.

[8] D. R. Amancio, "A complex network approach to stylometry," *PloS one*, vol. 10, no. 8, p. e0136076, 2015.

[9] 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.

[10] S. Segarra, M. Eisen, and A. Ribeiro, "Authorship attribution through function word adjacency networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5464–5478, 2015.

[11] M. A. Boukhaled and J.-G. Ganascia, "Using function words for authorship attribution: Bag-of-words vs. sequential rules," *Natural Language Processing and Cognitive Science: Proceedings 2014*, pp. 115–122, 2015.

[12] R. Ramezani, N. Sheydaei, and M. Kahani, "Evaluating the effects of textual features on authorship attribution accuracy," in *International eConference on Computer and Knowledge Engineering (ICCKE) 2013*. IEEE, 2013, pp. 108–113.

[13] E. Ferracane, S. Wang, and R. Mooney, “Leveraging discourse information effectively for authorship attribution,” in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, vol. 1, 2017, pp. 584–593.

[14] A. Dumalus and P. Fernandez, “Authorship attribution using writers rhythm based on lexical stress,” in *Proceedings of the 11th Philippine Computing Science Congress*, 2011, pp. 82–88.

[15] P. Plecháč, K. Bobenhausen, and B. Hammerich, “Versification and authorship attribution. a pilot study on czech, german, spanish, and english poetry,” *Studia Metrica et Poetica*, vol. 5, no. 2, pp. 29–54, 2018.

[16] R. Hou and C.-R. Huang, “Robust stylometric analysis and author attribution based on tones and rimes,” *Natural Language Engineering*, pp. 1–23, 2019.

[17] A. Zenkov, “New statistical method of text attribution,” *International Journal Of Professional Science*, no. 1, pp. 6–21, 2017.

[18] A. Jamak, A. Savatić, and M. Can, “Principal component analysis for authorship attribution,” *Business systems research journal: international journal of the Society for Advancing Business & Information Technology (BIT)*, vol. 3, no. 2, pp. 49–56, 2012.

[19] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, and M. Potthast, “Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection,” in *Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings*, vol. 2125, 2018, pp. 1–25.

[20] M. Llorens and S. J. Delany, “Deep level lexical features for cross-lingual authorship attribution,” in *Proceedings of the First Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine 2016) co-located with the 38th European Conference on Information Retrieval (ECIR 2016)*. Dublin Institute of Technology, 2016, pp. 16–25.

[21] L. M. Stuart, S. Tazhibayeva, A. R. Wagoner, and J. M. Taylor, “On identifying authors with style,” in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 3048–3053.

[22] Y. Sari, A. Vlachos, and M. Stevenson, “Continuous n-gram representations for authorship attribution,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, 2017, pp. 267–273.

[23] U. Sapkota, S. Bethard, M. Montes, and T. Solorio, “Not all character n-grams are created equal: A study in authorship attribution,” in *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2015, pp. 93–102.

[24] H. Gómez-Adorno, G. Sidorov, D. Pinto, D. Vilariño, and A. Gelbukh, “Automatic authorship detection using textual patterns extracted from integrated syntactic graphs,” *Sensors*, vol. 16, no. 9, p. 1374, 2016.

[25] E. Stamatatos, “Authorship attribution using text distortion,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 2017, pp. 1138–1149.

[26] K. Sundararajan and D. Woodard, “What represents style in authorship attribution?” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2814–2822.

[27] M. Cristani, G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, and V. Murino, “Conversationally-inspired stylometric features for authorship attribution in instant messaging,” in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 1121–1124.

[28] Y. Seroussi, I. Zukerman, and F. Bohnert, “Authorship attribution with topic models,” *Computational Linguistics*, vol. 40, no. 2, pp. 269–310, 2014.

[29] A. Sharma, A. Nandan, and R. Ralhan, “An investigation of supervised learning methods for authorship attribution in short hinglish texts using char & word n-grams,” *ArXiv*, p. 1812.10281, 2018.

[30] A. Johnson and D. Wright, “Identifying idiolect in forensic authorship attribution: an n-gram textbite approach,” *Language and Law*, vol. 1, no. 1, pp. 37–69, 2014.

[31] S. Ruder, P. Ghaffari, and J. G. Breslin, “Character-level and multi-channel convolutional neural networks for large-scale authorship attribution,” *ArXiv*, p. 1609.06686, 2016.

[32] N. Potha and E. Stamatatos, “Intrinsic author verification using topic modeling,” in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM, 2018, pp. 20–26.

[33] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, “Authorship verification for short messages using stylometry,” in *2013 International Conference on Computer, Information and Telecommunication Systems*. IEEE, 2013, pp. 1–6.

[34] J. S. Li, L.-C. Chen, J. V. Monaco, P. Singh, and C. C. Tappert, “A comparison of classifiers and features for authorship authentication of social networking messages,” *Concurrency and Computation: Practice and Experience*, vol. 29, no. 14, p. e3918, 2016.

[35] C. Jebari, “A segment-based weighting technique for url-based genre classification of web pages,” *Polibits*, no. 53, pp. 43–47, 2016.

[36] H. Gómez-Adorno, J.-P. Posadas-Duran, G. Ríos-Toledo, G. Sidorov, and G. Sierra, “Stylometry-based approach for detecting writing style changes in literary texts,” *Computación y Sistemas*, vol. 22, no. 1, pp. 47–53, 2018.

[37] R. Kern and M. Granitzer, “Efficient linear text segmentation based on information retrieval techniques,” in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM, 2009, pp. 167–171.

[38] A. Rexha, S. Klampfl, M. Kröll, and R. Kern, “Towards authorship attribution for bibliometrics using stylometric features,” in *CLBib@ ISSI*, 2015, pp. 44–49.

[39] ——, “Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features,” in *BIR 2016 Workshop on Bibliometric-enhanced Information Retrieval*, 2016, pp. 26–31.

[40] A. Rexha, M. Kröll, H. Ziak, and R. Kern, “Extending scientific literature search by including the author’s writing style,” in *BIR 2017 Workshop on Bibliometric-enhanced Information Retrieval*, 2017, pp. 93–100.

[41] ——, “Authorship identification of documents with high content similarity,” *Scientometrics*, vol. 115, no. 1, pp. 223–237, 2018.

[42] S. Bergsma, M. Post, and D. Yarowsky, “Stylometric analysis of scientific articles,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 327–337.

[43] S. Ashraf, H. R. Iqbal, and R. M. A. Nawab, “Cross-genre author profile prediction using stylometry-based approach.” in *Working Notes Papers of the CLEF 2016 Evaluation Labs*, 2016, pp. 992–999.

[44] T. S. Melka and M. Místeckỳ, “On stylometric features of H. Beam Pipers Omnilingual,” *Journal of Quantitative Linguistics*, pp. 1–40, 2019.

[45] L. De Bruyne, B. Verhoeven, and W. Daelemans, “Stylometric text analysis for dutch-speaking adolescents with autism spectrum disorder,” *Computational Linguistics in the Netherlands Journal*, vol. 8, pp. 3–23, 2018.

[46] E. Gianitsos, T. Bolt, P. Chaudhuri, and J. Dexter, “Stylometric classification of ancient greek literary texts by genre,” in *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2019, pp. 52–60.

[47] M. Balint, M. Dascalu, and S. Trausan-Matu, “Classifying written texts through rhythmic features,” in *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 2016, pp. 121–129.

[48] M. Balint and S. Trausan-Matu, “A critical comparison of rhythm in music and natural language,” *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information*, vol. 9, no. 1, pp. 43–60, 2016.

[49] R. T. Anchiêta, F. A. R. Neto, R. F. de Sousa, and R. S. Moura, “Using stylometric features for sentiment classification,” in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 189–200.

[50] J. Robinson, “General and individual style in literature,” *The Journal of aesthetics and art criticism*, vol. 43, no. 2, pp. 147–158, 1984.