algorithms

MDPI

Article

Digital Authorship Attribution in Russian-Language Fanfiction and Classical Literature

Anastasia Fedotova , Aleksandr Romanov *, Anna Kurtukova and Alexander Shelupanov

Department of Security, Tomsk State University of Control Systems and Radioelectronics, 634050 Tomsk, Russia
* Correspondence: ras@fb.tusur.ru

Abstract: This article is the third paper in a series aimed at the establishment of the authorship of Russian-language texts. This paper considers methods for determining the authorship of classical Russian literary texts, as well as fanfiction texts. The process of determining the author was first considered in the classical version of classification experiments using a closed set of authors, and experiments were also completed for a complicated modification of the problem using an open set of authors. The use of methods to identify the author of the text is justified by the conclusions about the effectiveness of the fastText and Support Vector Machine (SVM) methods with the selection of informative features discussed in our past studies. In the case of open attribution, the proposed methods are based on the author's combination of fastText and One-Class SVM as well as statistical estimates of a vector's similarity measures. The feature selection algorithm for a closed set of authors is chosen based on a comparison of five different selection methods, including the previously considered genetic algorithm as a baseline. The regularization-based algorithm (RbFS) was found to be the most efficient method, while methods based on a complete enumeration (FFS and SFS) are found to be ineffective for any set of authors. The accuracy of the RbFS and SVM methods in the case of classical literary texts averaged 83%, which outperforms other selection methods by 3 to 10% for an identical number of features, and the average accuracy of fastText was 84%. For the open attribution in cross-topic classification, the average accuracy of the method based on the combination of One-Class SVM with RbFS and fastText was 85%, and for in-group classification, it was 75 to 78%, depending on the group, which is the best result among the open attribution methods considered.

Keywords: authorship identification; natural language processing; machine learning; feature selection; fastText; support vector machine; genetic algorithm

Citation: Fedotova, A.; Romanov, A.; Kurtukova, A.; Shelupanov, A. Digital Authorship Attribution in Russian-Language Fanfiction and Classical Literature. *Algorithms* **2023**, *16*, 13. https://doi.org/10.3390/a16010013

Academic Editors: Melania Susi and Alwin Poulose

Received: 2 December 2022
Revised: 22 December 2022
Accepted: 22 December 2022
Published: 26 December 2022

1. Introduction

The identification of a social media user and the process of his authentication can be performed based on the outgoing text information. For text analysis, it is necessary to use natural language processing (NLP) methods. In this context, attribution of the user is one of the most relevant areas of NLP [1,2]. In general, attribution methods are initially applied to classical literary works [3] due to the classical writer's well-established writing style, based on which it becomes possible to select a feature space, clearly distinguishing one author from another. Such space includes a set of features carrying information about the distinctive peculiarity of the writer [4].

Fanfiction texts relate to fiction prose. Nowadays, these texts are created by fans of a particular work and its storyline (novel, television series, etc.). In most cases, authors of fanfics use the original characters, places, and style, and complement the text with their vision of the original work's plot lines by adding details. Fanfiction is published online in informal community platforms to make such literature accessible to a wide audience.

As an approbation, the features selected for classical literary texts can be applied to fanfiction prose written by authors with an actively progressive style [5]. The set of features can be reduced by removing uninformative and unstable features or by expanding with other authors' characteristics. The practical application of effective methods for the text's author attribution is useful in the following fields:

1. Forensics, in particular, for actions against pedophilia, extremism, enmity or hatred in social media, conducting expertise, solving disputes about copyright and intellectual property, as well as dispute resolutions for other forensics and information security problems.
2. Detection and resistance of plagiarism in the educational process.
3. Linguistic research, including research for social media marketing (SMM) purposes and in general linguistic fields.

Nowadays, copyright law includes the protection of the property and non-property rights of the creator. Since the copyright law came into existence, the institution has undergone significant changes in the protection of intellectual property. Significant modifications were made based on world digitalization and the resulting decrease of handwritten texts in favor of digital ones due to the ease of creating, transferring, and editing the latter. However, the creator's interest in the wide dissemination of his work to extract material benefits or gain popularity remains stable.

It is always in the author's interest to ensure that his work does not become a source of income for third parties or be subject to plagiarism. Despite this, many textual works are published in the public domain. Thus, the problem arises not only of appropriating the intellectual property of another creator but also of attempting to create texts on his behalf. Such texts may contain calls to an action prohibited by law, offenses, and other content that negatively affects the author's reputation. When solving these problems in general practice, it is customary to involve forensic experts. Such examinations can take a long time when there are many candidate authors, in which case, automated solutions can act as recommendation systems for forensics.

This work is organized as follows: Section 2 provides an analysis of modern works devoted to the authorship of fanfiction and classical texts, including those written in Russian. Section 3 describes the formal state of the problem, including both closed and open attribution. Section 4 describes the datasets used in the experiments, as well as its statistical characteristics. Section 5 describes the open and closed set methods with the support of the authors' previous works aimed at the determination of the authorship of Russian-language texts. Section 6 presents the results. Section 7 contains the conclusions and a discussion of the results and disadvantages of the considered methods, as well as plans for further research.

The scientific novelty of this research lies in the development of open set methods for Russian-language fanfiction texts based on the combination of fastText and SVM, as well as the complication of the close attribution of the author of a classical literary text by testing the method with confusing samples including artificially generated texts. Closed set attribution was used as an additional step for solving the problem of determining the author using an open set since the latter takes into account all the factors from a closed set and introduces additional complexity.

2. Related Works

2.1. Our Previous Research

In our previous papers [1,5] we found that, regarding classical machine learning (ML) methods, the best ML choice was the support vector machine (SVM) trained on a carefully selected feature space using a genetic algorithm (GA). The GA helped in feature reduction from 1168 to 400. Regarding neural networks (NNs), good results for the classical literary texts were found using fastText. In some cases, the results were inferior in accuracy compared to deep NNs (convolutional NNs (CNNs); Long Term Memory (LSTM) and their hybrids, bidirectional LSTM (BiLSTM); and two modifications of Bidirectional Encoder Representations from Transformers: RuBERT and MultiBERT) by 5%, but these methods had better training time for the maximum number of classes mentioned in all deep NN models. However, the issue of feature selection was not solved because, although the GA managed to improve the classification accuracy, we did not compare it with other feature selection methods (filter, wrapper, embedded) or their combinations. In this study, we pay attention to this issue.

For closed set attribution, the previously considered classical literary texts are used to solve the authorship attribution task in a man–machine case, where machine texts were generated using RuGPT-2 [6] and RuGPT-3 [7] based on texts written by ten authors with the highest separating ability, as identified in the results of past work [5].

2.2. Related Works Aimed at Authorship Attribution

Using fanfiction text in attribution tasks is not new for the global community. Every year at the PAN conference, fanfics in English, Spanish, German, and other languages are offered to participants as datasets. This choice is justified by the fact that, unlike classical literary texts, fanfiction data are substantially updated each year as new works by fiction writers are published daily in large volumes.

At the last PAN conference [8], the focus was on cross-fandom attribution. Based on this, all texts of unknown authorship were fanfiction texts of the same fandom (i.e., the thematic category), while documents with known authorship were fanfiction texts of several different fandoms. Importantly, the participants should have considered the task as an open attribution, where the true author of the text in the target thematic category may not be present in the list of candidate authors. Similar competitions were held until 2019, when the task was formulated as authorship verification. In this case, the participants had to determine whether a pair of texts were written by the same author.

The Authorship Verification 2022 competition used Aston 100 Idiolects Corpus [9] in English as a dataset. The dataset included essays, e-mail correspondences, text messages, and fragments of business letters. In total, 22,922 texts were written by 56 authors. The organizers noted the solutions of the najafi22 [10], galicia22, and jinli22 teams, which used the naive Bayesian classifier (NB). The winning team, najafi22, upgraded the NB by combining it with a distance-based character n-gram model, which achieved over 60% accuracy for the cross-genre case.

The main feature of PAN-2021 [11] was a dataset consisting of text sample pairs from two different fanfictions. Information about fandom (i.e., the thematic category) data for each text in the pair was also provided. The larger and smaller datasets differed in the number of classes and samples. The best result for the larger dataset was obtained by the team led by Benedikt Boenninghoff [12], with an accuracy of 95.45%. For the smaller set, the Weerasinghe team [13] achieved an accuracy of 92.8%, which was the best result in the competition.

For research aimed at Russian-language text analysis, fanfiction remains an understudied area since authors currently favor classic texts, scientific works, and social media comments. However, the creation of linguistic studies about fanfiction's role in the hierarchy of modern prose [14] and other studies of other fanfiction phenomena [15] indicates the importance of this text type.

There are many studies aimed at establishing authorship of natural text for forensic purposes [16]. Most of such publications applied different features of writing style and used SVM and LR [17]. Text features, including lexical, syntactic, structural, and specific to the genre and subject of the document, were applied both separately and as a single vector.

According to one of the recent extensive analytical works devoted to the approaches of determining the author of the text and publicly available datasets [18], several studies that apply semantic methods stand out in addition to those using the traditional syntactic and frequency groups of features. The authors assume that the semantic analysis of the text allows for identifying deeper features of the writing style, as they are directly related to the writing style and are not controlled by the writer when creating the text.

The work [19] is devoted to determining the authorship of translated and native texts. The authors used not only lexical and syntactic features but also investigated the influence of semantic frames on accuracy. The dataset of native texts consisted of 71 articles by

English-speaking authors in *The Federalist Papers*. The dataset of translated texts included works of 19th-century Russian classical literature translated into English (30 original texts, 4 authors, and 12 translators). For each text, 400 features were selected (frequencies of words and symbols *n*-grams, semantic frames of the most frequent words in the text, and frequencies of frames). The classifiers chosen were SVM and NB. SVM accuracy for translated texts ranged from 26.5% to 57.6%, depending on the author. Experiments were conducted on a set of native texts using different values of *n* when calculating *n*-gram frequencies. The most effective was the choice of *n* = 5 with lexical features and the naive Bayesian classifier. In the case of five authors, the accuracy was 98.6%. According to the results, semantic features worsen the classification accuracy of the native texts, while for the translated texts, on the contrary, semantic features increase the classification accuracy. Using only semantic features performs worse for all cases compared to the standard feature sets; the best result was shown using NB trained on lexical features.

The authors of [20] note that the statistical features used in many works ignore syntactic (e.g., dependencies of text levels relative to each other) or semantic information. In recent years, some researchers used syntactic trees or latent-semantic terms using NNs. However, only a few works consider them together. This paper proposes a Multi-Channel Self-Attention Network (MCSAN) that includes both inter-channel and inter-position interactions to extract *n*-grams of characters, words, parts of speech, phrase structures, dependency relations, and topics from multiple dimensions (style, content, syntactic and semantic features) to distinguish different authors. Experiments were conducted using well-known datasets for text author detection: CCAT10 (10 authors, 100 texts per author), CCAT50 (50 authors, 100 texts per author), and IMDb62 (62 authors, 1000 texts per author). To compare the proposed approach with other methods using the same datasets, CNN and BiLSTM were used both alone and in combination with the MCSAN. Experimental results show that the extracted features are effective–their application improves CCAT10 and CCAT50 results by 2.1% and 3.2%, respectively, compared to using BiLSTM and CNN without the proposed MCSAN.

In [21], authorship determination is used as a module of the recommendation system for book selection. According to the authors, one of the factors influencing reading preferences is the writing style. The article proposes a system that recommends books after studying the style of their authors. The system has two components: authorship determination and book recommendation. A CNN was used to determine the authorship. Vector representations of words and bigram character sequences were used as features. The input data were the text of the book. When an accuracy over 80% was achieved using the test sample, the authors extracted features from the hidden layer and used them for another task (book recommendation). Users of the system read the books and rated them on a 10-point scale. A list of books was generated based on the user's evaluations, and the feature vector was calculated to find the author with the closest values. When identifying the author, the architecture of the NN included an input layer, three hidden layers, fully connected, combining, and output layers. The latter outputs a vector with dimensions equal to the number of authors. Each element of the vector represents the probability that the input book belongs to a particular author. The values of all elements of one vector in total are equal to 1. The Litrec dataset [22], containing 1927 authors and 3710 literary works in English, was used as the dataset. Latent semantic analysis was used for the recommendation module. The final accuracy of the recommendations was counted as the proportion of matches of the system's advice and the user's actual choice. Depending on the volume of the text, the accuracy varied from 19.5% to 68.9%.

2.3. Related Works Aimed at Authorship Attribution of a Russian-Language Text

As for Russian-language texts, since the scientific community has not come to a consensus on the choice of the most effective feature space, different features are used. Under the notion of informative feature selection methods, we understand the extraction of dependencies at different levels of the text: in sentences, paragraphs, chapters, and whole texts within the author. Hierarchical dependencies are in the form of graphs, latent semantic analysis, and models of vector representation of words, based on distributive semantics. The latter is of the greatest interest since these methods aggregate all the others.

The work [23] is devoted to determining the author using psychological characteristics based on the text. Special attention was paid to aggressiveness. Identification of authors with increased personal emotionality in the text allows for improved analysis of social media platforms to identify hate speech, trolling, cyberbullying, and different types of manipulation, etc. The analysis of the emotional component of the text can be conducted at the level of the whole text or a single element. In this paper, such analysis was conducted at the level of the user, because in this case aggression could be a characteristic of the personality. Essays written by 487 individuals were used as a dataset. Features were expanded by upgrading a group of semantic features. Each author was assigned a category of physical aggression: angry, hostile, or neutral. Thus, the task of predicting the value of the target feature was reduced to a two-stage classification: in the first stage, the target label was the author label, and in the second stage, the target label was the aggressiveness label. For the same classifiers, the quality metric was the proportion of correct answers. The highest value of 59.1% was obtained using semantic and frequency features using SVM.

Many studies are aimed at scientific publications. The work [24] presents a system for resolving copyright disputes of Russian-language scientific articles. This solution improves the citation index calculation and article search results. When forming the dataset, the link.springer database was used as the initial repository of publications, and the data of the scientific electronic library, eLIBRARY, were used to obtain reliable information about the authors and their articles. The main purpose of the system is to act as an auxiliary tool for decision-making by experts. The system provides interactive visualization of the results to improve the quality of expert analysis. One of the problems was the ambiguity in the interpretation of transliterations of authors' names from Russian into English: the generation of all possible transliterations is not possible since real data may not obey the rules of transliteration. However, the greater the coverage of variants, the more data will be available during the search. In the current work, various transliterations of Russian letters used in state standards were studied. Based on the selected transliterations of individual letters, the generation of all possible transliterations of the author's name in Russian was implemented. For all transliteration options, the Springer database is accessed. Another database, eLIBRARY.ru, is also used according to the full name originally proposed by the expert in Russian. Then the data are compared using keys: author's full name, title, date of publication, co-authors, and authors' affiliations. In the absence of some parameters in one of the sources or their mismatch, the *k*-means clustering algorithm was used. Here, a pairwise comparison of articles and merging of groups took place if the similarity coefficient of articles exceeds a given threshold. To calculate the coefficient of similarity of natural language texts, the texts are represented as vectors in a multidimensional space. Then the measure of proximity between them is defined as the cosine distance. To improve the quality of comparison of texts in natural language, as well as to reduce the dimension of the vector representation of texts, they are preprocessed, which includes the removal of stop-words and stemming. The similarity was calculated using the word2vec tool. A module was added to the system for clustering articles that were not recognized at the stage of comparison with publications from eLIBRARY, which made it possible to improve the result of identifying the authorship of articles up to 92%. The approaches used in this system are applicable for disambiguating the authorship of publications from various bibliographic databases.

Forensic experts note the need to set the parameters for establishing the author's style for the attribution of Internet communication products [25]. In correspondence, users actively use non-verbal communication (stickers, emoticons, etc.), explaining or introducing a completely different meaning of the text. Therefore, the semantics of the message depends on the graphic level, while the importance of spelling and punctuation features for identifying authorship is reduced. The reason is the simplification of interaction in the digital space–even authors with a high level of spelling and punctuation skills do not always follow the rules of spelling and punctuation. Criminologists also note the importance of such features as deliberate errors. The main point stated in the work of T. P. Sokolova aimed at the modification of the key features of the text towards semantic ones. This will allow not only to distinguish between different authors by quantitative components (frequency distributions and aggregated statistical values) but also to determine the main intentions of the creator of the text, which will make it possible to optimize the production of forensic author's examinations.

There can be a large number of parameters describing the author's style: preferred words, local speech features, sentence length, use of turns of speech, and vocabulary. However, changing these parameters leads to a change in the frequency characteristics of the text, i.e., these changes affect the frequency of occurrence of the characters with which the text is written. The study of character frequencies separately allows for solving many problems related to text attribution, but the question arises of determining the set of the most informative frequencies, as well as eliminating redundant ones.

There are at least three reasons for the negative impact of a large number of non-informative features. First, with an increase in the number of features, the statistical reliability of the algorithm decreases. Second, the more features, the more learning objects are needed for reliable classification. Third, the more features, the longer the algorithm runs.

Feature selection is applied to find an optimal set of features and decrease the number of uninformative features and dimensionality while, at the same time, increasing the accuracy of the classification. The selected set of features reduces the computational cost, in particular for time and memory, but at the same time, the efficiency of the algorithm will increase.

Feature selection can be divided into three steps:

1. Generation of a subset of features;
2. Estimation of the generated subset;
3. If the generated subset is relevant or a break condition has occurred, then the process is terminated. Otherwise, return to step 1.

In the estimation step, the following methods are used:

1. Filter [26]. This method is based on identifying the generalizing properties of data and is not focused on specific algorithms and statistics. They are fast and less computationally expensive, so these methods are more often applied to datasets with a large number of features. However, there is a significant disadvantage of these methods: it considers each feature in isolation, not taking into account their mutual influence on each other. A simple example of a filter would be sorting the features in descending order of correlation with the target feature and selecting the *k*-features with the highest correlation.
2. Wrapper methods [27]. The peculiarity of this method is the search for all possible feature subsets and the assessment of their quality by "running" through the model. This method is included in the process of building a classifier, so it uses a measure of the effectiveness of the classifier to evaluate the selected set. Due to such work, the method significantly increases the complexity of the algorithm as well as computational costs, but at the same time, gives better results than the filter method. The essence of the wrapper method is quite simple: the classifier is run on different subsets of features of the original training set, and the best subset is selected.

At each stage, inaccuracies may appear, for example, due to the noisiness of the analyzed text, language features, etc., which will further lead to serious errors at higher levels. In addition, the larger the items of text, the larger the volume that may be required for the statistics to stabilize and for the analysis to be carried out. Character and word-level characteristics allow for the modeling of complex relationships within words and sentences, making them more promising than higher-level elements. However, to determine informative groups of features in a particular problem, it is necessary to research all potentially possible characteristics using promising methods.

2.4. Related Works Aimed at Author Profiling

Author profiling is understood as the establishment of the personality of the writer using the characteristics of the author. Among these characteristics, gender, age, education, profession, and personal qualities are distinguished. Gender and age are considered primary features because their concretization will allow the establishment of the remaining features (secondary) and reduce the set of candidates when determining the author. The feature set forms a unique linguistic personality. This concept is widely used today in forensic examinations [28], classification, psychological and marketing research, sociolinguistics [29], and diagnosing a person by text, as well as in other applied areas. As part of the world-famous PAN conference, participants were offered tasks both for profiling the author [30] and for determining gender and age separately [31].

The study [32] is devoted to gender profiling the author of a substandard text. By a substandard text, the authors considered a limited scope of use, for example, professional texts. The study was divided into three stages: feature selection; reducing the set of features in order to obtain only those suitable for determining a gender; and selection of features applicable to substandard texts. In the first stage, various groups of features were selected, including frequency distributions of characters, punctuation marks, parts of speech, numbers, average values of word lengths, and sentences. To determine the gender of the author, the final feature set included: the average values of word lengths, indexes of lexical diversity, the particle composition of the texts under study, and the type of sentences. Frequency distributions and lexical diversity indices are particularly applicable to substandard texts. Despite the many conclusions and the specific listing of the selected features, the authors did not provide information about the dataset, classification methods, or results of the study.

The article [33] is devoted to the methods used to analyze Russian-language fiction from the end of the XX century (1960–2000), namely, the determination of the gender of the authors of these materials. To solve the problem, it is necessary to determine the features of the text that the classification methods can use to distinguish between female and male texts. A list of frequently used punctuation marks and parts of speech was compiled for the analysis. To solve the problem of classification, a dataset of 87 works by 9 authors was collected. These authors were chosen based on differences in style and subject matter of the texts. Regarding the determination of the gender of the author, it was found that the frequency of using different parts of speech in the female and male texts of the given period were as follows: nouns, verbs, prepositions, pronouns, conjunctions, and adjectives, which reflects the specificity of the artistic style. The authors concluded that, in the literature of the 20th-century, women more actively used punctuation as an expressive means: the proportion of the use of exclamation marks, question marks, and commas by women writers significantly exceeded the value obtained through the analysis of men's texts. Four classifiers were used in the experiment to determine the gender of the author of the text: SVM naive Bayesian classifier, random forest, and k-nearest neighbor method. The classification results were as follows: SVM (75% male and 68% female), Bayesian classifier (65% male and 55% female), k-nearest neighbor method (75% male texts and 55% female), and random forest (70% male and 64% female).

Determining an author's gender using Twitter posts was discussed in [34]. The data included 4126 Twitter users and 6500 entries. When selecting features, the authors focused on semantic ones, and they extracted semantic categories from the tweets. The semantic categories were combined with tags of parts of speech, emoticons, and mentions of other users. The classification models used were AdaBoost, random forest, and recurrent NNs. For the analysis, 67% of the data were used for training and the rest were used for model testing. The resulting F-measure values ranged from 82% (random forest) to 94% (AdaBoost).

In the study [35], the sentiment of the text was taken into account when profiling the author. The study used an EmoGraph approach, which made it possible to identify six types of emotions. The dataset was based on a sample from the PAN-13 corpus of the author profiling task. The selected texts were written in Spanish and included labels for the author's age and gender. When extracting the features, the authors relied on research for English, as the field has not been sufficiently studied for Spanish. As a result, the feature vector included the frequencies of unique words in the text, capitalized words, average word length, number of capital letters and multiple vowel spellings, frequencies of punctuation marks, and emoticons. EmoGraph, which is a NN with five hidden layers, served as both a feature selection tool and a classifier. The authors of the study were able to achieve results comparable to the winners of the 2013 PAN competition (64% vs. 65% winners in the age determination task, respectively).

The authors of [36] considered age and gender not only as biological characteristics of the author but also as a way of expressing his social identity on the Internet. This approach is based on the fact that the data provided by users during registration may differ from the real state of affairs, which is an attempt to reflect self-positioning in society. Part of the research included the creation of an online game where Twitter users on a voluntary basis were invited to leave a link to their profile and specify their real data. Participants in the game tried to guess gender and age based on the tweet. A total of more than 3000 users–women and men aged under 20, 20 to 40, and over 40–participated. The average success rate on the first attempt was 32%, and the overall accuracy, counted as the proportion of correct answers, was 71%. It was found that more than 10% of Twitter users do not use pronouns that correspond to their biological sex, and older Twitter users are often perceived as younger. The automation of determining age and gender was carried out using the method of logistic regression. The signs in this case were the frequencies of the unigram characters. The accuracy of the system was 84% for age and 81% for gender.

2.5. Review of Publicly Available Datasets

For publicly available datasets, it may be noted that the organizers of the PAN contest provide the data used for the attribution cases [37]. In addition to the texts, the features, and the main statistical indicators were also attached. However, for this study, it is impossible to use the proposed PAN datasets since they do not contain Russian-language texts. Researchers can find English, Spanish, German, and Italian texts written in various genres including essays, emails, classics, and fanfiction.

In addition, for the English language, there are some specialized sets, for example, texts from authors of the Victorian era [38], where 50 authors and 1000 features for each text are presented. The disadvantage, in this case, is the inability to use raw data and independently extract features. A more modern dataset [39] includes texts by English-speaking bloggers and various related information about authors (gender, age, zodiac sign), which makes the dataset useful for author profiling. However, such a dataset still does not contain Russian texts.

Regarding the Russian language, the Kaggle platform provides two datasets [40,41]. Both datasets are aimed at identifying classical authors. Among the shortcomings, one can note the amount of data, where the first dataset contains only three classes, and the second dataset contains only two classes. Furthermore, at the moment, there are no datasets containing Russian-language fanfiction texts, nor are there any studies devoted to determining the author based on such data, which makes our work unique.

3. Problem Statement

When setting the mathematical formulation of the closed attribution problem, the process of establishing text authorship is given by three finite sets: $\boldsymbol{A} = \{a_1, \ldots, a_n\}$, $\boldsymbol{T} = \{t_1, \ldots, t_m\}$, $\boldsymbol{T}' = \{t'_1, \ldots, t'_s\}$ representing the authors, texts with known authorship, and anonymous texts, respectively. Each text, including anonymous ones, corresponds to a feature vector. The solution of the problem is reduced to the calculation of the objective function $f(t'_i) = [p(a_1), p(a_2), \ldots, p(a_n)]$, which takes an anonymous text as input and represents the probability distribution of belonging to each of the authors based on the featured text.

The final step in authorship identification for an anonymous text is to select the author with the maximum probability based on the objective function. Texts with certain authorship are used as a training sample. An equal number of non-anonymous texts are provided for each candidate author, while anonymous texts, in contrast, are unevenly distributed across all authors. All texts used for these experiments were written in Russian.

For closed set attribution, the previously considered classical literary texts are used to solve the authorship attribution task in a man–machine case, where machine texts were generated using RuGPT-2 [6] and RuGPT-3 [7] based on texts written by ten authors with the highest separating ability, as identified in the results of past work [5].

When solving open set attribution, it is proposed that some anonymous texts were written by none of the candidate authors. Thus, the size of the authors' training set is over that authors' training set $(|A_{test}| > |A_{train}|) \cup (A_{test} \cap A_{train} = A_{train})$. In this case, any new authors are defined as -1 class, and in the absence of sufficient similarity of the feature vector with the known samples of the authors, the new anonymous text is assigned to the negative class. Sufficient similarity in this context is determined depending on the method of open attribution: it can be a threshold value of probability, a measure of similarity between vectors calculated based on the Euclidean distance, or others.

In addition, special attention is paid to the data specifics. There are thematic texts with special keywords, such as names of characters, locations, special phrases, etc., that are repeated in all texts. In the case of open attribution, the experiments were devoted to the thematic of fanfics, that is, we divided the task into two cases, simple and complex, including:

- Cross-topic, in which we use authors who write fanfiction for various thematic groups;
- In-group classification, including classifications within each thematic group.

Simplicity and complexity depend on the vocabulary of the thematic group. Let there be a set of thematic groups $\boldsymbol{G} = \{g_1, \ldots, g_l\}$. That is, for the simple case: if author a_1 writes all his texts based on the thematic group g_1, and author a_2 based on the group g_2, then we assume that it will be easy to distinguish such authors because each author repeats in his texts words of his thematic category which do not occur in the other one. In the complex case, both a_1 and a_2 write based on only one thematic group g_1.

4. Dataset and Data Collection

4.1. Literary Dataset

As a dataset of classical literary texts, we used the previously described dataset [5], which includes 100 authors and 1100 literary texts. Detailed characteristics of the dataset are given in [5]. In the current study, in addition to the existing dataset, a set of texts generated using the RuGPT-2 [6] and RuGPT-3 [7] models is presented to solve the case of distinguishing authorship between a professional writer and a generative model. These models were chosen since the original texts are written in Russian. The generation took place for a sample of ten authors with the highest separating accuracy (Dostoevsky, Leskov, Tolstoy, Krylov, Bunin, Astafiev, Rasputin, Karamzin, Mamin-Sibiryak, Lermontov), whose works were classified with better separation ability in the previous work. Since we mean better separation ability, not better quality of works. The presence of artificial texts complicates the task and model's decisions.

The final dataset of classical literary texts for this study consists of 100 true authors and 20 generated fake authors. For each of the ten authors mentioned above, four generated texts were obtained. The statistics of the dataset are presented in Table 1.

Table 1. Information about the literary dataset.

Characters	Original Texts	Generated Texts (RuGPT-2)	Generated Texts (RuGPT-3)
Number of authors	100	10	10
Number of texts	1000	40	40
Average length of text, symbols	8,364,982	631,024	619,257
Dataset size, symbols	3,751,471,683	147,058,867	128,111,547
Dataset size, words	56,911,945	2,076,477	1,856,746
Dataset size, sentences	4,742,667	189,706	143,469
Average length of text, words	13.6	12.8	13.1

4.2. Fanfiction Dataset

Since this research is conducted using Russian-language texts, it is necessary to choose Russian data when forming the dataset and selecting a source. The online library ficbook [42] contains many subject areas and texts and has an active audience.

The criteria for selecting thematic categories (fandoms) were the popularity and availability of voluminous texts. By popularity, we mean the renewal of fanfics and the appearance of new authors and texts. It is important because, if a fandom is not popular, it is impossible to collect a sufficient amount of data. The second criterion chosen is the presence of texts with different lengths, because if all texts are written in the essay format and there are only a few such samples per author, there will not be enough data for training.

Based on provided criteria, five of the most popular fandoms from ficbook were selected for dataset collection including:

1. Harry Potter (HP);
2. Marvel Universe (MU);
3. Sherlock BBC;
4. Naruto;
5. Star Wars (SW).

In addition to the fandoms criteria, several requirements for the texts were defined when forming the dataset including:

1. Texts without a "translation" tag. It is important to collect the texts originally written in Russian because translation distorts the author's style;
2. Texts without co-authors. It is important to use texts written by a single author since, for co-authored works, it is not possible to establish what part of the text was written by each of the co-authors, which will introduce "noise" in calculating the features of authorial style and training the models.
3. Texts without two or more fandoms. Texts written within the same fandom were used; cross-fandom works were excluded because such texts make it easier to classify within a fandom.
4. Authors with a number of text samples. For the dataset collection, we used authors with at least five text samples in each fandom.

Based on the criteria, a dataset of fanfiction texts was obtained. The dataset contains texts by 686 authors and a total of 6569 texts. Most of the authors wrote texts devoted to Harry Potter since this fandom is one of the most popular not only for Russian but also for world fanfiction. This makes the final dataset unbalanced. However, taking into account the fact that in most works on determining the author of a fanfiction text, the classification is made within the category rather than cross-topic, this unbalanced design is not a disadvantage of the dataset. The minimum number of works per author throughout the dataset is 5 texts and the maximum is 23. General information about the dataset is presented in Table 2.

Table 2. Information about the fanfiction dataset.

Characters	All Data	HP	MU	Naruto	SW	Sherlock BBC
Number of authors	686	435	63	68	67	54
Number of texts	6569	4138	657	663	585	526
Dataset size, symbols	10,076,425	6,348,148	1,032,879	1,032,830	622,836	755,732
Dataset size, words	4,357,128	2,744,991	711,713	431,356	415,953	533,115
Dataset size, sentences	2,091,662	1,317,747	237,249	188,751	157,949	158,966
Average length of text, symbols	153,473	103,425	156,742	147,934	140,995	155,461
Average length of sentence, words	25.6	31.2	22.1	25.1	25.3	28.5
Maximum number of texts per author	34	21	19	22	23	22
Minimal number of texts per author	5	5	5	5	5	5

5. Methods

The specifics of open attribution are expressed by the appearance of new authors in the test set without including them in the training one; thus, it is not reasonable to apply methods suitable for closed attribution in its original form. Therefore, each group of problems (open and closed set of authors) requires an appropriate method. Section 5.1 describes closed attribution methods, including block diagrams and the presentation of methods as pseudocode. Section 5.2 describes the open attribution methods used in this study.

The choice of methods for closed attribution is based on the results obtained in our past work [5]. A comparison of classical ML methods (SVM, logistic regression (LR), decision trees (DT), random forest (RF), k-nearest neighbors (KNN)) and deep NNs (CNN, LSTM and their hybrids, BiLSTM, BERT, and RuBERT) with fastText, determined that it was possible to establish the effectiveness of the fastText model in the most difficult classification case including 50 authors (the maximum number of classes). Among the classical ML methods, SVM was identified in combination with feature selection by the genetic algorithm (GA).

5.1. Closed Set Attribution

According to the results of our previous study, SVM was particularly effective with a GA feature selection. The initial set of features was 1168 elements, but the GA reduced the feature space dimensionality and allowed a gain in accuracy. Reducing the feature space leads to an improvement in the classifier's training time. When selecting features, it is fundamentally important not just to select only a certain number of features but to keep only the informative ones. In the last paper, we did not consider other selection methods in addition to the GA, which was the task of the current study.

Formally, the original model uses N = 1168 text features, in this way each i-th feature corresponds to Xi-th element from a set of input features, and the full feature space is set as $\boldsymbol{F} = X_1 \times X_2 \times X_3 \times \ldots \times X_n$. Not all N features can help solve practical tasks, so the size of the finite informative features set, $|\boldsymbol{F}'| = M$ where $N > M$, and the subset, $R = N/M = \{r|r \in Nur \notin M\}$, consists of redundant features.

5.1.1. Genetic Algorithm

The genetic algorithm (GA) [5] was considered earlier and is now used a baseline in the current study. A detailed description of the method and the results obtained for 2, 5, 10, 20, and 50 authors were presented in the previous paper [5]. Figure 1 shows a block diagram of the genetic algorithm, and pseudocode is shown in Algorithm 1.

Algorithm 1. Genetic Algorithm

```
Set N:> Numbers of features
Set a:> Crossover ratio
Set Pop:> Feature set
Set Estimator:> SVM
procedure Crossover (N, a)
    number of crossover’s iters N = (a − N)/2
    for $j$ = 1 to N do
        select by random two gens $(A, B)$ from Pop;
        generate $(C, D)$ by one-point crossover $(A, B)$;
        save $(C, D)$ -> $P2$;
    return $P2$;
procedure Mutation (Population):
    for j in Population do:
    select j
    mutate each bit in vector-> new solution j
    update j in Population
    return Population
procedure Fitness (Estimator, chromosome):
    fitness_value = Estimator.test(chromosome)
    return fitness_value
begin
    Stopping_criteria = False
    while Stopping Criteria False:
        fitness_vector = []
        for each_chromosome in Chromosome:
            fintess_vector += Fitness(each_chromosome);
        Selection (fitness_vector);
        Crossover (N,crossover_ratio);
        if (length of crossover >0):
            Mutation (Pop);
            Pop = Mutated population
end
```

5.1.2. Regularization-Based Feature Selection

Regularization-based feature selection (RbFS) [43] is a heuristic-embedded method based on the evaluation of each feature measure. A linear model is fed to the model input, and a threshold value of weights is set for the selection. A feature is considered informative if the corresponding importance is not lower than the given threshold parameter or if it corresponds to the selected range of the value range, e.g., exceeds the average importance or median for all features. Limiting the power of the output feature set allows us to set a limit on their number.

The key to the RbFS method is the regularization parameter, which is why the embedded method is suitable for linear classifiers. The regularization process consists of adding a penalty to various model parameters to avoid overfitting, which leads to the model “fitting” to the solution of a specific problem, and, therefore, to its inapplicability for other cases. With the L1 regularization of a linear model, a penalty is applied to the coefficients that reduce each of the predictors. As a result of the experiments, it was found that in the case of SVM, the parameter C controls sparsity: with a smaller value of the parameter, fewer features are selected, and vice versa. A large C represents weak regularization, and a small C represents strong regularization. The scheme of the method is shown in Figure 2, and pseudocode is shown in Algorithm 2.

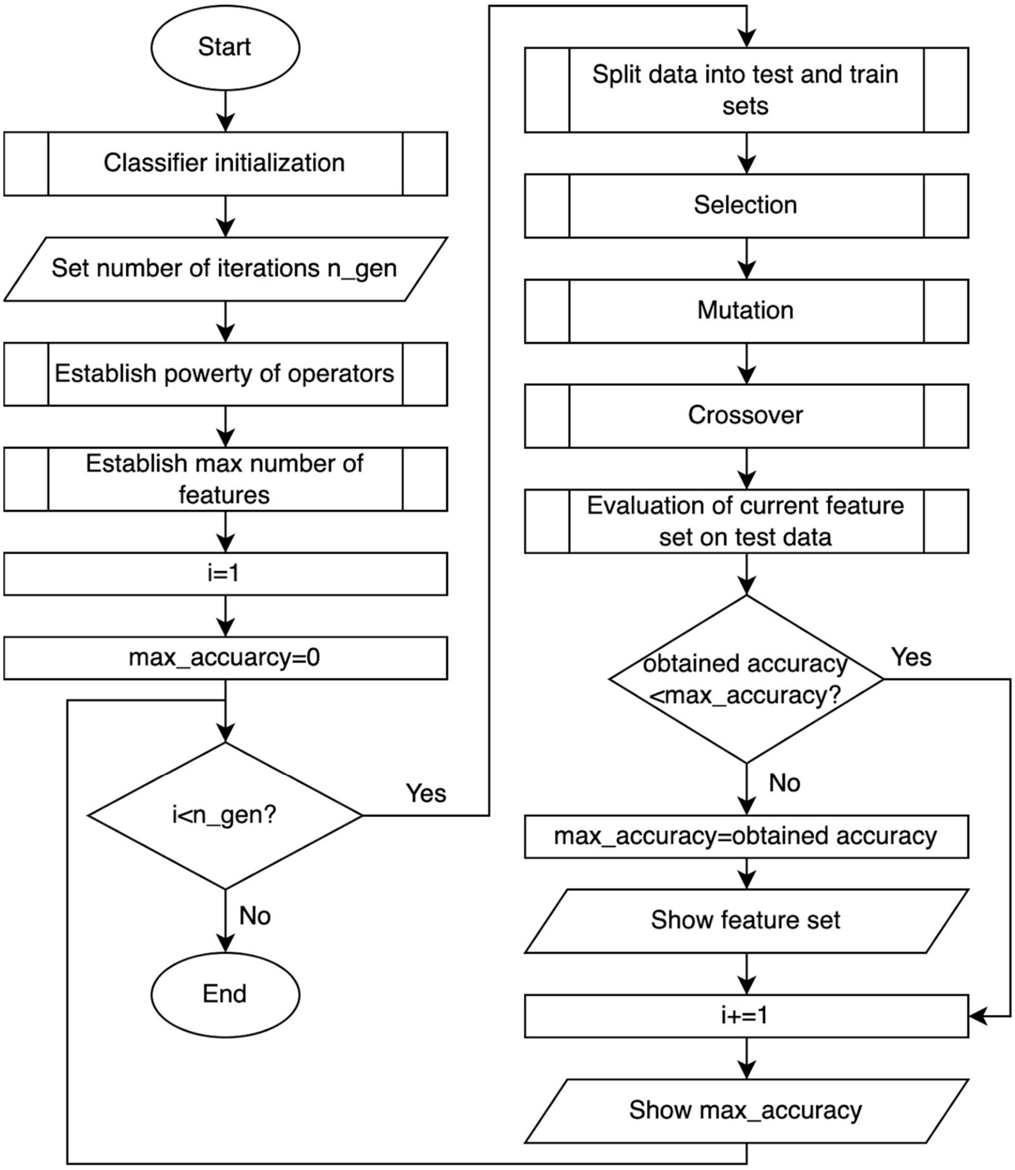
Start
Classifier initialization
Set number of iterations n_gen
Establish powerty of operators
Establish max number of features
i=1
max_accuarcy=0
i<n_gen?
Yes
No
End
Split data into test and train sets
Selection
Mutation
Crossover
Evaluation of current feature set on test data
obtained accuracy <max_accuracy?
Yes
No
max_accuracy=obtained accuracy
Show feature set
i+=1
Show max_accuracy

Figure 1. Genetic Algorithm block diagram.

Algorithm 2. RbFS Algorithm

```
Set N:> Numbers of features
Set Importance:> 0.95
Set Estimator:> SVM
Set Limit> 400
procedure Regularization (j, x, y, b):
    return $a_jx_jy_j + b$
begin
    feature_subset = []
    for i in N do:
        model = Estimator
        x, y, b = Estimator_coefficients:
        weights = Regularization (i, x, b):
        feature_subset += i, weights;
    select first Limit from festure_subset as best_subset;
end
```

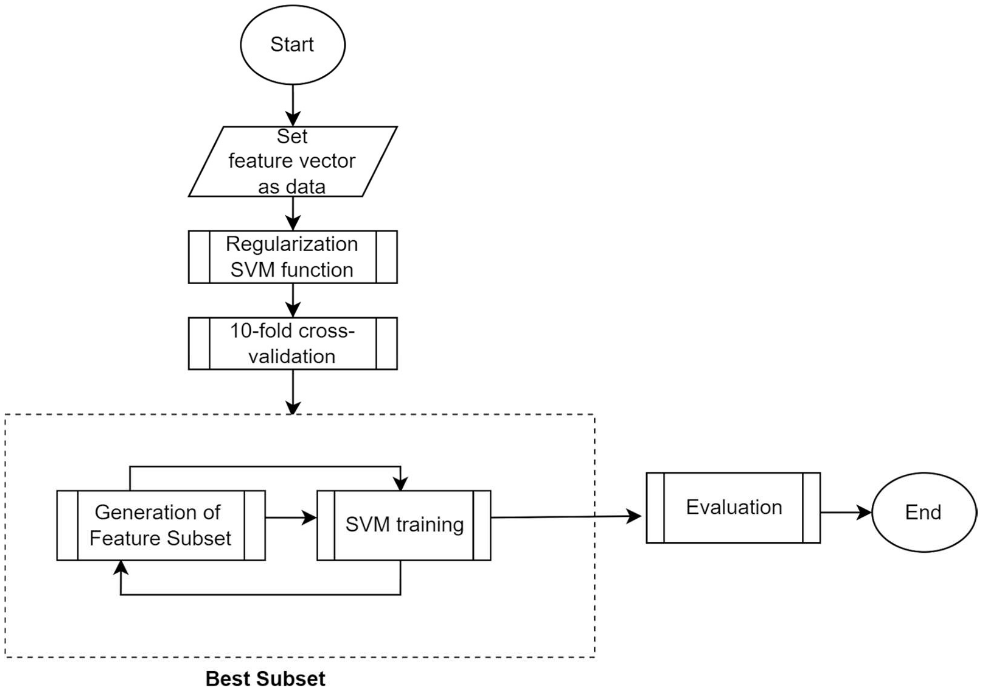
Start
Set feature vector as data
Regularization SVM function
10-fold cross-validation
Generation of Feature Subset
SVM training
Evaluation
End
Best Subset

Figure 2. RbFS block diagram.

5.1.3. Forward Feature Selection

Methodologically, forward selection (FFS) [44] is the simplest way to reduce the feature set dimensionality. FFS is an iterative algorithm with a stepwise regression application. In this method, the first feature is selected randomly, and the second and others are then added based on best performance (in the case of classification, best performance means maximum accuracy) then the process is repeated until a given number of features or accuracy threshold is reached. Each new iteration adds a feature to the model. The disadvantage of the method is the fact that the selection of the second and subsequent features essentially involves a complete oversampling, which negatively affects the selection time. Because of this, direct selection can cause overfitting of the model, which occurs when the selected features have a strong correlation but carry information that is close in meaning. The scheme of the method is shown in Figure 3, and pseudocode is shown in Algorithm 3.

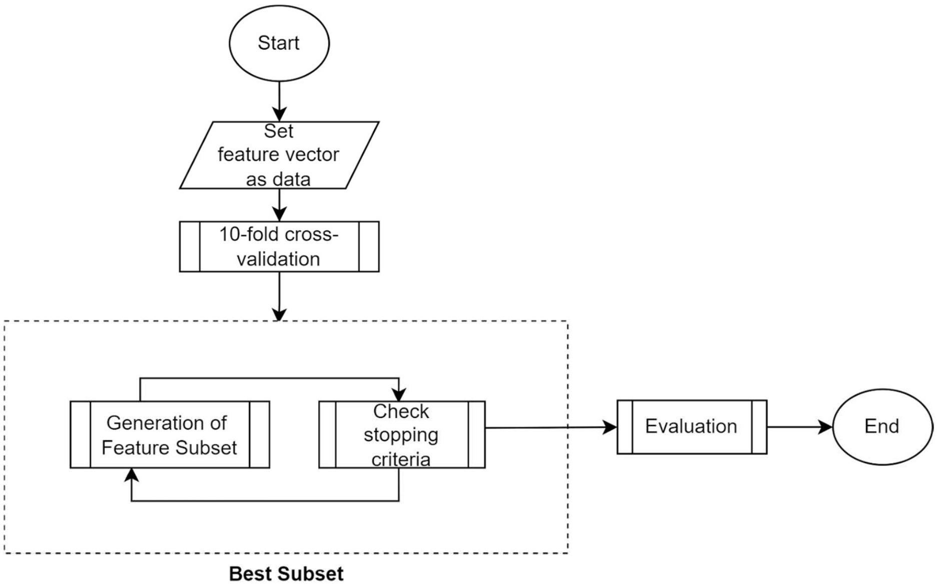

Figure 3. FFS block diagram.

Algorithm 3. FFS Algorithm

```
Set N:> Numbers of features
Set Estimator:> SVM
Set Feature_set:> []
begin
    for I in N do:
        accuracy = []
        best_accuracy = 0:
        model = Estimator:
            feature_set += feature[i]s;
            model.train(train_set_x[f_set], train_set_y)
            final_accuracy += accuracy (model, test_x[f_set], test_y), feature))
            for j in N[1::] do:
                current_accuracy = 0
                    if feature[j] not in feature_set do:
                        feature_set += feature[j]
                        model.train(train_set_x[f_set], train_set_y)
                        accuracy += accuracy(model(feature))
                        if accuracy > current_accuracy do:
                        final_accurcy += accuracy (model, feature))
                        feature_set.append(feature[j]);
end
```

5.1.4. Sequential Feature Selection

As the name of the method implies, sequential selection (SFS) is the opposite of FFS. In the beginning, the algorithm works using all available features, which are then excluded step-by-step. At each iteration, a feature that is of no value to the model is excluded. The exclusion of features can be based on the selected metric (accuracy, acceptable error rate, loss, or the *p*-value of the initial model). There is also uncertainty in this method when removing strongly correlated variables.

The algorithm is similar to FFS with the exception that during generation, the power of a subset of features decreases rather than increases.

5.1.5. Shapley Additive exPlanations

Shapley Additive exPlanations (SHAP) [45] is not designed for feature selection, but it is a good tool for interpreting the contribution of each feature to model predictions. As a result, SHAP allows us to estimate the importance of features. In the selection task, SHAP was used to select the most important features for accurate SVM training.

The method is based on Shapley values from coalitional game theory. The features generated for each text act as players in a coalition. Shapley values are used for the proportional distribution of accuracy between features. SHAP is a modification of the Shapley values since the explanation of the Shapley value is presented as an additive method of attribution, a linear model. This is formally shown in the following form (1):

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j {z'}_j \tag{1}$$

where g is a linear model; $z' = \{0,1\}^M$ is a coalition vector, where 1 corresponds to presence of the current feature, and 0 means absence; M is the maximum coalition size; $\phi_j \in R$ is the Shapley values set as an attributive function for the j-th feature; and ϕ_0 is the Shapley value without the j-th feature.

The condition of using a linear model is due to the possibility of ϕ calculation. For the text fragment x, authorship is determined using the model and the coalitional vector X'

consisting of N = 1068 features, i.e., all the values of the features are "present". The formula is simplified to the following (2):

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{2}$$

That is, we consider the average efficiency gain from adding the j-th feature to the coalition (a group of informative features); in the terminology of game theory, the result is the model accuracy obtained with this combination on a particular example. The fundamental difference of SHAP is that it works with a trained model and determines the importance of features based on the evaluation of test samples, while other methods determine the importance of a feature taking into account model training. The scheme of the method is shown in Figure 4, and pseudocode is shown in Algorithm 4.

Algorithm 4. SHAPc Algorithm

```
Set N:> Numbers of features
Set Players:> Limit number of features
Set Max_coalition_value:> 0
Set Estimator:> trained SVM
Set M:> maximum coalition size
Set start_accuracy:> 0
Set feature_set:> []
Set shap_values:> 0
procedure Win(feature, feature_set):
    return expected value of feature_set and feature
procedure shap_value(i, model):
    f_set = Coalition_vector
    importance = model.test(test_set_x[f_set],test_set_y)
    f_set += feature[i]
    feature_importance_ = model.test(test_x[f_set], test_y)
    return Win(importance, feature_importance)
begin
    for i in N! do:
        M = 1
        model = Estimator
        if M <= Players do begin:
            for j in (N-2)! do:
                M += 1;
                shap_values += shap_value(j, model)
        if shap_values >= Max_coalition_value do:
            Max_coalition_value = shap_values
    informative_features = ArgMax(max_coalition_value
end
```

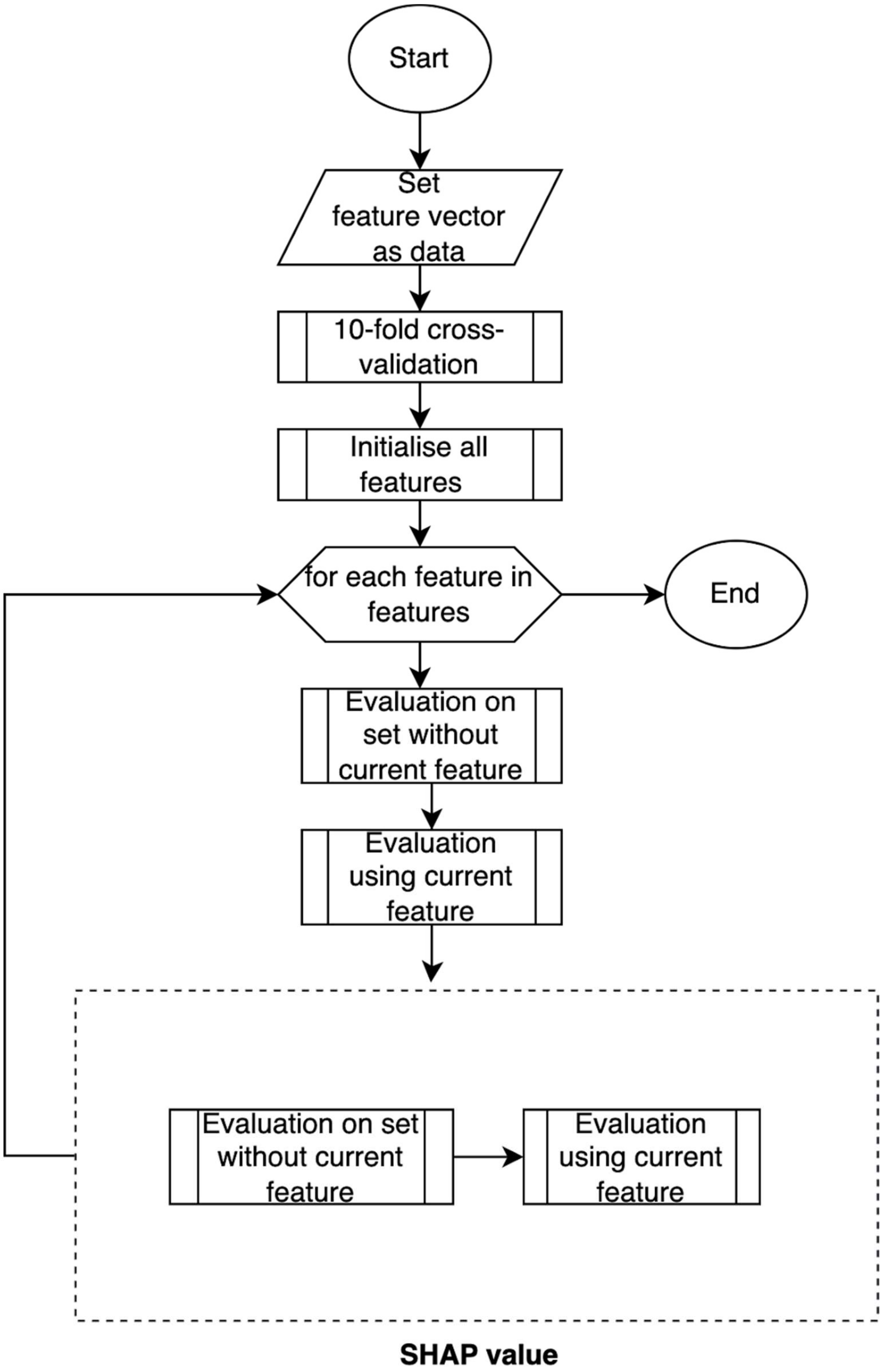
Start
Set feature vector as data
10-fold cross-validation
Initialise all features
for each feature in features
End
Evaluation on set without current feature
Evaluation using current feature
Evaluation on set without current feature
Evaluation using current feature
SHAP value

Figure 4. FFS block diagram.

5.2. Open Set Attribution

Nowadays, there are no classical, generally accepted open attribution methods. When solving these problems, the researcher must independently develop the necessary algorithms based on the specifics of the data and previous experience.

Researchers take into account their experience in solving closed set attribution and the results obtained when choosing open set attribution methods. In this study, a combination of One-Class SVM with feature selection + fastText was chosen as a baseline. Feature selection was provided using the method with proven effectiveness when solving classical tasks: closed set attribution. Thus, we do not conduct feature selection in open set methods and only use informative feature sets. This fact reduces the complexity of the experiment. The IDEF0 diagram of our methodology for identification of the text's author in open set attribution is shown in Figure 5. A detailed description of the method is provided below.

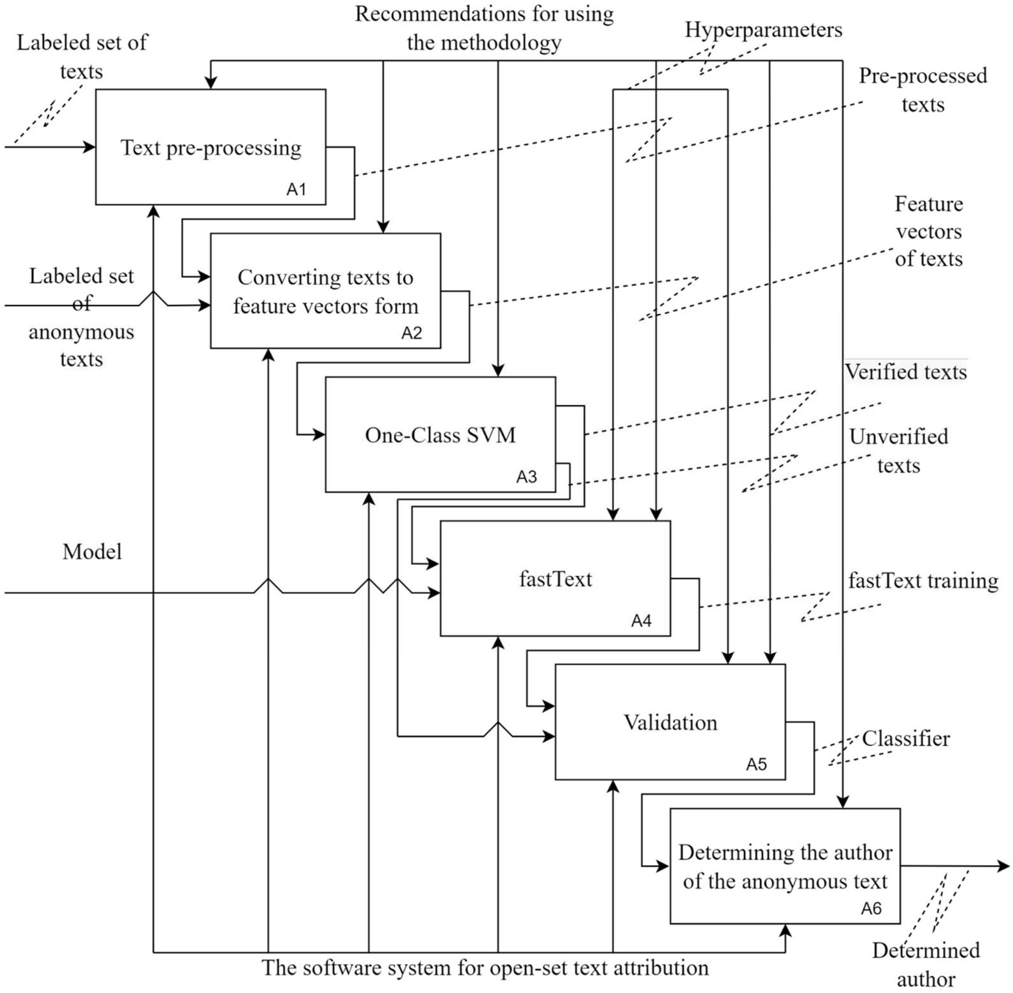

Figure 5. Open set IDEF0 diagram.

5.2.1. Method Based on One-Class SVM and fastText

This method involves two steps. In the first step, the texts, represented as a feature vector, are fed to the input of One-Class SVM. This method was chosen due to One-Class SVM's ability to detect anomalies, where anomalies are texts belonging to the negative class. Detection of anomalies is possible when the data in the training set follow a normal distribution and the test set contains anomalies, which are negative class samples. Based on normal and extra observations, One-Class SVM builds a non-linear space, where the anomalous data are cut off using a boundary. Thus, the goal of the first step is to determine whether the input data belong to one of the known authors whose texts were used in both training and testing (positive classes) or to an unknown class (authors whose texts are used only in testing).

In the second step of the method, the data that were classified as positive, i.e., as belonging to the set of authors from the training set and not being an anomaly, will be passed to the second stage including classification using fastText, which is already in the form of natural-language text. Using fastText, we obtain the assigned author label for samples of positive classes. At the evaluation stage of the method's performance, both negative and positive classes are included in the accuracy calculation.

5.2.2. Method Based on Thresholding Using Softmax Probabilities

The Softmax activation function is applied to fastText, and when solving the classification problem, the output is a vector, $\boldsymbol{P} = [p_1, p, \ldots, p_n]$, where n is the number of classes in the training sample, consisting of the probability distribution of the list of potential authors, while the sum of all probabilities is equal to: $\sum_{i}^{n}(p_i) = 1$. Based on the training samples of each class, the minimum threshold value of the probability of belonging to the true class, $p_{\min}$, is revealed. The smallest predicted probability, $p_{\min}$, for class X will serve as a threshold value, if the probability of a test sample belonging to class $X < p_{\min}$, it will be classified as a negative class. If the threshold value of $p_{\min}$ is exceeded, the text will be determined as belonging to the positive class X.

5.2.3. Statistical Methods

Method Based on Thresholding Using the Euclidean Distance

The Euclidean distance method is also based on a threshold value. A vector of 400 informative features is calculated for each text. Feature selection is completed using the method chosen as the best according to the results of closed attribution. Then, for each author (positive class), based on the training sample, the centroid is calculated as the arithmetic mean for each feature from the vector. As a result, for each of the known authors, a centroid vector of the i-th class is formed (3):

$$C_i = \bigcup_{j}^{400} \left[\frac{\sum_{j}^{n}(j)}{n} \right] \tag{3}$$

where n is a number of test samples, and j is the index of the feature vector element $\in [1;400]$.

The Euclidean distance is used to predict the class of the new text. That is, the i-th test sample calculates its own feature vector, $\mathrm{T_i}$, and then determines the distance of this vector to each centroid according to the Euclidean distance (4):

$$dist(T_i, C_j) = \sqrt{\left(x_{T_i} - x_{C_j}\right)^2 + \left(y_{T_i} - y_{C_j}\right)^2 + \ldots + \left(z_{T_i} - z_{C_j}\right)^2}, \tag{4}$$

where $\mathrm{T_i}$ is the feature vector of the anonymous text, $\mathrm{C_j}$ is the centroid of any positive class, and j corresponds to positive classes set size.

The threshold distance is defined for each positive class as the maximum deviation of one of the training samples from the centroid. If the Euclidean distance of the current test pattern exceeds this distance for all classes, it will be classified as an example belonging to the set of negative classes. An illustration of the method principle is shown in Figure 6.

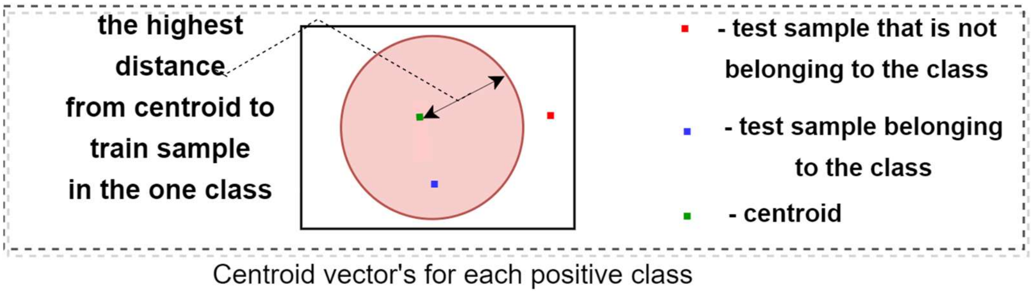
the highest distance from centroid to train sample in the one class
- test sample that is not belonging to the class
- test sample belonging to the class
- centroid
Centroid vector's for each positive class

Figure 6. Illustration of the thresholding using Euclidean distance method.

Method Based on Thresholding Using Cosine Similarity

The last open set attribution method is also based on a thresholding value considered as a cosine similarity. In this method, similar to the previous one, the feature vector is calculated for each text. After that, the centroid is calculated for each positive class. A cosine similarity measure is calculated for each training sample regarding class centroids (5):

$$similarity = \frac{T * C}{\|T\|\|C\|} = \frac{\sum\limits_{i=1}^{n} T_i \times C_i}{\sqrt{\sum\limits_{i=1}^{n} (T_i)^2} \times \sqrt{\sum\limits_{i=1}^{n} (C_i)^2}} \quad (5)$$

The minimum value per class is recognized as the threshold value. Test samples that do not exceed the threshold value for any class are classified as the negative class.

6. Results

This section presents the results of the experiments using all previously described methods on two datasets. For fanfiction texts, the experiments used open attribution methods. For classical literary texts, the experiments considered the task of closed attribution.

In all cases, a cross-validation procedure was used. To assess the quality of classification, accuracy, calculated as the proportion of correctly classified texts to the total number of texts in the test sample, was used as a metric:

$$Accuracy = \frac{Number\ of\ correct\ classified\ samples}{Number\ of\ all\ samples} \quad (6)$$

The feature vector calculated for each text contains the frequency distributions of unigrams, bigrams, trigrams, high-frequency words from the frequency dictionary of the Russian language, and punctuation marks. The process of feature vector formation is described in detail in the previous paper [1].

6.1. Closed Set Attribution

The process used to conduct the closed set attribution experiment and to split the data was standard: the dataset was split into training and test samples in 80:20 proportion, where 80% corresponded for four texts, and 20% of the test included a single sample. Experiments were conducted for 2, 5, 10, 20, and 50 classes, and artificial (generated texts) for the case of comparing the original texts with the generated generative model of 1, 2, 4, and 10 classes, respectively. Thus, 20% to 50% of the confounding samples were introduced to complicate the task. The length of each text for the closed set attribution experiments was 15,000 characters.

All five described selection methods were applied in combination with SVM. For closed attribution, classification using fastText and SVM with selected features are considered. Five different methods including the GA, presented earlier, as a baseline were used to select informative features and to establish the best subset of informative features and the most efficient method of selection.

The SVM hyperparameters include:

1. Sequential optimization method;
2. Linear kernel;
3. Regularization parameter = 1;
4. The acceptable error rate is 0.00001;
5. Additional options: normalization and compression heuristics.

The fastText hyperparameters include:

1. Number of *n*-grams = 3;
2. Learning rate parameter = 0.6;
3. Dimension for text = 500;

4. Loss function defined as "ova" (Softmax loss for multi-label classification);
5. The maximum number of allocated memory segments was 2,000,000.

The first experiment compared the GA method with RbFS, FFS, SFS, and SHAP. SVM was used as a classifier in all cases since earlier results [5] confirmed high accuracy in comparison with other classical ML methods (LR, RF, DT, KNN, NB) and a gain in training time of tens of times in comparison with deep NN models (CNN, LSTM and their hybrids, BiLSTM). The experiment aims to improve the classification accuracy for 20 and 50 authors.

The maximum number of features set equaled 400 since the GA can work well on all of the presented sets of authors with 500 and more features. Therefore, 50, 100, 200, and 400 features were considered, as in the earlier experiment with the GA. The results for GA and RbFS are presented in Tables 3 and 4, for FFS in Table 5, for SFS in Table 6, and for SHAP in Table 7. Figure 7 shows the most informative features according to the SHAP method. Figures 8–11 show the five most informative features for the cases of five authors, reflecting the influence of the feature on the classification of authors according to the GA, FFS, SFS, RbFS methods, respectively.

Table 3. Feature selection methods (GA-baseline).

Method	**GA-Baseline**			
Number of Features **Number of Authors**	**50**	**100**	**250**	**400**
2	95.9 ± 3.1	98.3 ± 3.9	96.2 ± 1.6	98.6 ± 2.7
5	94.8 ± 2.0	97.4 ± 1.9	95.0 ± 3.5	97.5 ± 3.1
10	84.3 ± 3.2	85.9 ± 3.8	87.1 ± 3.4	88.0 ± 2.9
20	60.7 ± 2.6	63.2 ± 2.5	72.9 ± 2.8	73.7 ± 3.3
50	33.8 ± 2.3	38.1 ± 3.7	42.4 ± 4.2	44.4 ± 2.6
Average accuracy	73.9 ± 2.6	76.6 ± 3.2	78.3 ± 3.1	80.4 ± 2.9

Table 4. Feature selection methods (RbFS).

Method	**RbFS**			
Number of Features **Number of Authors**	**50**	**100**	**250**	**400**
2	96.1 ± 3.2	95.9 ± 4.5	96.6 ± 4.0	98.9 ± 2.3
5	88.7 ± 2.9	90.3 ± 3.3	96.3 ± 2.5	98.6 ± 3.6
10	75.4 ± 4.0	84.2 ± 3.6	89.2 ± 3.2	92.5 ± 4.1
20	64.6 ± 5.2	66.6 ± 3.7	74.1 ± 4.7	78.9 ± 4.1
50	35.1 ± 5.5	36.4 ± 5.0	43.2 ± 4.8	47.3 ± 6.3
Average accuracy	71.9 ± 4.8	74.7 ± 4.2	79.9 ± 3.7	83.2 ± 4.8

Table 5. Feature selection methods (FFS).

Method	**FFS**			
Number of Features **Number of Authors**	**50**	**100**	**250**	**400**
2	95.0 ± 4.9	95.0 ± 3.5	97.2 ± 3.6	96. ±3.1
5	83.1 ± 3.7	88.8 ± 2.2	94.0 ± 3.8	94.1 ± 2.9
10	76.7 ± 3.8	80.4 ± 5.4	85.4 ± 4.1	84.7 ± 3.8
20	65.2 ± 4.9	68.1 ± 5.0	69.2 ± 3.6	69.9 ± 4.2
50	37.2 ± 6.3	35.8 ± 4.3	40.2 ± 4.8	41.3 ± 5.5
Average accuracy	71.4 ± 4.8	73.6 ± 4.7	77.2 ± 3.9	77.3 ± 4.3

Table 6. Feature selection methods (SFS).

Method	FFS			
Number of Features	50	100	250	400
Number of Authors				
2	89.9 ± 5.7	92.4 ± 5.4	97.2 ± 3.3	98.6 ± 3.1
5	83.4 ± 3.8	88.4 ± 5.7	95.2 ± 1.9	97.2 ± 4.0
10	65.2 ± 3.4	69.3 ± 4.0	90.0 ± 2.9	92.7 ± 3.8
20	58.4 ± 3.3	62.3 ± 5.1	74.5 ± 3.8	71.1 ± 2.2
50	29.2 ± 4.6	35.9 ± 4.8	40.1 ± 3.9	44.9 ± 5.2
Average accuracy	65.2 ± 2.4	69.7 ± 4.7	79.4 ± 2.8	80.9 ± 4.4

Table 7. Feature selection methods (SHAP).

Method	FFS			
Number of Features	50	100	250	400
Number of Authors				
2	97.5 ± 2.3	98.2 ± 4.9	97.6 ± 3.5	93.3 ± 2.3
5	94.0 ± 3.4	90.0 ± 4.8	92.4 ± 2.1	89.7 ± 4.3
10	88.5 ± 3.6	84.2 ± 3.8	85.1 ± 4.6	83.1 ± 3.6
20	65.2 ± 3.8	72.9 ± 4.1	65.6 ± 4.7	62.9 ± 4.1
50	37.3 ± 5.0	42.1 ± 4.7	41.1 ± 3.6	39.1 ± 4.7
Average accuracy	76.5 ± 4.1	77.5 ± 4.2	76.4 ± 3.8	73.6 ± 4.0

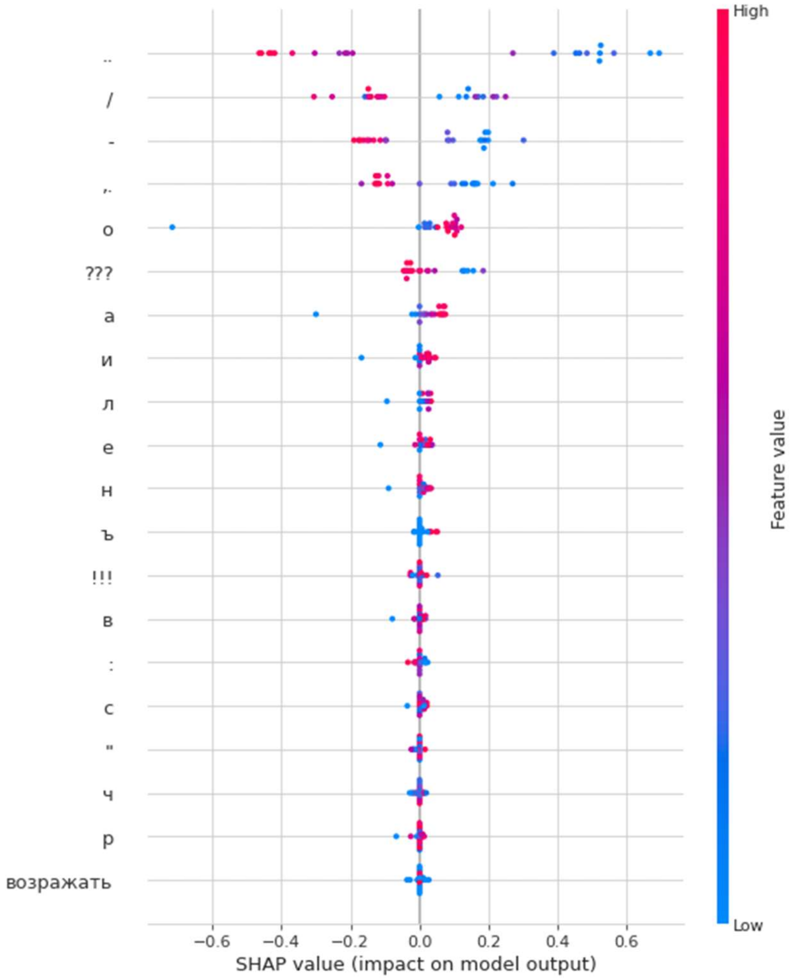

Figure 7. SHAP informative features.

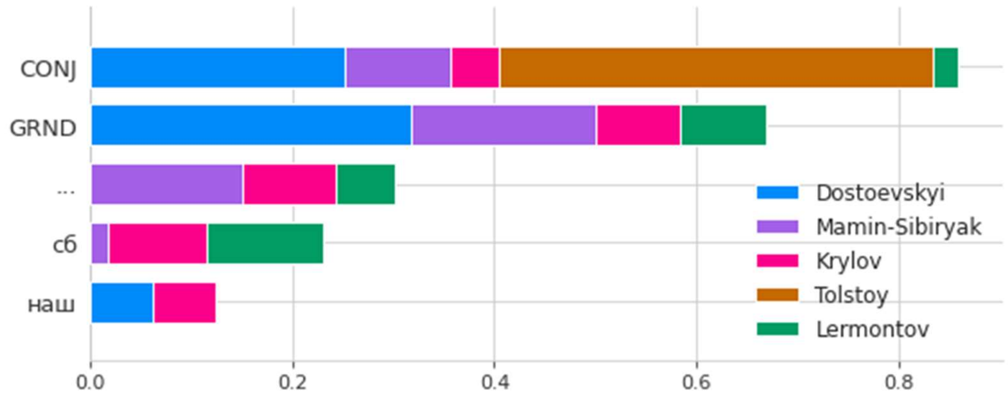

Figure 8. Top five informative features based on GA decision.

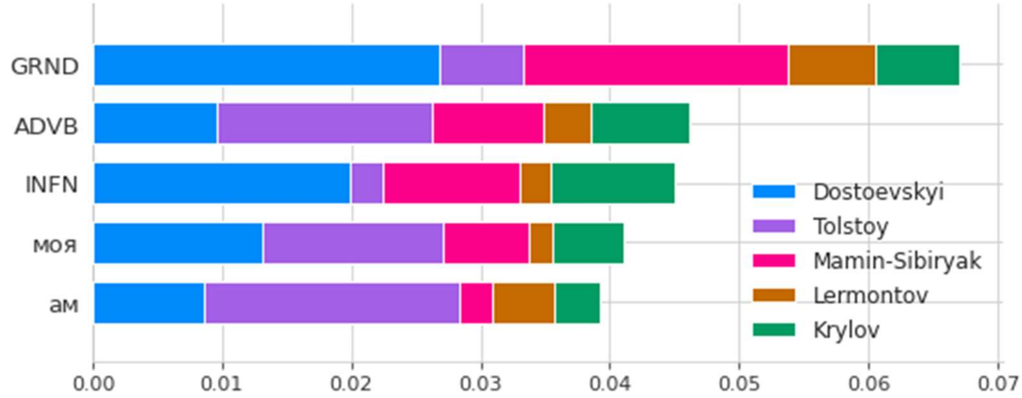

Figure 9. Top five informative features based on FFS decision.

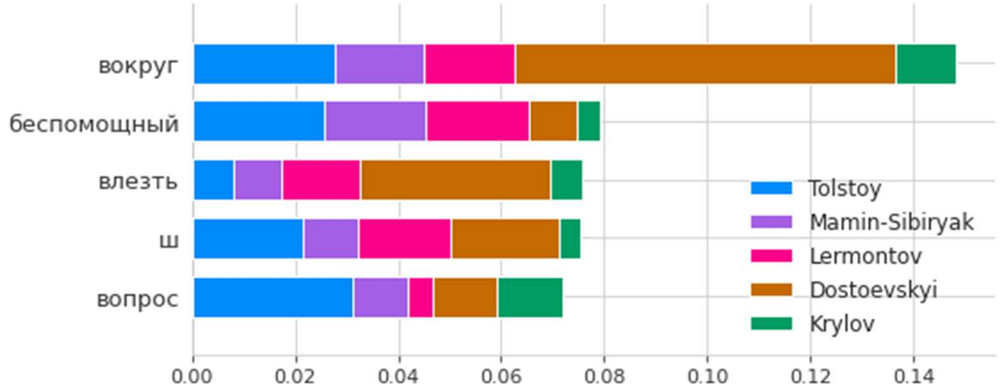

Figure 10. Top five informative features based on SFS decision.

Figure 6 demonstrates the feature importance in classification using SVM in combination with SHAP. Importance is regarded as an increase in accuracy using the given feature. The y-axis shows the specific names of the feature, and the x-axis shows the impact on model output for the given feature. The graphs shown in Figures 7–10 reflect the most informative features of the five authors' classification and show the impact on model output from the GA, FFS, SFS, and RbFS methods, respectively. The legend of each graph shows the names of the authors, which were set as classes in the experiments. The y-axis shows the features, and the x-shows the impact on model output, with the value of the impact divided for each class according to the color label of the author shown in the legend of the graph. It should be noted that RbFS identified punctuation features as the most informative and that such features were also found to be informative using the SHAP and GA methods. In contrast, punctuation features were not included in the top 5 informative features identified using the FFS and SFS methods. The smallest increase even for the most informative features is observed for the results of the FFS method (Figure 8), where the frequencies of parts of speech are among the most informative features. Part of speech frequencies was also found using GA, but with a significantly larger increase. For example, the impact of the feature GRND together with GA is 0.66 in total for five authors, while for FFS, the impact is only 0.067, which confirms the results shown in Tables 3 and 5, and the conclusions about the inefficiency of selection methods with a total selection.

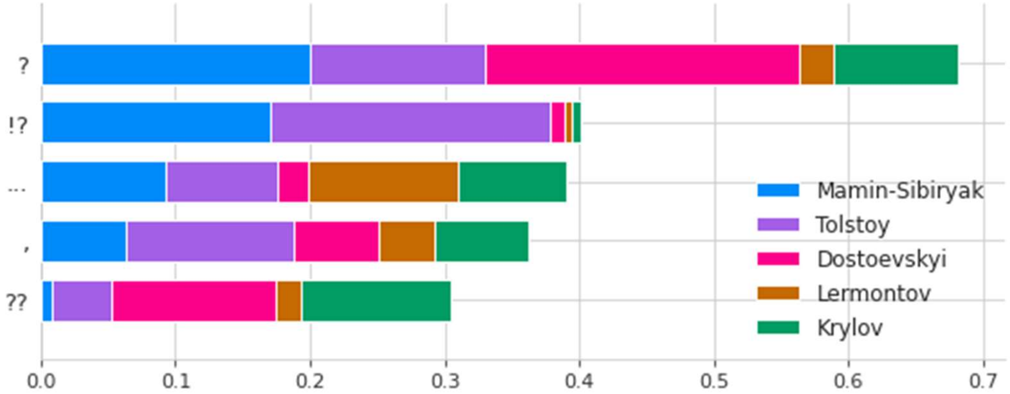

Figure 11. Top five informative features based on RbFS decision.

The results confirmed the inefficiency of the FFS and SFS methods because estimated accuracy is inferior to the GA’s results for the same number of authors (5, 10, 20, and 50). It should be noted that for the last two cases (20 and 50 authors), FFS and SFS accuracies are comparable with GA’s. However, this fact is not a reason to set FFS or SFS as the best feature selection method since the goal was to improve accuracy for difficult cases. As a reason for the low accuracy, we suggest simple algorithms based on an exhaustive search of these methods. The features identified as informative by the SHAP method demonstrate competitive results only with a small number of features (50, 100) and classes (2, 5, 10). The main advantage of the method is a demonstration of good results using a low feature space dimension. This is justified by the fact that features are selected based on their importance, and in the case of expanding the set of features, with each new feature, the importance of a new element is lower than that of the past. In addition, features can correlate with each other, but at the same time, individually, they do not provide additional benefits.

The method based on the regularization (RbFS) demonstrates the objective achievement and improvement of the accuracy for 20 and 50 authors to 5%. As a reason for the obtained result, we note the limitation of this method. RbFS works with linear models, so SVM was used only with the linear kernel, as well as the influence of the regularization parameter. Thus, we assume that such efficiency is due to the choice of the classifier, which can also serve as a drawback of the method as it uses a limited list of classification models (only linear), while the GA can work with other methods as well.

The limitation of statistical methods based on distances is a high dependence on the training data. When such approaches are used, texts created by the same author but based on a different topic or created using an attempt to imitate the writing style of another writer, are highly likely to be misclassified. A limitation of FFS and SFS is the randomness factor in the choice of the first feature based on which the further set is formed. If the first element of the informative features set is “unsuccessfully” selected, the subsequent components may be selected based on “fitting” to the specific test data used in performing the selection but be a little informative in the last iterations of the classifier’s work when using the selected set.

Table 8 shows the results of an experiment aimed at using SVM with RbFS for the classification of classical literary texts in three variations: the standard classification process and the complication of the task by introducing artificial texts generated by RuGPT-2 and RuGPT-3 based on the original dataset. For a generation, we used the texts of ten authors classified with the highest separating power (Dostoevsky, Leskov, Tolstoy, Krylov, Bunin, Astafiev, Rasputin, Karamzin, Mamin-Sibiryak, Lermontov). Table 9 shows similar results of the experiments for classification using fastText.

Table 8. Best FS method + generated texts.

	Dataset		
Number of Authors	**Original TEXTS**	**Original + Generated (RuGPT-2)**	**Original + Generated (RuGPT-3)**
2	98.9 ± 2.3	96.5 ± 4.2	95.7 ± 3.8
5	98.6 ± 3.6	94.4 ± 3.4	90.2 ± 4.1
10	92.5 ± 4.1	89.2 ± 3.6	86.7 ± 3.9
20	78.9 ± 4.1	71.8 ± 4.3	70.0 ± 2.6
50	47.3 ± 6.3	35.6 ± 5.2	35.1 ±3.8
Average accuracy	83.2 ± 4.1	77.5 ± 4.0	75.5 ± 3.6

Table 9. fastText.

	Dataset		
Number of Authors	**Original Texts**	**Original + Generated (RuGPT-2)**	**Original + Generated (RuGPT-3)**
2	98.2 ± 4.5	95.4 ± 5.2	95.0 ± 3.7
5	95.0 ± 3.7	90.3 ± 4.5	90.4 ± 5.0
10	92.2 ± 6.3	87.2 ± 4.4	85.3 ± 3.7
20	69.9 ± 4.3	63.2 ± 3.7	62.5 ± 4.4
50	56.8 ± 6.2	48.4 ± 5.1	40.9 ± 4.8
Average accuracy	84.2 ± 5.0	76.9 ± 4.5	73.4 ± 4.3

Using artificial-generated texts complicates the process of determining the text's author in both cases when using the classical methods of ML represented by SVM and NNs represented by fastText. Due to the careful formation of the feature space and the establishment of a feature selection method, SVM turned out to be more resistant to model obfuscation. The maximum accuracy loss for fastText was 16%, and for SVM it was 12%. The greatest loss of accuracy was recorded for 50 and 20 numbers of classes, while for 2 and 5 classes, despite the complication of the task, the losses did not exceed 3% of the result obtained without the use of generated samples. The average classification accuracy of the original samples was higher in the case of fastText, but the feature selection SVM outperformed fastText when testing the method using the injection of generated samples, which indicates the effectiveness of the fitted feature space.

6.2. Open Set Attribution

It is worth noting that the texts used in the practical part of the work are not written by professional writers and are classified within their thematic category, which indicates an increase in the complexity of experiments and closeness to real forensic tasks.

In our series of studies, we take into account the earlier work [1,5] and continue to use methods that demonstrate the best results. When solving the open attribution of the author of a text, the goal was to develop methods based on a single-class modification of SVM and fastText, a single use of fastText, and also to add statistical methods that are not based on ML, since the representation of the text as a numeric vector allows us to work with them.

The fastText hyperparameters are the same as described in Section 6.1. The One-Class SVM hyperparameters include:

1. Linear kernel;
2. Regularization parameter = 1;

3. Gamma = 'scale';
4. Tolerance for stopping criterion = 1×10^{-5}.

Training for open attribution was completed using the texts of N authors, and text samples from another M authors were introduced in the test stage, so the size of the set of authors for training is N, and the size of the set for testing is $N + M$. The division of the initial set N into training and test samples, as in the case closed attribution, was completed in a ratio of 80:20. The experiments were carried out for 2 + 1, 4 + 1, 9 + 1, 19 + 1 and 49 + 1 authors, where the second term indicates an additional negative class, including authors whose texts are added at the testing stage. The number of added anonymous texts was 30% of the test sample. Each text for open attribution experiments was 25,000 characters.

Two cases of experiments were also considered for open set attribution. The first, a cross-topic case, consisted of the classification of authors who published works for different thematic categories, and the second, a more complex case, consisted of in-group classification. For the first case, all possible pairs of thematic categories and all open attribution methods described above were considered including the One-Class SVM + fastText, the threshold value of the SoftMax function for fastText classification and two statistical methods including the Euclidean distance and cosine similarity measure (Tables 10–13). For the in-group classification (Tables 14–17), the same methods were used, but all authors wrote texts in only one thematic group, so the task was complicated by the use of similar special words, names, and other attributes of thematic groups.

In an attempt to imitate the original, the use of pre-fixed proper names and significant words should be noted as the complexity of fanfiction data. As a result, the texts become similar in their general content; for classification, we should find differences, analyze the semantics, and rely only on the writing style features of a particular author. The method is based on the combination of One-Class SVM and fastText, which makes it possible to take into account the generated feature space and complete semantic text analysis using neural networks. Due to this, the results obtained using this method are superior to those obtained using other methods. Statistical methods, including the Euclidean distance thresholds and cosine similarity, were the least effective even for a simple, cross-topic, classification case, and the difference in accuracy compared to the One-Class SVM-based method was up to 15%. In-group classification imposes additional complexity due to the imitation of the original data. Regarding the influence of thematic groups, it can be noted that the results vary by no more than 8% from group to group with the same method and number of classes, which indicates that the developed methods are independent of the choice of a particular thematic category.

Table 10. One-Class SVM + fastText result for cross-topic case.

Number of Authors	HP-Naruto	HP-Sherlock BBC	HP-MU	HP-SW	Naruto-Sherlock BBC	Naruto-SW	Naruto-MU	Sherlock BBC-SW	Sherlock BBC-MU	MU-SW
1 + 1	98.4 ± 3.2	98.9 ± 1.3	99.0 ± 1.3	98.4 ± 2.2	98.2 ± 0.9	98.3 ± 1.9	98.0 ± 2.0	98.6 ± 1.8	99.0 ± 1.5	99.1 ± 2.0
3 + 2	96.2 ± 2.6	95.3 ± 2.1	96.1 ± 2.7	95.6 ± 2.3	95.3 ± 3.4	96.5 ± 1.8	96.3 ± 3.0	96.2 ± 2.2	96.2 ± 2.5	95.8 ± 1.8
7 + 3	90.3 ± 2.4	91.3 ± 3.3	93.0 ± 2.8	92.4 ± 2.3	92.4 ± 1.9	91.8 ± 2.9	92.6 ± 2.7	90.5 ± 3.6	91.4 ± 2.8	91.7 ± 1.9
15 + 5	81.8 ± 3.9	80.3 ± 3.5	83.6 ± 2.6	82.0 ± 3.6	82.1 ± 3.7	82.2 ± 4.5	82.3 ± 4.4	81.4 ± 2.8	82.3 ± 4.1	82.3 ± 4.2
40 + 10	59.0 ± 5.7	55.5 ± 3.4	54.4 ± 3.6	57.3 ± 3.9	54.4 ± 3.8	58.5 ± 4.0	55.4 ± 4.4	56.0 ± 5.1	55.2 ± 4.0	56.3 ± 4.2
Av. accuracy	85.1 ± 3.6	84.3 ± 2.7	85.2 ± 2.6	85.1 ± 2.9	84.4 ± 2.7	85.5 ± 3.0	85.6 ± 3.3	84.5 ± 3.1	84.7 ± 2.9	85.0 ± 2.8

Table 11. Method based on SoftMax thresholding using fastText for cross-topic case.

Number of Authors	HP-Naruto	HP-Sherlock BBC	HP-MU	HP-SW	Naruto-Sherlock BBC	Naruto-SW	Naruto-MU	Sherlock BBC-SW	Sherlock BBC-MU	MU-SW
1 + 1	98.0 ± 2.3	98.5 ± 1.4	98.2 ± 0.9	99.2 ± 0.5	98.0 ± 1.1	99.0 ± 2.8	97.9 ± 1.6	98.9 ± 2.2	98.5 ± 2.4	98.8 ± 1.1
3 + 2	95.1 ± 2.5	95.2 ± 1.9	97.3 ± 1.8	96.5 ± 3.0	97.0 ± 2.7	96.4 ± 2.0	96.9 ± 2.1	97.3 ± 1.9	97.1 ± 2.2	96.7 ± 2.5
7 + 3	93.2 ± 3.3	91.8 ± 4.0	92.1 ± 2.9	91.0 ± 4.1	93.2 ± 3.7	92.0 ± 3.5	91.9 ± 3.9	93.6 ± 4.4	93.3 ± 3.3	92.8 ± 2.8
15 + 5	81.2 ± 4.5	78.4 ± 4.6	81.0 ± 3.7	82.2 ± 3.0	81.4 ± 3.8	79.9 ± 4.6	81.2 ± 4.7	78.1 ± 3.7	82.2 ± 5.1	82.1 ± 4.8
40 + 10	54.6 ± 5.3	50.1 ± 4.8	51.7 ± 4.7	55.9 ± 4.4	52.1 ± 4.7	57.4 ± 4.3	52.0 ± 4.9	55.6 ± 4.7	52.5 ± 4.6	53.6 ± 4.8
Av. accuracy	82.6 ± 3.6	82.8 ± 3.3	84.0 ± 2.8	83.6 ± 3.0	84.3 ± 3.2	84.9 ± 3.4	83.9 ± 3.3	85.1 ± 3.4	85.2 ± 3.5	84.8 ± 3.2

Table 12. Method based on Euclidean distance for cross-topic case.

Number of Authors	HP-Naruto	HP-Sherlock BBC	HP-MU	HP-SW	Naruto-Sherlock BBC	Naruto-SW	Naruto-MU	Sherlock BBC-SW	Sherlock BBC-MU	MU-SW
1 + 1	95.7 ± 1.5	96.9 ± 1.4	96.2 ± 2.0	94.3 ± 1.8	95.1 ± 1.7	97.0 ± 1.4	95.2 ± 1.9	97.3 ± 2.2	97.0 ± 1.9	98.2 ± 1.4
3 + 2	92.2 ± 3.5	93.4 ± 1.0	92.6 ± 2.2	91.6 ± 1.9	92.1 ± 2.7	93.2 ± 1.9	90.5 ± 2.2	92.8 ± 2.7	91.7 ± 3.5	92.6 ± 1.6
7 + 3	89.3 ± 2.7	86.2 ± 3.9	89.8 ± 3.0	86.2 ± 3.4	90.9 ± 3.4	89.8 ± 2.6	86.7 ± 2.8	90.7 ± 3.8	85.8 ± 2.9	88.4 ± 3.0
15 + 5	75.2 ± 4.4	80.2 ± 4.2	80.1 ± 3.8	77.1 ± 2.4	82.0 ± 2.4	80.2 ± 4.7	76.8 ± 3.8	80.6 ± 3.9	74.3 ± 4.3	80.3 ± 3.7
40 + 10	49.3 ± 5.1	46.2 ± 4.5	47.3 ± 3.5	50.2 ± 4.6	45.8 ± 4.3	45.2 ± 3.4	44.3 ± 3.8	50.3 ± 3.9	46.6 ± 4.3	41.2 ± 2.9
Av. accuracy	78.1 ± 3.4	80.2 ± 3.0	81.2 ± 2.9	79.8 ± 2.8	80.2 ± 2.9	81.0 ± 2.8	78.7 ± 2.9	82.3 ± 3.2	77.8 ± 3.4	80.1 ± 2.5

Table 13. Method based on cosine similarity thresholding for cross-topic case.

Number of Authors	HP-Naruto	HP-Sherlock BBC	HP-MU	HP-SW	Naruto-Sherlock BBC	Naruto-SW	Naruto-MU	Sherlock BBC-SW	Sherlock BBC-MU	MU-SW
1 + 1	96.2 ± 2.3	94.8 ± 1.4	96.1 ± 2.2	97.2 ± 1.7	96.1 ± 1.8	95.2 ± 0.8	97.1 ± 1.6	96.8 ± 0.9	95.9 ± 1.7	96.4 ± 2.3
3 + 2	93.3 ± 2.6	93.5 ± 3.1	94.2 ± 1.8	92.4 ± 1.9	93.5 ± 4.3	92.6 ± 2.7	93.6 ± 2.9	91.6 ± 3.4	90.6 ± 3.1	89.6 ± 3.2
7 + 3	89.2 ± 3.5	90.2 ± 1.6	83.9 ± 3.5	84.9 ± 3.2	91.0 ± 2.2	83.4 ± 3.0	82.7 ± 2.2	79.6 ± 2.4	81.1 ± 2.9	77.9 ± 2.1
15 + 5	80.3 ± 3.8	75.4 ± 2.7	72.7 ± 2.7	75.1 ± 2.4	77.4 ± 2.8	80.5 ± 4.2	75.6 ± 4.3	74.5 ± 2.4	71.4 ± 4.0	70.0 ± 3.3
40 + 10	44.9 ± 3.3	46.7 ± 2.6	45.5 ± 4.2	47.4 ± 3.3	45.5 ± 4.7	48.8 ± 5.1	44.5 ± 3.8	47.8 ± 4.7	46.7 ± 3.8	48.1 ± 5.0
Av. accuracy	80.7 ± 3.1	77.1 ± 2.3	75.4 ± 2.8	79.4 ± 2.5	80.7 ± 3.1	80.9 ± 3.2	78.7 ± 2.8	78.2 ± 2.8	77.1 ± 3.1	76.4 ± 3.5

Table 14. One-Class SVM + fastText for in-group classification.

	Dataset				
Number of Authors	**HP**	**Naruto**	**Sherlock BBC**	**MU**	**SW**
2 + 1	95.1 ± 2.7	97.3 ± 1.5	98.2 ± 1.4	96.0 ± 3.3	96.8 ± 2.5
4 + 1	88.0 ± 2.9	93.2 ± 4.4	92.5 ± 3.6	94.1 ± 4.0	92.1 ± 2.7
9 + 1	80.1 ± 3.4	82.5 ± 4.1	83.2 ± 3.7	85.3 ± 4.2	84.2 ± 4.0
19 + 1	72.2 ± 5.1	69.5 ± 2.4	67.6 ± 3.4	70.7 ± 2.9	71.4 ± 4.6
49 + 1	42.5 ± 2.6	47.3 ± 3.9	51.1 ± 5.2	45.8 ± 3.2	44.4 ± 4.8
Average accuracy	75.5 ± 3.4	77.9 ± 3.3	78.6 ± 3.5	78.3 ± 3.4	77.1 ± 3.7

Table 15. Method based on SoftMax thresholding using fastText for in-group classification.

	Dataset				
Number of Authors	**HP**	**Naruto**	**Sherlock BBC**	**MU**	**SW**
2 + 1	94.4 ± 1.4	98.0 ± 1.6	97.8 ± 2.2	94.9 ± 3.1	93.9 ± 2.7
4 + 1	87.5 ± 3.6	86.8 ± 3.9	89.0 ± 2.4	90.9 ± 3.7	86.9 ± 3.8
9 + 1	80.2 ± 2.6	77.4 ± 5.0	72.3 ± 2.6	74.4 ± 2.3	72.4 ± 2.2
19 + 1	71.3 ± 4.1	65.1 ± 3.4	68.7 ± 4.3	65.5 ± 3.2	64.8 ± 5.1
49 + 1	41.3 ± 3.4	42.2 ± 2.8	40.1 ± 6.3	46.1 ± 4.4	41.8 ± 3.9
Average accuracy	74.9 ± 3.0	73.9 ± 3.3	73.5 ± 3.8	74.3 ± 3.3	71.9 ± 3.5

Table 16. Method based on Euclidean distance for in-group classification.

	Dataset				
Number of Authors	**HP**	**Naruto**	**Sherlock BBC**	**MU**	**SW**
2 + 1	90.5 ± 3.7	91.7 ± 2.3	89.7 ± 3.0	90.1 ± 3.1	88.2 ± 1.9
4 + 1	81.4 ± 3.2	84.3 ± 5.2	78.1 ± 3.7	84.5 ± 3.3	82.6 ± 3.8
9 + 1	68.9 ± 4.3	66.7 ± 1.4	68.6 ± 4.3	83.6 ± 2.4	79.6 ± 2.4
19 + 1	56.4 ± 3.2	55.4 ± 5.2	59.3 ± 4.6	61.9 ± 3.7	60.6 ± 4.3
49 + 1	31.9 ± 5.2	32.8 ± 3.1	35.2 ± 5.4	33.2 ± 5.7	32.9 ± 2.4
Average accuracy	65.8 ± 3.9	66.2 ± 3.4	66.2 ± 4.4	70.7 ± 3.6	68.8 ± 3.0

Table 17. Method based on cosine similarity thresholding for in-group classification.

	Dataset				
Number of Authors	**HP**	**Naruto**	**Sherlock BBC**	**MU**	**SW**
2 + 1	86.3 ± 1.3	89.7 ± 2.1	88.1 ± 3.5	90.0 ± 3.0	87.2 ± 3.1
4 + 1	80.8 ± 2.6	78.2 ± 6.3	74.4 ± 4.4	81.3 ± 4.2	80.1 ± 2.9
9 + 1	72.3 ± 3.9	67.5 ± 2.7	69.3 ± 6.4	70.4 ± 3.6	66.4 ± 4.6
19 + 1	63.1 ± 3.2	55.4 ± 4.2	54.2 ± 3.9	53.6 ± 2.8	49.8 ± 4.8
49 + 1	35.5 ± 7.0	31.7 ± 6.3	30.5 ± 6.2	27.5 ± 3.8	29.2 ± 4.9
Average accuracy	67.6 ± 3.9	64.5 ± 4.5	63.3 ± 4.8	32.5 ± 3.5	62.5 ± 4.0

7. Discussion and Conclusions

The paper considers methods for determining the authorship of classical Russian literary texts, as well as fanfiction data based on popular literature and cinema works. The process of determining the author was considered in the classical version of the classification experiments using a closed set of authors, and experiments were also carried out for a complicated modification of the problem using an open set of authors.

For the case of a closed set of authors, the RbFS method was chosen using a comparison to the GA, FSS, SFS, and SHAP methods. Methods based on full enumeration and sequential inclusion/exclusion of features turned out to be the least effective for any number of features and classes, especially for 20 and 50 authors, where the loss in accuracy compared to the GA reaches 10%. The features selected using the SHAP and RbFS methods were also considered in the comparison. SHAP can outperform the results obtained using the genetic algorithm for a small number of classes, such as 2, 5, and 10 classes, and no more than 100 features.

With a greater number of features, the accuracy is inferior to all considered methods. RbFS is recognized as the main selection method since its combination with SVM made it possible to improve accuracy rates for the maximum of the considered number of classes, i.e., 50 authors, up to 5%. Regarding the more complicated version of closed attribution, the introduction of artificial texts generated using RuGPT-2 and RuGPT-3 based on the original corpus of classics, the maximum accuracy loss for fastText was 9%, and for SVM it was 12%. The greatest loss of accuracy was recorded during experiments for 50 and 20 classes, while for 2 and 5 classes, despite the complication of the task, the losses did not exceed 3%. The average classification accuracy of the original samples was higher in the case of fastText, but the feature selection SVM outperforms fastText in the case of testing the method using the injection of generated samples, which indicates the effectiveness of the feature space.

In the case of open attribution, when testing the model, authors who did not participate in the training were added to the test sample. Thus, in addition to classifying the initially specified authors, we included an additional task: to determine added anonymous texts. Also, the limitation was imposed by the specifics of the data: fanfiction is the result of a fanfiction writer's work. Such texts are based on the original plot, which is the reason why such texts contain features common to all texts of the category including the special names, words, phrases, and expressions inherent in the original characters.

When developing the methods, the importance of the formed feature space containing the frequency distributions was taken into account, as well as the undoubted advantages of NNs that have proven themselves in text classification tasks. Thus, the method based on the joint use of One-Class SVM and fastText combines the above features. Another method based on the probability threshold of the Softmax function for fastText shows comparable results in classification for pairs of thematic categories but does not perform as well as the first method when working with complex cases. For example, One-Class SVM + fastText accuracy reaches 80% for 20 classes and 51% for 50 classes even in the complex case, while the results of the second method do not exceed 75% and 46% for the same cases. In addition, compared to statistical methods based on Euclidean distance and cosine similarity, One-Class SVM + fastText can improve the results by up to 15%. At the same time, the assessment of the One-Class SVM + fastText method in the case of classification within the thematic group allows us to declare independence from the specific subject of amateur texts, since the results vary by no more than 8% from group to group with the same number of classes. Thus, One-Class SVM + fastText was chosen as the recommended method for open attribution. When determining the authorship of amateur texts, the accuracy within categories reaches 98%, and for pairwise comparison of various thematic categories, the accuracy reaches 99%.

It is necessary to highlight informative features when forming the author's style of writing, where the author is considered both a professional writer and a fanfiction writer when writing a text. With the right approach, the combination of such features will make it possible to fully describe the author's style and establish the differences between one author and another. In some related fields, the ability to find writers close in style, and, conversely, unique, unlike others, and to determine the author's age and gender, level of education, etc., can be helpful.

Further research plans include the definition of plagiarism in scientific papers in Russian, as well as the continuation of research aimed at short comments from social network users.

Author Contributions: Supervision, A.R. and A.S.; writing—original draft, A.F. and A.K.; writing—review and editing, A.R., A.S. and A.K.; conceptualization, A.K., A.F. and A.R.; methodology, A.F. and A.R.; software, A.K. and A.R.; validation, A.F. and A.R.; formal analysis, A.S. and A.F.; resources, A.K. and A.S.; data curation, A.F. and A.R.; project administration, A.R.; funding acquisition, A.S. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of Russia, Government Order for 2020–2022, project no. FEWM-2020-0037 (TUSUR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting reported results including links to publicly archived datasets and analysis code. Available online: https://github.com/afedotowaa/authorship_attribution/ (accessed on 20 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Romanov, A.; Kurtukova, A.; Shelupanov, A.; Fedotova, A.; Goncharov, V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* **2021**, *13*, 3. [CrossRef]
2. Romanov, A.S.; Kurtukova, A.V.; Sobolev, A.A.; Shelupanov, A.A.; Fedotova, A.M. Determining the Age of the Author of the Text Based on Deep Neural Network Models. *Information* **2020**, *11*, 589. [CrossRef]
3. Jafariakinabad, F.; Hua, K.A. Unifying Lexical, Syntactic, and Structural Representations of Written Language for Authorship Attribution. *SN Comput. Sci.* **2021**, *2*, 481. [CrossRef]
4. Mahor, U.; Kumar, A. A Comparative Study of Stylometric Characteristics in Authorship Attribution. In *Information and Communication Technology for Competitive Strategies*; ICTCS Springer: Singapore, 2021; pp. 71–81.
5. Fedotova, A.; Romanov, A.; Kurtukova, A.; Shelupanov, A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. *Future Internet* **2022**, *14*, 4. [CrossRef]
6. Russian GPT-2 Model. Available online: https://github.com/vlarine/ruGPT2 (accessed on 19 October 2022).
7. Russian GPT-3 Model. Available online: https://developers.sber.ru/portal/products/rugpt-3?attempt=1 (accessed on 19 October 2022).
8. PAN: Series of Scientific Events and Shared Tasks on Digital Text Forensics and Stylometry. Available online: https://pan.webis.de/ (accessed on 20 October 2022).
9. The 100 Idiolectic Project. Available online: https://fold.aston.ac.uk/handle/123456789/17 (accessed on 20 October 2022).
10. Najafi, M.; Tavan, E. Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis. In Proceedings of the CLEF 2022—Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2021; Available online: https://ceur-ws.org/Vol-3180/paper-215.pdf (accessed on 22 October 2022).
11. PAN at CLEF 2021. Available online: https://pan.webis.de/clef21/pan21-web/index.html (accessed on 25 October 2022).
12. Boenninghoff, B.; Nickel, R.M.; Kolossa, D. O2D2: Out-of-distribution detector to capture undecidable trials in authorship verification. *arXiv* **2021**, arXiv:2106.15825.
13. Weerasinghe, J.; Singh, R.; Greenstadt, R. Feature Vector Difference based Authorship Verification for Open-World Settings. In Proceedings of the CLEF 2021—Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021; pp. 2201–2207.
14. Drozdova, A.; Petrov, V. Modern Classic in the Web Environment: Narrative Variations of V. Nabokov's in Fanfiction. Acta Universitatis Sapientiae. *Film Media Stud.* **2020**, *18*, 89–107.
15. Shafirova, L.; Cassany, D.; Bach, C. Transcultural literacies in online collaboration: A case study of fanfiction translation from Russian into English. *Lang. Intercult. Commun.* **2020**, *20*, 531–545. [CrossRef]
16. Apoorva, K.A.; Sangeetha, S. Deep neural network and model-based clustering technique for forensic electronic mail author attribution. *Appl. Sci.* **2021**, *3*, 348. [CrossRef] [PubMed]
17. Wang, H.; Riddell, A.; Juola, P. Mode effects' challenge to authorship attribution. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 1146–1155.
18. Swain, S.; Mishra, G.; Sindhu, C. Recent approaches on authorship attribution techniques—An overview. In Proceedings of the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; Volume 1, pp. 557–566.
19. Hedegaard, S.; Simonsen, J.G. Lost in translation: Authorship attribution using frame semantics. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 65–70.
20. Wu, H.; Zhang, Z.; Wu, Q. Exploring syntactic and semantic features for authorship attribution. *Appl. Soft Comput.* **2021**, *111*, 107815. [CrossRef]

21. Alharthi, H.; Inkpen, D.; Szpakowicz, S. Authorship identification for literary book recommendations. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 390–400.
22. The Litrec Dataset. Available online: https://www.inesc-id.pt/publications/8386/pdf (accessed on 2 November 2022).
23. Kovalev, A.K.; Kuznetsova, Y.M.; Minin, A.N.; Penkina, M.Y.; Smirnov, I.V.; Stankevich, M.A.; Chudova, N.V. Methods for identifying the psychological characteristics of the author in the text (on the example of aggressiveness). *Cyber Secur. Issues* **2019**, *4*, 72–79. [CrossRef]
24. Isachenko, V.V.; Apanovich, Z.V. Analysis and visualization system for cross-language identification of authors of scientific publications. Bulletin of the Novosibirsk State University. *Ser. Inf. Technol.* **2018**, *16*, 49–61.
25. Sokolova, T.P. Problems of expert identification in forensic autonomy. *Bull. O.E. Kutafin Univ.* **2022**, *2*, 67–76. [CrossRef]
26. Bardamova, M.; Hodashinsky, I. Hybrid Algorithm for Tuning Feature Weights in a Fuzzy Classifier. In Proceedings of the 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Yekaterinburg, Russia, 13–14 May 2021; pp. 354–357.
27. Feofanov, V.; Devijver, E.; Amini, M.R. Wrapper feature selection with partially labeled data. *Appl. Intell.* **2022**, *52*, 12316–12329. [CrossRef]
28. Anwar, W.; Bajwa, I.S.; Choudhary, M.A.; Ramzan, S. An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution. *IEEE Access* **2018**, *7*, 3224–3234. [CrossRef]
29. Morales Sánchez, D.; Moreno, A.; Jiménez López, M.D. A White-Box Sociolinguistic Model for Gender Detection. *Appl. Sci.* **2022**, *12*, 2676. [CrossRef]
30. Rangel, F.; Giachanou, A.; Ghanem, B.H.H.; Rosso, P. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR Workshop Proceedings*; Sun SITE Central Europe: Aachen, Germany, 2020; Volume 2696, pp. 1–18.
31. Bevendorff, J.; Chulvi, B.; Fersini, E.; Heini, A.; Kestemont, M.; Kredens, K.; Zangerle, E. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Bologna, Italy, 5–8 September 2022; Springer: Cham, Switzerland, 2022; pp. 382–394.
32. Krassa, S.I.; Kalugina, E.N. Gender profiling of the author of the subprime text. *Bull. South Ural State Univ. Ser. Linguist.* **2014**, *11*, 19–22. (In Russian)
33. Khazova, A.B. Automatic determination of the gender of the author of the text: The phenomenon of Russian women's prose. Bulletin of the Novosibirsk State University. *Ser. Linguist. Intercult. Commun.* **2020**, *18*, 22–32.
34. Kovács, G.; Balogh, V.; Mehta, P.; Shridhar, K.; Alonso, P.; Liwicki, M. Author Profiling Using Semantic and Syntactic Features: Notebook for PAN at CLEF 2019. 2019. Available online: https://core.ac.uk/download/pdf/287813157.pdf (accessed on 21 December 2022).
35. Alvarez-Carmona, M.A.; Villatoro-Tello, E.; Villasenor-Pineda, L. A comparative analysis of distributional term representations for author profiling in social media. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4857–4868. [CrossRef]
36. Nguyen, D.; Trieschnigg, D.; Doğruöz, A.S.; Gravel, R.; Theune, M.; Meder, T.; de Jong, F. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland, 23–29 August 2014; pp. 1950–1961.
37. PAN Data. Available online: https://pan.webis.de/data.html (accessed on 21 December 2022).
38. Victorian Era Authorship Attribution Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/Victorian+Era+Authorship+Attribution (accessed on 21 December 2022).
39. Blog Authorship Corpus. Available online: https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus (accessed on 21 December 2022).
40. Russian Literature. Available online: https://www.kaggle.com/datasets/d0rj3228/russian-literature (accessed on 21 December 2022).
41. Authorship Attribution for Russian Literature. Available online: https://www.kaggle.com/code/d0rj3228/authorship-attribution-for-russian-literature (accessed on 21 December 2022).
42. Ficbook: Fanfiction Book. Available online: https://ficbook.net/ (accessed on 19 November 2022).
43. Zhao, H.; Hu, Q.; Zhu, P.; Wang, Y.; Wang, P. A recursive regularization based feature selection framework for hierarchical classification. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 2833–2846. [CrossRef]
44. Ren, J.; Qiu, Z.; Fan, W.; Cheng, H.; Yu, P.S. Forward semi-supervised feature selection. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Osaka, Japan, 20–23 May 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 970–976.
45. Marcílio, W.E.; Eler, D.M. From explanations to feature selection: Assessing shap values as feature selection mechanism. In Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Porto de Galinhas, Brazil, 7–10 November 2020; pp. 340–347.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.