information

MDPI

Article

Genre Classification of Books in Russian with Stylometric Features: A Case Study

Natalia Vanetik *,†, Margarita Tiamanova †, Genady Kogan † and Marina Litvak †

Department of Software Engineering, Shamoon College of Engineering, Beer Sheva 84500, Israel; margati@ac.sce.ac.il (M.T.); genadko@ac.sce.ac.il (G.K.); marinal@sce.ac.il (M.L.)

* Correspondence: natalyav@sce.ac.il; Tel.: +972-8-647-5015

† These authors contributed equally to this work.

Abstract: Within the literary domain, genres function as fundamental organizing concepts that provide readers, publishers, and academics with a unified framework. Genres are discrete categories that are distinguished by common stylistic, thematic, and structural components. They facilitate the categorization process and improve our understanding of a wide range of literary expressions. In this paper, we introduce a new dataset for genre classification of Russian books, covering 11 literary genres. We also perform dataset evaluation for the tasks of binary and multi-class genre identification. Through extensive experimentation and analysis, we explore the effectiveness of different text representations, including stylometric features, in genre classification. Our findings clarify the challenges present in classifying Russian literature by genre, revealing insights into the performance of different models across various genres. Furthermore, we address several research questions regarding the difficulty of multi-class classification compared to binary classification, and the impact of stylometric features on classification accuracy.

Keywords: text classification; genre classification; Russian literature; stylometry; genres dataset

Citation: Vanetik, N.; Tiamanova, M.; Kogan, G.; Litvak, M. Genre Classification of Books in Russian with Stylometric Features: A Case Study. *Information* **2024**, *15*, 340. https://doi.org/10.3390/info15060340

Academic Editor: Katsuhide Fujita

Received: 5 May 2024
Revised: 24 May 2024
Accepted: 4 June 2024
Published: 7 June 2024

1. Introduction

In the realm of literature, the concept of genre serves as a fundamental organizational principle, providing readers, publishers, and scholars with a cohesive framework. Within any collection of works, a genre stands as a distinct category, characterized by shared stylistic, thematic, and structural elements. It functions as a conceptual tool, simplifying the process of categorization and enhancing the comprehension of a diverse array of literary expressions. Genres encompass a broad spectrum, ranging from timeless and conventional categories like fiction and non-fiction to more nuanced classifications such as mystery, romance, fantasy, and beyond.

Genres extend beyond mere categorization; they act as guiding beacons, directing readers toward narratives that align with their preferences and expectations. Recognizing the genre of a literary work becomes akin to following arrows that point toward stories tailored to individual tastes.

A text's stylistic features serve as markers of different genres and are frequently employed for automatic analysis in this field because they represent a text's structural quirks, among other things [1,2].

Automatic genre classification makes it possible to solve several computational linguistics problems more quickly, including figuring out a word or phrase's meaning or part of speech, locating documents that are pertinent to a semantic query [3], improving authorship attribution [4–6], and more [2,7].

Only a small number of the numerous studies that address the subject of automatic genre classification (explained in more detail in Section 2) focus on Russian literature, and just three corpora have been generated for the job of genre classification in Russian, even

though there are over 258 million Russian speakers in the world [8]. The corpus introduced in [9] contains texts collected from the Internet that belong to six genre segments of Internet texts, namely, contemporary fiction, poetry, social media, news, subtitles for films, and a collection of thematic magazines annotated as "the rest". These texts span 5 billion words; however, out of the six genres, only two can be attributed to literature—fiction and poetry. The corpus of [2] contains 10,000 texts assigned to five different genres—novels, scientific articles, reviews, posts from the VKontakte social network [10], and news texts from OpenCorpora [11], the open corpus of Russian texts. Only one genre in this corpus is a literature genre—the novels. In [12], the authors have developed a corpus of the A.S. Pushkin Lyceum period (1813–1817) that contains ten different genres of his poems. However, no prosaic texts are contained in this corpus, and the texts are limited to a single author.

None of the above corpora covers a significant amount of modern and historical genres in Russian literature. To overcome this gap, we present a new dataset comprising Russian books spanning eleven diverse literature genres, aimed at facilitating research in text classification. The dataset encompasses eleven different literature genres, thereby providing a comprehensive resource for studying genres in Russian literature. We evaluate several traditional machine learning models, alongside state-of-the-art deep learning models, including transformers [13] and dual contrastive learning [14], for both binary and multi-class genre classification tasks. Furthermore, we provide insights into the strengths and limitations of each model, shedding light on their applicability to real-world genre classification scenarios. This dataset can serve as a valuable resource for researchers interested in advancing the understanding and development of genre studying and classification systems for Russian texts.

We perform an extensive evaluation of binary and multi-class genre classification on a subset of our dataset and analyze the results; we employ a wide range of text representations, including stylometric features [2]. The purpose of this evaluation is to show that genre classification of Russian books is a highly nontrivial task. We also analyze what text representations work better for what task, and the difference in classification of different genres.

We address the following research questions in our work.

- RQ1: Do stylometric features improve genre classification accuracy?
- RQ2: What genres are easier to classify?
- RQ3: Does contrastive learning perform better for genre classification than fine-tuned transformer models and traditional models?
- RQ4: Does removing punctuation decrease classification accuracy for genre classification?
- RQ5: Does a transformer model pre-trained on Russian perform better than a multilingual transformer model?

This paper is organized as follows. Section 2 describes the related work. Section 3 describes our dataset, the process of its collection, and the data processing we performed. In Section 4, we describe text representations and classification models we used to perform genre classification on our data. Section 5 describes the hardware and software setup and full results of our experimental evaluation. Finally, Sections 6 and 7 discuss the conclusions and limitations of our approach.

2. Related Work

The task of categorizing Russian literary genres has garnered significant attention in the domains of literary analysis and Natural Language Processing (NLP). Many studies have attempted to enhance the understanding and automation of genre classification, focusing on the unique linguistic and cultural components present in Russian literature. We must acknowledge the growing body of research focusing on topics other than English literature, even though a lot of work has helped us comprehend how genres are applied to English-language novels. Renowned research on Arabic [15] and Spanish [16] genre classification has highlighted both the benefits and drawbacks of linguistic and cultural diversity.

Genre classification occurs not only at the level of texts but also based on book titles. The authors of [17] presented a method for genre classification based on the book's title. The dataset (available at https://github.com/akshaybhatia10/Book-Genre-Classification, accessed on 1 January 2024) constructed by the authors contains 207,575 samples of data assigned to 32 different genres.To represent the data, the texts were converted to lowercase, tokenized, and stemmed. Punctuation and English stopwords were removed. Word embeddings were used as word representations, and five different machine learning models were applied for the task of genre classification by title. The best-performing model was Long Short-Term Memory (LSTM) with dropout, achieving an accuracy of 65 %.

The work of [16] addresses genre classification in Spanish [16]. The authors introduce a method for automatic detection of books' themes that does not rely on a pre-defined list of genres. The authors construct the dataset by scraping the books from two undisclosed Latin American publishers. Their approach clusters key categories and results in 26 thematic categories. Models such as SVM [18] and BERT [13] achieve F1 scores ranging from 57% to 65.26%.

The work of [19] uses Recurrent Neural Networks (RNNs) as a deep learning method to classify book plots and reviews. The successful classification of 28 genres, including action, adventure, comedy, drama, family, mystery, romance, and science, is demonstrated by the testing findings. For the top 10 recommendations, the RNN-based recommendation system outperforms the matrix factorization technique with a precision of 82%, compared to 77%. In comparison to conventional artificial neural network techniques, the study indicates that combining a deep learning model with an RNN enhances accuracy and lowers validation loss percentage, improving Root Mean Squared Error (RMSE).

Several corpora for genre classification have been developed over the years in multiple languages, such as English, Arabic, Spanish, and more [20–23]. Not much analogous research has been conducted on datasets in the Russian language. The authors of [2] investigated contemporary vector text models, such as ELMo embeddings, the BERT language model, and a complex of numerical rhythm features, for genre categorization of texts written in the Russian language. Their experiments used ten thousand texts from five different genres: OpenCorpora news, Vkontakte communications, reviews, scientific publications, and novels. The study differentiated between genres with limited rhythms and those with diverse rhythms, such as novels and reviews, using rhythm features and LSTM. The multi-classification F-score of 0.99 attests to the effectiveness of contemporary embeddings. In [12], an automated technique for classifying Russian poetry writings' genre type and semantic properties is proposed. Based on the relationship between genre and text style, it describes a combined classifier for genre kinds and stylistic coloring. Computational tests with A.S. Pushkin's Lyceum lyrics show good results in determining stylistic colors and genres. The author of [24] proposed a method for genre classification of Russian texts that relies on different feature types—lexical, morphological, and syntactic features, as well as readability, text and word length, and symbolic features (n-grams). Support Vector Machine [25] is then used as a classifier for different genre types. All the experiments in this work are performed on the "Taiga" webcorpous [9] that has undergone morphological and syntactic annotation and covers six different genres of Internet texts.

As a representation learning technique, contrastive learning seeks to maximize similarity between samples in the same class and minimize it across samples in different classes [26]. By including both input data and their related labels, dual contrastive learning expands on this strategy and makes it easier to learn discriminatory representations for both aspects [27].

Our study presents a substantial dataset for the genre classification of Russian books, and the evaluation of this dataset with binary genre classification. The dataset is extensive, with genres represented in varying proportions, due to some genres being more popular while others are less so. Our dataset covers 11 genres and more than 8K books.

We perform our study with multiple classifiers, including traditional ones (Random Forest [28], logistic regression (LR) [29], and Extreme Gradient Boosting [30]), their ensem-

ble, transformers, and contrastive learning. The RF model is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes of the individual trees; the LR model finds the probability of a binary outcome based on one or more predictor variables, and it uses a logistic function to represent the output as a probability; XGB performs gradient boosting by building the models sequentially with each new model correcting the errors made by previous ones. The ensemble learning approach involves combining the predictions of multiple models (RF, LR, and XGB) by combining their predicted probabilities. This method uses the strengths of each model to enhance the overall prediction performance. Fine-tuned transformers are the pre-trained transformer models that have been further trained on our dataset, allowing the model to adapt to specific tasks. A contrastive learning model is trained to differentiate between similar and dissimilar pairs of data points. Full descriptions of these models are provided in Section 4.4.

Stylometry, also called computational stylistics, encompasses a wide-ranging field of studies that involves analyzing linguistic features extracted from texts to characterize the style of authors, documents, or document groups. Using statistical techniques, subtle differences and similarities between texts that may be imperceptible to the naked eye can be identified, allowing the delineation of text groups based on their linguistic affinity. Stylometry subtasks include authorship attribution, authorship verification, authorship profiling, stylochronometry, and adversarial stylometry [31]. In some cases, authorship profiling helps narrow the search space by identifying different variables such as genre, age, or gender [32–34]. Stylometry is valuable for genre classification due to its ability to capture unique stylistic features that distinguish different genres. Word frequency, sentence length, and syntactic patterns are examples of stylometric characteristics that may be used to distinguish between genres by emphasizing writers' consistent stylistic choices within each genre. Furthermore, the efficiency of stylometry in genre categorization has been further validated by its successful application in tasks such as author profiling across genres [35].

Several tools for stylometric analysis have been developed, and a number of them support the Russian language. Stylo [36] is an R package that supports various languages, including Russian, and provides functionalities for authorship attribution, stylistic analysis, and text classification. WebSty [37] is another stylometric analysis tool that supports Russian language processing. It is an easily accessible open-source tool and forms part of the CLARIN-PL research infrastructure. While primarily focused on English texts, Coh-Metrix [38] is a comprehensive tool for analyzing various aspects of text cohesion, coherence, and readability. It has inspired versions for analyzing other languages, including Russian, albeit with fewer functionalities compared to its English counterpart. Finally, StyloMetrix [39] is an open-source multi-lingual tool specifically designed for representing stylometric vectors. It supports Russian language analysis, as well as English, Polish, Ukrainian, and German. This tool offers a multitude of features for syntactic and lexical vector representation.

Stylometry subtasks include authorship attribution, authorship verification, authorship profiling, stylochronometry, and adversarial stylometry [31]. In some cases, authorship profiling helps narrow the search space by identifying different variables such as genre, age, or gender [32–34].

We use the StyloMetrix [39] package to compute over 90 stylometric features for Russian texts and use them further in our text representation (described in Section 4.3.3). A full list of these features is provided in the Appendix A.

Language models, such as RuBERT and multi-lingual BERT (mlBERT), are applied to evaluate the dataset quality and see how they behave, with and without contrastive learning, compared to traditional models. In the context of genre classification, transformers are used to automatically learn representations of text documents that capture the stylistic and semantic features associated with different genres [40,41]. Transformers have been already applied to genre classification in Spanish [16].

There are several pre-trained transformer models available in the literature and in the HuggingFace repository [42], and we describe them here. The BERT multi-lingual model [13] is a multi-lingual variant of BERT (Bidirectional Encoder Representations from Transformers) which supports multiple languages, including Russian, and the "ruBERT-base-cased" transformer model [43] which is an adapted BERT for Russian language-processing tasks. The ruBERT model extends the original BERT architecture [13] by fine-tuning pre-trained multi-lingual embeddings specifically for the Russian language. The adapted model demonstrates improved performance on Russian language tasks compared to the original multi-lingual BERT model.

We use contrastive learning to improve representation learning, in which a model is trained to minimize similarity between samples from different genres and increase similarity between samples of the same genre [26]. By considering both the input data and their corresponding labels, dual contrastive learning encourages the model to learn representations that are discriminatory for both the input data and their labels (genres in our case).

The dual contrastive learning (DualCL) method is suggested in [27] to adapt the contrastive loss to supervised contexts. The DualCL framework learns both the classifier parameters and the properties of input samples simultaneously in the same space. To be more specific, DualCL uses contrastive learning to differentiate between the input samples and these augmented samples, treating the classifier parameters as improved samples associated with unique labels. In Natural Language Processing (NLP), contrastive learning has been already applied to tasks such as text classification, sentiment analysis, and language modeling [14].

In addition to providing a novel dataset and evaluating its quality, our contribution addresses the applicability of stylometry, contrastive learning, and modern language models in genre categorization.

3. The SONATA Dataset

We have considered several online sources of Russian literary texts. Our requirements were as follows:

- A wide, up-to-date, and legitimate selection of titles, and agreements with leading Russian and international publishers;
- Clear genre labels and a wide selection of genres;
- The option to freely and legally download a significant number of text samples in .txt format;
- A convenient site structure that allows automated data collection.

The following options were examined. The LitRes site [44] contains a wide range of e-books across various genres, including fiction, non-fiction, educational materials, and more. For most of the books, text fragments but not the whole texts can be downloaded. However, to use LitRes API and to automatically download multiple text samples, a user is required to pay with a credit card issued in Russia, which may not be suitable for some researchers. http://royallib.com/ (accessed on 1 January 2024) is an online library that offers a large collection of free electronic books [45]. The site provides access to a wide range of e-books, including classic literature, modern novels, non-fiction, educational materials, and more. This site offers books for free, making literature accessible to a broad audience. However, this feature also implies that the text collection is outdated because most modern Russian books are the subject of a copyright. Finally, https://knigogo.net/ (accessed on 1 January 2024) is a Russian-language website that provides fresh news about literature, reviews, and feedback on popular books. It contains a large selection of audiobooks and online books in formats such as fb2, rtf, epub, and txt for iPad, iPhone, Android, and Kindle. It has clear genre labels and a convenient structure that allows efficient parsing. Moreover, it provides free access to text samples.

For these reasons, we have chosen to collect our data from https://knigogo.net/ (accessed on 1 January 2024). We name the resulting dataset SONATA for ruSsian bOoks

geNre dATAset. Genre categories were translated into English for the reader's benefit because the source website [46] supports the Russian language only. The dataset is available on GitHub at https://github.com/genakogan/Identification-of-the-genre-of-books.

3.1. The Genres

To build our dataset, we used the genres provided by the Knigogo website at https://knigogo.net/zhanryi/ (accessed on 1 January 2024). Because not all genres and sub-genres provided by the website have a sufficient amount of data for analysis, we filtered out the less-represented genres and combined sub-genres where appropriate, aimed to streamline the classification for more meaningful insights.

As a result, 11 genres were selected for the dataset: science fiction, detective, romance, fantasy, classic, action, non-fiction, contemporary literature, adventure, novel and short stories, and children's books. In non-fiction, we encompassed all genres that do not belong to fiction literature. Due to the relatively small number of books in each sub-genre within non-fiction, considering each sub-genre separately would not be productive for our experiment. The original list of genres on Knigogo and the list of selected genres are depicted in Figure 1. All genres, covered by the SONATA dataset, and their translations are shown in Table 1.

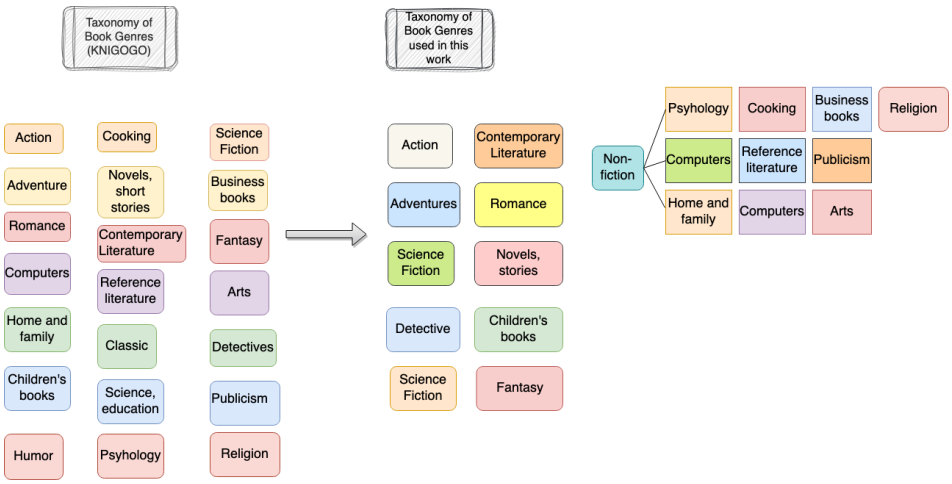

Figure 1. Genre taxonomy in Knigogo and a subset of genres used in the SONATA dataset.

Table 1. Genres in Russian and their translation to English.

Genre (Russian)	Translation
фантастика	Science fiction
Adventure	Detective
Роман	Romance
фэнтези	Fantasy
Классика	Classics
Боевик	Action
Нехудожественная литература	Non-fiction
Современная литература	Contemporary literature
Приключения	Adventure
Новеллы, рассказы	Short stories
Детские книги	Children's books

3.2. Data Processing and Statistics

Book downloading was performed by a Python script; the script extracts URLs leading to pages where individual books can be downloaded first; then, it extracts the URLs for

direct download of the text files of these books. Then, the script attempts to download each book, saving them as text files in a directory. The execution of the script is initiated using a specific URL associated with fantasy genre books. A more detailed description of the script is available in the Appendix A.

As a result, 8189 original books were downloaded. However, because some books belong to multiple genres, the total amount of book instances with a single genre label is 10,444.

During the data processing, a series of challenges arose, such as text formatting, removing author names and publisher information, and the volume of texts. We applied the following steps: (1) using a custom Python script, we re-encoded the files into a UTF-8 format; (2) we parsed the second line of text containing the meta-data and removed the authors' names; (3) finally, we extracted books without author names and split them into small parts (denoted by chunks) of 300 words each. Splitting text into chunks allows us to process long texts that exceed the length limit in many pre-trained language models (LMs).

The amount of books and book chunks per genre in the SONATA dataset appears in Table 2. Because some books are attributed to several genres on the original site, the total unique number of books is smaller than the sum of books per genre. We report both of these values in the table.

Table 2. Size of the collected data.

Genres	Books Number	Chunks Number
action	640	59,040
adventure	120	6969
children's	282	4801
classic	463	20,919
detective	1303	51,022
science-fiction	1909	244,217
fantasy	2595	113,044
non-fiction	896	22,632
contemporary	811	28,648
short stories	206	3798
romance	1219	39,779
all genres	10,444	594,869
all genres, unique	8189	414,574

4. Binary and Multi-Class Genre Classification

In binary genre classification, the task involves categorizing texts into two distinct genres. For example, a text can be classified as either fiction or non-fiction, romance or thriller, positive or negative sentiment, etc. In multi-class genre classification, texts are classified into more than two genres or categories. The task is usually more complex than binary genre classification, as the classifier needs to differentiate between multiple classes and assign the most appropriate genre label to each text. Multi-class genre classification problems are often encountered in large-scale text categorization tasks, where texts can belong to diverse and overlapping genres.

4.1. The Pipeline

To evaluate the SONATA dataset for tasks of binary and multi-class genre classification, we first processed and sampled the data (see details in Section 4.2) and generated the appropriate text representation (see details in Section 4.3). Then, we split the text representations into training and test sets, trained the selected model (see Section 4.4) on the training set and evaluated the model on the test set. This pipeline is depicted in Figure 2.

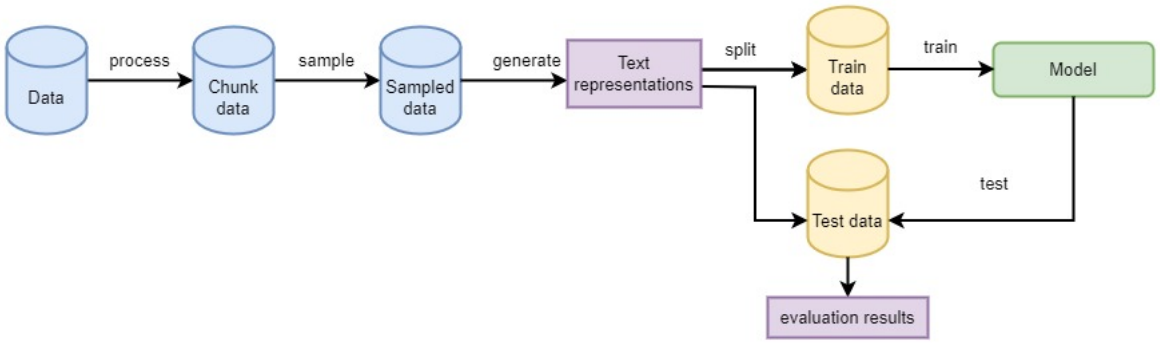

Figure 2. Evaluation pipeline.

4.2. Preprocessing and Data Setup

We did not change the case of the texts and did not remove punctuation in the main setup. The effect of these optional operations on evaluation results is addressed in the Appendix A.

Because of the hardware limitations and data size, we could not apply classification models to the whole dataset. To construct a sample of our dataset, we first selected one random text chunk from every book to avoid the case of author recognition instead of genre recognition. Then, we sampled N chunks at random from every genre, where N is a user-defined parameter. In our experiments, we used $N = 100$. The number of text chunks, average character counts, and the number of unique words for sample size $N = 100$ and every genre is shown in Table 3 (we do not report the average number of words per chunk because all the chunks in our data contain exactly 300 words). The effect of smaller and larger values of N is addressed in the Appendix A.

Table 3. Data statistics.

Genre	Total Chunks	Avg Chars per Chunk	Unique Words	Word(s) with the Highest Frequency	Translation
action	630	2604.4	48,058	просто	simply
adventure	120	2596.1	16,241	время, сказал	time, said
children's	279	2549.8	26,423	сказал	said
classic	479	2164.9	40,127	сказал	said
contemporary	802	2587.3	61,526	очень	very
detective	989	2578.1	63,937	очень	very
fantasy	992	2621.2	67,054	просто	simply
non-fiction	886	2730.2	61,922	которые	which/what
romance	994	2528.7	58,459	просто	simply
science-fiction	989	2632.1	70,551	просто	simply
short-stories	206	2511.0	21,788	очень	very
all	7366	2576.5	200,285	просто	simply

For the binary genre classification task, we select text chunks as a balanced random sample of size N where N is a user-defined parameter. If a genre contains fewer than N book chunks, we select all of them. In each sample, half of the chunks represent positive samples belong to the selected genre, and the other half contain book chunks that represent negative samples and are chosen in a uniformly random fashion from all the other genres. We ensure that no chunks belonging to the same book fall into different sample categories. The positive samples are labeled 1, and the negative samples are labeled 0. For the multiclass genre classification task, we select a random sample of N text chunks from every genre, where N is a user-defined parameter. If a genre contains fewer than N book chunks, all of them are added to the data. The label of every sample is determined by its genre and is a number in the range $[0 \ldots 10]$.

For the evaluation, the obtained balanced dataset is randomly divided into training and test sets with the ratio 80%/20%. This process is illustrated in Figure 3.

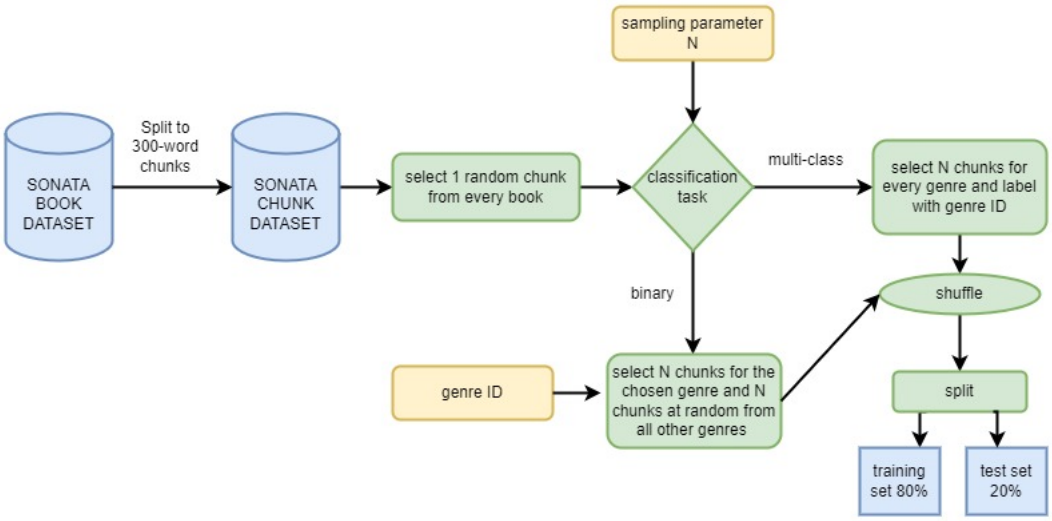

Figure 3. Evaluation pipeline.

4.3. Text Representations

We represent texts as vectors and optionally enhance them with stylometric features. Details are provided in the subsections below. Figure 4 depicts the general pipeline of text representation construction.

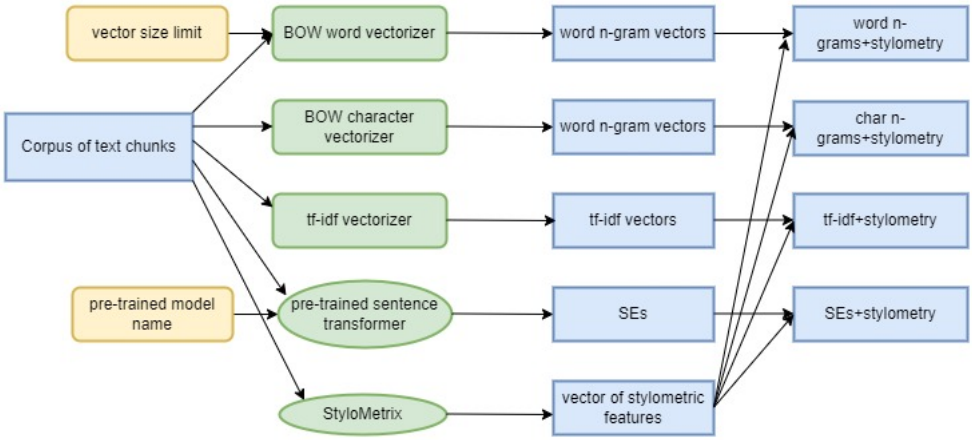

Figure 4. Text representations.

4.3.1. Sentence Embeddings

BERT sentence embeddings [13] are vector representations of entire sentences generated using the BERT model. These embeddings can then be used as features for various downstream NLP tasks. The sentence embeddings (SEs) we use in this work were obtained using one of the pre-trained BERT models (a multi-lingual model of [13] or a Russian BERT model [43]). With both models, the SE vector size is 768.

Figure 5 shows the distribution of books from different genres, where every book is represented by its SE vector computed with ruBERT. For data visualization, we used t-Distributed Stochastic Neighbor Embedding (t-SNE), a common dimensionality reduction technique [47]. It is designed to preserve pairwise similarities, making it more effective at capturing non-linear structures and clusters in high-dimensional data. We can see that contemporary literature (top left) and science fiction (top right) are the only genres for which the data points are partially clustered together. This plot demonstrates that genre classification is a non-trivial task and relying solely on SE can be challenging.

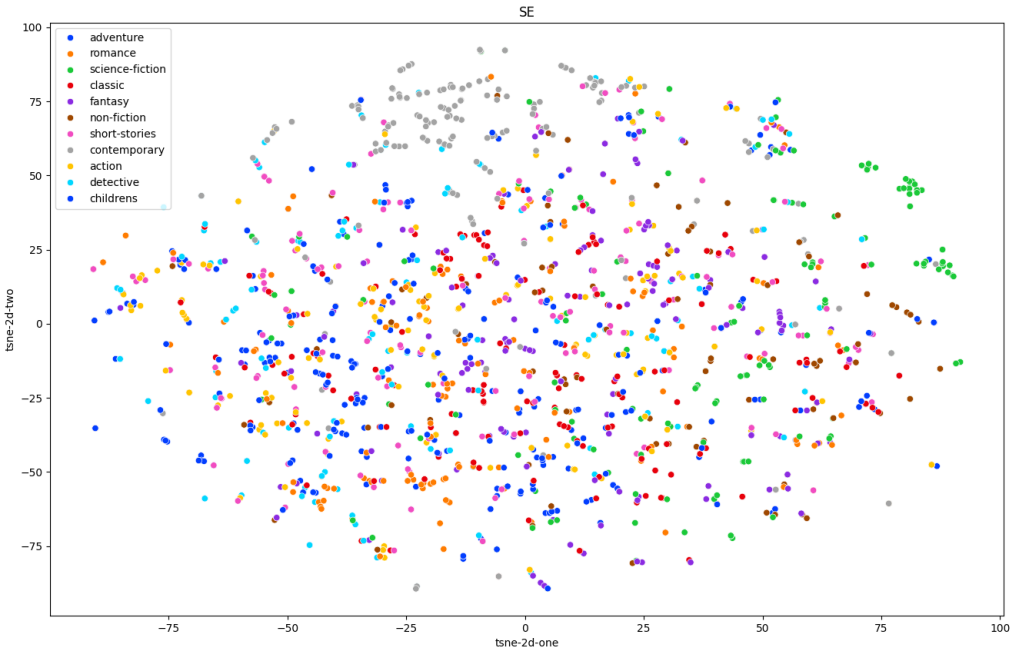

Figure 5. Sentence embedding features of 11 genres represented by t-SNE for samples of size $N = 100$.

4.3.2. BOW Vectors with tf-idf and n-Gram Weights

The concept of Term Frequency-Inverse Document Frequency (tf-idf) constitutes a quantitative measure designed to signify the significance of a term within a document set or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. In our methodology, book chunks were treated as discrete documents, and the entire dataset was regarded as the corpus. We filtered out Russian stopwords using the list provided by the NLTK package [48], to which we added the word 'это' (meaning 'this'); details are provided in the Appendix A.

N-grams are the sequences of n consecutive words or characters seen in the text, where n is a parameter. In our evaluation, we used the values $n = 1, 2, 3$ for both character and word n-grams. In the case of word n-grams, we filtered out Russian stopwords as well using the list provided by NLTK [48]. Vector sizes for these representations for the samples of size $N = 100$ are shown in Table 4. For the multi-class setup, we empirically limited the number of word n-gram features to 10,000 due to very large vector sizes. This does not affect our analysis because this text representation does not provide the best performance.

Table 4. BOW vector lengths for samples of size $N = 100$.

Genre	Char n-Grams Vector Sizes $n = [1, 2, 3]$	Word n-Grams Vector Sizes $n = [1, 2, 3]$	tf-idf Vector Size
action	14,665	46,877	21,781
adventure	13,319	29,283	16,241
children's	13,727	36,189	17,489
classic	16,278	42,994	22,414
contemporary	16,053	46,980	23,588
detective	13,962	46,734	21,841
fantasy	13,389	47,811	22,830
non-fiction	19,688	48,554	23,176
romance	13,237	44,325	19,902
science-fiction	14,759	48,227	23,806
short-stories	14,765	31,974	16,561
all	32,500	403,194	101,350

4.3.3. Stylometric Features

StyloMetrix [49] is a multi-lingual tool designed for generating stylometric vectors for texts in Polish, English, German, Ukrainian, and Russian introduced in [39]. These vectors encode linguistic features related to writing style and can be used for authorship attribution, and genre classification. A total of 95 metrics are supported for the Russian language. The metrics describe lexical forms, parts of speech, syntactic forms, and verb forms. The lexical metrics provide information on plural and singular nouns, and additional morphological features such as animacy (animate/inanimate), gender (feminine, masculine, and neutral), distinguishing between first and second names, and diminutive. Direct and indirect objects as well as cardinal and ordinal numerals are also included in the metrics. Six distinctive lexical forms of pronouns such as demonstrative, personal, total, relative, and indexical are reported, as well as qualitative, quantitative, relative, direct and indirect adjectives. Other lexical features include punctuation, direct speech, and three types of adverb and adjective comparison. A partial list of stylometric features for Russian is provided by the authors of the StyloMetrix package on their GitHub repository [50]; we present a compact list of these features in the Appendix A. We compute stylometric features with StyloMetrix for text chunks in our dataset and use them alone or in conjunction with text representations described in previous sections.

Data visualization of stylometric features with t-SNE for data samples of size $N = 100$ and 11 genres is shown in Figure 6. We can see that, similarly to SEs, contemporary literature (bottom left) and science fiction (bottom right) are the only genres for which the data points are partially clustered together. Therefore, relying solely on stylometric features for genre classification is not expected to produce a good result.

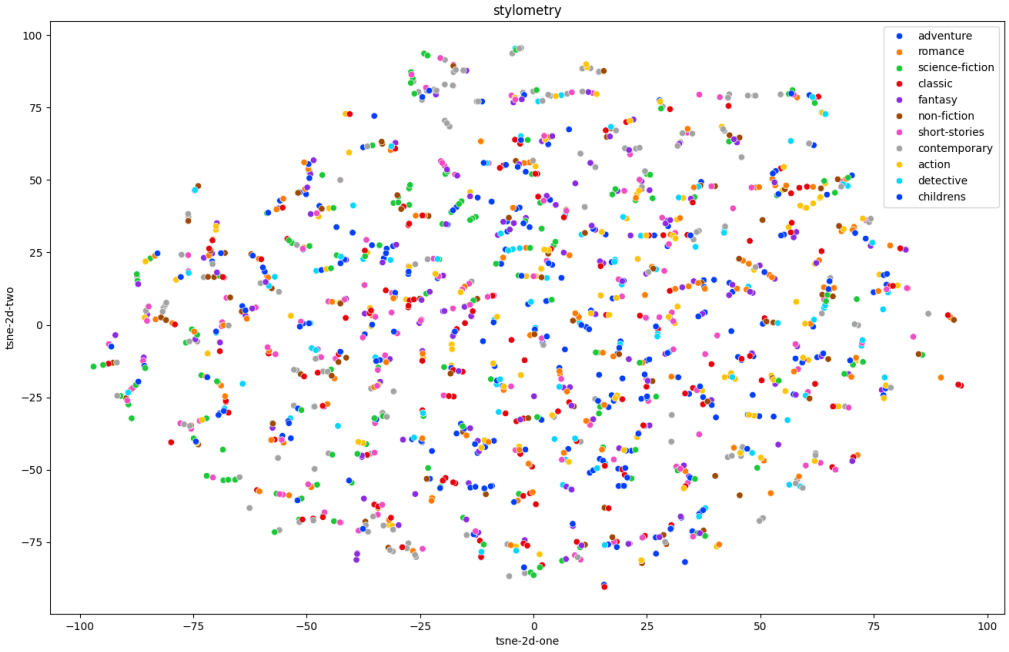

Figure 6. Stylometric features of 11 genres represented by t-SNE for samples of size $N = 100$.

4.4. Classification Models

4.4.1. Traditional Models

An ensemble learning technique called the Random Forest (RF) [28] classifier works by building several decision trees during training. A bootstrapped sample of the training data and a random subset of the characteristics are used to build each tree in the forest. The logistic function, which converts input values into probabilities between 0 and 1, is used in logistic regression (LR) [29] to represent the relationship between the predictor variables and the probability of the result. Gradient boosting is used in the Extreme Gradient Boosting (XGB) [30] classifier, an ensemble learning technique that creates a predictive model. It

creates a sequence of decision trees repeatedly, trying to fix the mistakes of the preceding ones with each new tree. We apply RF, LR, and XGB models to all data representations described in Section 4.3.

4.4.2. Voting Ensemble of Traditional Models

A voting-based ensemble classifier approach combines the predictions of multiple base classifiers to make a final decision. Each base classifier is trained independently on the same dataset or subsets of it using different algorithms or parameter settings [51]. During the prediction phase, each base classifier provides its prediction for a given instance. The final prediction is then determined by aggregating the individual predictions through a voting mechanism, where the most commonly predicted class is selected as the ensemble's prediction.

In our voting setup, we use RF, LR, and XGB classifiers described in Section 4.4.1. For the binary genre classification setup, the decision is made based on the majority vote of the three classifiers (i.e., max voting). For the multi-class genre classification setup, we computed the sum of probabilities of every class based on the individual probability distributions produced by each classifier and assigned each sample to the class with the highest accumulated probability.

4.4.3. Fine-Tuned Transformers

We fine-tune and apply fine-tuned transformer models to the texts in our dataset for both tasks—binary genre classification and multi-class genre classification.

The main transformer model we employ is the ruBERT (Russian BERT) model, specifically the `ruBERT-base-cased` variant, which is trained on large-scale Russian derived from the Russian part of Wikipedia and news data [43]. The baseline transformer model is the BERT multi-lingual base model `bert-base-multilingual-uncased` developed by GoogleAI [13]. The model is pre-trained on the top 102 languages (including Russian) with the largest Wikipedia using a masked language modeling objective. We denote this model by mlBERT. Both models utilize a BERT transformer architecture, which employs a bidirectional approach that allows to capture of contextual information from both left and right contexts.

4.4.4. Dual Contrastive Learning

To tackle the problem of genre classification, we also apply the advanced text classification method DualCL [27] that uses contrastive learning with label-aware data augmentation.

The objective function used in this method consists of two contrastive losses, one for labeled data and another for unlabeled data. Contrastive loss is computed for each labeled instance (x_i, y_i) as

$$\mathcal{L}_L = -\log \frac{e^{f(x_i)\cdot g(y_i)}}{\sum_{j=1}^{N} e^{f(x_i)\cdot g(y_j)}}$$

where $f(x_i)$ is the feature representation of input x_i, y_i is the corresponding label of x_i, $g(y_i)$ is the embedding of label y_i, and N is the total number of classes. In our experiments, we use the pre-trained transformer model ruBERT [43] as the basis for feature representation computation. The number of classes N is set to 2 for the task of binary genre classification, and to 11 for the multi-class genre classification. Contrastive loss for unlabeled data is computed as

$$\mathcal{L}_U = -\log \frac{e^{f(x_i)\cdot f(x_j)/\tau}}{\sum_{k=1}^{M} e^{f(x_i)\cdot f(x_k)/\tau}}$$

where $f(x_i)$ and $f(x_j)$ are the feature representations of inputs x_i and x_j, M is the total number of unlabeled instances, and τ is a temperature parameter that controls the concentration of the distribution. We use the default value of τ provided by [52].

Dual contrastive loss is the combination of the contrastive losses for labeled and unlabeled data, along with a regularization term:

$$\mathcal{L} = \mathcal{L}_L + \lambda \mathcal{L}_U + \beta ||\theta||^2$$

where λ and β are hyperparameters that control the trade-off between the supervised and unsupervised losses, and the regularization term θ represents the model parameters. The values of these parameters we use are of the native implementation in [52].

5. Experimental Results

5.1. Hardware Setup

Experiments were performed on a cloud server with a 2-core Intel Xeon CPU, 16 GB of RAM, and 1 NVIDIA TU104GL GPU. The runtime for every experiment setting (binary or multi-class classification) was less than 10 min.

5.2. Software Setup

All non-neural models were implemented in sklearn [53] python package. Our neural models were implemented with PyTorch [54]. NumPy and Pandas libraries were used for data manipulation. For contrastive learning, we utilized the publicly available Python implementation DualCL [52]. Pre-trained transformer models mlBERT [55] and ruBERT [56] were applied.

5.3. Models and Representations

We applied traditional models denoted by RF, LR, and XGB described in Section 4.4.1. We also used the voting model described in Section 4.4.2 and denoted it by 'voting'. Additionally, we fine-tuned the two transformer models described in Section 4.4.1 and denoted the Russian-language model by RuBERT, and the multi-lingual BERT model by mlBERT. Finally, we also applied the dual contrastive model of [27] and denoted it by DualCL. All of the above models were used for the binary classification task and the multi-class classification task.

For the traditional and voting models, we used the eight representations described in Section 4.3. Transformer models and DualCL were applied to the raw text.

5.4. Metrics

We report the metrics described below for all the models.

Precision measures the accuracy of positive predictions made by the model, and it is computed as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall or sensitivity measures the ability of the model to correctly identify all positive instances. It is computed as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 measure combines precision and recall into a single metric and is computed as

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is the ratio of correctly predicted instances to the total number of instances in the data:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}}$$

When assessing the genre classification results, it is essential to employ all of these metrics because each statistic represents a distinct facet of model performance. The accuracy of positive predictions is the main focus of precision, which is important for situations where false positives can be expensive. In situations where missing a relevant instance is very undesirable, recall evaluates the model's capacity to find all relevant examples and ensures that true positives are not overlooked. Because it combines recall and precision, the F1 score offers a balanced metric that is especially helpful in situations when class distributions are unbalanced.

Although accuracy provides a measure of overall correctness, it can be deceptive in situations where class distributions are not uniform since it may be excessively optimistic [57,58]. Moreover, for the task of genre classification, it is vital to see how a model performs on different classes. The most undesirable output would be assigning all text instances to a single genre, implying that the model does not learn anything except the majority rule. What we seek is a model that learns what the genres are and has moderate to high success in identifying all the genres. Thus, employing all four metrics ensures a comprehensive evaluation.

5.5. Binary Genre Classification Results

This section describes the evaluation of all of the 11 genres in the SONATA dataset with all the models applied to the task of binary genre identification.

5.5.1. Traditional Models

Table 5 shows the results of traditional model evaluation. Because of the large number of setups (11 genres, 8 representations, and 3 models), we show the representation and the classifier that achieved the best result for every one of the 11 genres. We use here the default setting of $N = 100$ samples for every genre and address different values of N in the Appendix A. We can see that all of the obtained accuracies are above the majority, which is 0.5, as we have balanced data samples. We can see a clear difference in the classification accuracy of different genres—the classic literature is easier to detect (with the accuracy of over 0.93), and the short stories genre is the hardest one (with the accuracy of 0.68). This may be because classic literature tends to employ a more formal and elaborate language style compared to short stories. The language in classic literature often includes archaic words, complex sentence structures, and sophisticated vocabulary, while short stories may use simpler language and have a more straightforward narrative style. In terms of text representation, sentence embeddings perform better for the majority of genres but not for all of them, and stylometric features are helpful in some but not in all of the cases. Tf-idf vectors work best for children's and contemporary literature—children's literature typically uses simpler language with shorter sentences, basic vocabulary, and straightforward syntax that can be captured with tf-idf vectors successfully. Contemporary Russian literature may employ innovative narrative techniques, non-linear storytelling, metafiction, and experimental forms that are also expressed in the vocabulary used. The traditional classifier that performs the best for the majority of genres (but not always) is RF.

Table 5. Results of binary genre classification with traditional models for sample size $N = 100$.

Genre	Representation	Classifier	P	R	F1	Acc
action	SE + stylometry	RF	0.8144	0.8084	0.7997	0.8000
adventure	SE + stylometry	LR	0.8548	0.8542	0.8541	0.8542
children's	tfidf + stylometry	RF	0.8486	0.8555	0.8397	0.8400
classic	SE + stylometry	RF	0.9355	0.9130	0.9179	0.9200
contemporary	tfidf + stylometry	RF	0.7192	0.7150	0.7159	0.7200
detective	SE	LR	0.7905	0.7456	0.7453	0.7600
fantasy	char n-grams	XGB	0.7420	0.7432	0.7399	0.7400
non-fiction	SE	LR	0.8800	0.8824	0.8798	0.8800
romance	SE	XGB	0.7603	0.7552	0.7565	0.7600
science-fiction	SE	LR	0.7890	0.7866	0.7799	0.7800
short-stories	SE	RF	0.6899	0.6899	0.6800	0.6800

5.5.2. The Voting Model

Table 6 shows the results of the evaluation for the ensemble voting models. Because of the large number of setups (11 genres and 8 representations), we show the results for the text representation that achieved the best result for every one of the 11 genres. The arrows indicate the increase (↑) or decrease (↓) of classification accuracy in comparison to the best traditional model for that genre. We can see that in all but one genre (non-fiction) the voting model does not outperform the best single traditional classifier. Non-fiction literature can include a wide range of sub-genres, including history, science, biography, memoirs, essays, and more. Each of these sub-genres may have distinct linguistic features that perhaps can be better captured by a voting ensemble. However, for other genres, single models seem to build different classification functions that do not separate between classes in the same way. The texts that fall to opposite "sides" of each separation function "confuse" the ensemble model.

Table 6. Results of binary genre classification with the voting model; grey color indicates the best result.

Genre	Classifier	P	R	F1	Acc	Comparison to the Best Trad Model
action	SE	0.7890	0.7866	0.7799	0.7800	↓
adventure	SE	0.7937	0.7917	0.7913	0.7917	↓
children's	tfidf + stylometry	0.8621	0.8621	0.8400	0.8400	↓
classic	n-grams	0.9200	0.9227	0.9199	0.9200	↓
contemporary	n-grams + stylometry	0.6782	0.6747	0.6753	0.6800	↓
detective	n-grams + stylometry	0.7585	0.7585	0.7585	0.7600	↓
fantasy	SE + stylometry	0.7388	0.7399	0.7391	0.7400	↓
non-fiction	SE + stylometry	0.9010	0.8977	0.8990	0.9000	↑
romance	SE + stylometry	0.7388	0.7399	0.7391	0.7400	↓
science-fiction	SE	0.7734	0.7681	0.7596	0.7600	↓
short-stories	SE	0.6659	0.6672	0.6599	0.6600	↓

5.5.3. Fine-Tuned Transformers

Table 7 shows the results of fine-tuning and testing BERT-based models—mlBERT and ruBERT—for every one of the 11 genres separately. We can see that both models produce results that are much worse than those of traditional models, and in most cases, these results fall below the majority. This outcome might be the result of several factors. First, our training data might be too small for efficient training of LLMs. Second, distinguishing between one specific genre against a mix of multiple genres is a difficult task based on semantics, without any stylistic features.

To our surprise, classification accuracy is higher for ruBERT for several genres only but not for all of them. This may be an indication that a cross-lingual training of mlBERT allows the model to utilize insights from other languages when classifying Russian texts.

Table 7. Results of binary genre classification with fine-tuned BERT models; the grey color indicates the best accuracy.

Genre	mlBERT F1	mlBERT Acc	ruBERT F1	ruBERT Acc
action	0.4762	0.5600	0.4156	0.5000
adventure	0.4045	0.4792	0.4678	0.4792
children's	0.3107	0.4000	0.5833	0.6000
classic	0.4156	0.5000	0.3189	0.3200
contemporary	0.4165	0.5400	0.3151	0.4600
detective	0.3506	0.5400	0.4746	0.5000
fantasy	0.3689	0.4600	0.4058	0.4400
non-fiction	0.4000	0.4000	0.4802	0.6000
romance	0.3151	0.4600	0.3151	0.4600
science-fiction	0.3506	0.5400	0.5833	0.6000
short-stories	0.4283	0.5600	0.3810	0.4800

5.5.4. Dual Contrastive Learning

Table 8 shows the results of binary genre classification with the DualCL model that employs either ruBERT or mlBERT as its base transformed model. We also indicate by the grey color the best accuracy among the two base transformer models.

We can see that while the results are worse than those of traditional models for every genre, there is a significant improvement over the fine-tuned transformer models. It is also evident that for all but one genre (romance), the DualCL model with ruBERT outperforms the same model with mlBERT.

Table 8. Results of binary genre classification for DualCL; the grey color indicates better results.

Model	Genre	ruBERT F1	ruBERT Acc	mlBERT F1	mlBERT Acc
DualCL	action	0.5703	0.6200	0.5331	0.5600
DualCL	adventure	0.5623	0.5625	0.5279	0.5625
DualCL	children's	0.5536	0.5600	0.5484	0.5600
DualCL	classic	0.6716	0.6800	0.4900	0.5800
DualCL	contemporary	0.6486	0.6600	0.3658	0.5000
DualCL	detective	0.6394	0.6400	0.5942	0.6000
DualCL	fantasy	0.6394	0.6400	0.3969	0.5600
DualCL	non-fiction	0.7391	0.7400	0.3867	0.5400
DualCL	romance	0.6162	0.6200	0.5066	0.6400
DualCL	science-fiction	0.5942	0.6000	0.4172	0.6000
DualCL	short-stories	0.5824	0.6200	0.5785	0.5800

5.6. Multi-Class Genre Classification Results

5.6.1. Traditional Models

Table 9 contains the results produced by traditional classifiers RF, LR, and XGB for all the text representations we employ. For every text representation, we report the results of the best model out of three (full results are contained in the Appendix A). Here, we can see a clear advantage of using sentence embeddings enhanced with stylometric features as text representation. The second best result is achieved by the sentence embeddings without stylometric features. These results indicate that capturing semantic information, linguistic patterns, and stylistic characteristics of the text is much more important for Russian genre classification than the vocabulary.

We can also see that the LR classifier achieves the best result for the majority of text representations. This may be because the logistic regression model is less prone to overfitting, especially when the dataset is small.

Table 9. Results of multi-class genre classification for traditional models; grey color indicates the best result.

Representation	Classifier	P	R	F1	Acc
SE	LR	0.4289	0.4332	0.4264	0.4293
SE + stylometry	LR	0.4415	0.4471	0.4386	0.4367
char n-grams	LR	0.2865	0.2650	0.2711	0.2705
char n-grams + stylometry	LR	0.2694	0.2471	0.2550	0.2506
n-grams	LR	0.3004	0.2866	0.2800	0.2754
n-grams + stylometry	RF	0.2962	0.2911	0.2617	0.2878
tfidf	LR	0.2996	0.2990	0.2372	0.2903
tfidf + stylometry	RF	0.2503	0.2617	0.2308	0.2581

Table 10 shows the per-genre precision, recall, and F-measure produced by the best model (LR and SE + stylometry text representation). We can see that the model does attribute all instances to a single class but is producing real predictions for all the genres. Some genres are identified better than others—adventure, contemporary literature, and science fiction. It may be because these genres typically adhere to clear conventions and tropes that provide consistent patterns and signals that a classification model can leverage.

In contrast, genres with more fluid boundaries pose greater challenges for classification due to their variability and ambiguity.

Table 10. Best result details of multi-class genre classification for traditional models (SE + stylometry, LR).

Genre	P	R	F1
action	0.3654	0.3878	0.3762
adventure	0.7353	0.5814	0.6494
children's	0.3519	0.5000	0.4130
classic	0.3478	0.3902	0.3678
contemporary	0.6000	0.7241	0.6562
detective	0.3600	0.3333	0.3462
fantasy	0.3514	0.4062	0.3768
non-fiction	0.4500	0.4500	0.4500
romance	0.4375	0.2593	0.3256
science-fiction	0.5610	0.6571	0.6053
short-stories	0.2963	0.2286	0.2581

5.6.2. The Voting Model

Table 11 shows the evaluation results of the voting ensemble model applied to different text representations. The arrow indicates an increase or decrease in accuracy concerning the results produced by individual traditional models. In general, the voting model does not reach the high scores of the best single traditional models; however, there is an improvement in accuracy for text representations that use word n-grams and character n-grams with stylometric features. N-grams capture local syntactic and semantic information, which might be beneficial for capturing genre-specific patterns, and using multiple classifiers may help to identify the most probable outcome that enhances classification accuracy.

Table 11. Results of multi-class genre classification for the voting model.

Representation	P	R	F1	Acc	vs. Best Trad. Model
SE	0.4076	0.4052	0.3995	0.3970	↓
SE + stylometry	0.3978	0.3977	0.3924	0.3921	↓
char n-grams	0.2779	0.2635	0.2651	0.2655	↓
char n-grams + stylometry	0.2740	0.2595	0.2630	0.2605	↑
n-grams	0.2915	0.2965	0.2799	0.2854	↑
n-grams + stylometry	0.3181	0.3044	0.2923	0.2953	↑
tfidf	0.2286	0.2459	0.2072	0.2382	↓
tfidf + stylometry	0.2935	0.2956	0.2640	0.2878	↓

The best text representation for the voting model is sentence embeddings. The detailed scores for all the genres produced by this model are shown in Table 12, with an arrow indicating a comparison with the F1 scores of the best single traditional model result shown in Table 10. We can see that while the combined score of this voting model is lower than that of its single-model counterpart, there are individual genres such as children's books, classic literature, fantasy, and romance novels that have higher F1 scores. Again, we observe that some genres such as adventure and science fiction are 'easier' to identify in this task.

5.6.3. Fine-Tuned Transformer Models

Table 13 shows the evaluation results of two fine-tuned transformer models, ruBERT and mlBERT, for the task of muli-class genre classification. The scores are low for both of the models, and the per-genre detailed scores show that both models failed to learn and classify all texts as belonging to a single class. There is a slight advantage to the mlBERT model over ruBERT, similar to the binary classification task. We believe that classifying books into one out of multiple genres based purely on their vocabulary and semantics is a very difficult task and something beyond embeddings learned by transformers needs to be

provided to a classification layer. One such thing is the stylistic characteristics of the text provided with stylometric features.

Table 12. Best result details of multi-class genre classification for the voting model (SE + stylometry); grey color indicates improvement.

Genre	P	R	F1	vs. Best Trad. Model
action	0.3000	0.3061	0.3030	↓
adventure	0.7931	0.5349	0.6389	↓
children's	0.3704	0.5263	0.4348	↑
classic	0.3542	0.4146	0.3820	↑
contemporary	0.5556	0.6897	0.6154	↓
detective	0.3200	0.2963	0.3077	↓
fantasy	0.3500	0.4375	0.3889	↑
non-fiction	0.4000	0.3000	0.3429	↓
romance	0.4103	0.2963	0.3441	↑
science-fiction	0.5366	0.6286	0.5789	↓
short-stories	0.2308	0.1714	0.1967	↓

Table 13. Multi-class classification results of ruBERT and mlBERT transformer models; grey color indicates better results.

classifier	P	R	F1	Acc
ruBERT	0.0087	0.0841	0.0158	0.0918
		per genre scores		
genre	P	R	F1	
action	0	0	0	
adventure	0	0	0	
children's	0	0	0	
classic	0	0	0	
contemporary	0	0	0	
detective	0	0	0	
fantasy	0	0	0	
non-fiction	0	0	0	
romance	0	0	0	
science-fiction	0	0	0	
short-stories	0.0956	0.925	0.1733	
classifier	P	R	F1	Acc
mlBERT	0.0091	0.0909	0.0166	0.0993
		per genre scores		
genre	P	R	F1	
action	0	0	0	
adventure	0	0	0	
children's	0	0	0	
classic	0	0	0	
contemporary	0	0	0	
detective	0	0	0	
fantasy	0.1005	1	0.1826	
non-fiction	0	0	0	
romance	0	0	0	
science-fiction	0	0	0	
short-stories	0	0	0	

5.6.4. Dual Contrastive Learning

Table 14 contains the evaluation results for the DualCL model with ruBERT and mlBERT backends. The model was trained for 100 epochs, and the best score that was achieved at epoch 97 is reported in the table, together with the per-genre scores. The scores are much higher than those of fine-tuned transformers, but they do not reach the results produced by the best single traditional model. Again, we see that some genres such as non-fiction and classic literature achieve higher scores than other genres. Similarly to the binary classification results and the fine-tuning results, the model that uses mlBERT as the backend, surprisingly, outperformed ruBERT in terms of the final accuracy. However, for specific genres, ruBERT-based DualCL performs better for 9 out of 11 genres (the grey color in the table indicates the best F1 genre for both variants of the DualCL model).

Table 14. Multi-class classification results for the DualCL model; grey color indicates better results.

classifier	P	R	F1	Acc
DualCL ruBERT	0.3706	0.3400	0.3400	0.3704
per genre scores				
genre	P	R	F1	
action	0.2761	0.3700	0.3162	
adventure	0.1020	0.2083	0.1370	
children's	0.5714	0.2857	0.3810	
classic	0.4497	0.6979	0.5469	
contemporary	0.2039	0.3100	0.2460	
detective	0.3220	0.1900	0.2390	
fantasy	0.4353	0.3700	0.4000	
non-fiction	0.8228	0.6500	0.7263	
romance	0.3763	0.3500	0.3627	
science-fiction	0.4561	0.2600	0.3312	
short-stories	0.0606	0.0476	0.0533	
classifier	P	R	F1	Acc
DualCL mlBERT	0.3582	0.3571	0.3354	0.3758
per genre scores				
genre	P	R	F1	
action	0.3684	0.2100	0.2675	
adventure	0.1400	0.2917	0.1892	
children's	0.3011	0.5000	0.3758	
classic	0.6092	0.5521	0.5792	
contemporary	0.1970	0.2600	0.2241	
detective	0.3409	0.1500	0.2083	
fantasy	0.3286	0.4600	0.3833	
non-fiction	0.7660	0.7200	0.7423	
romance	0.3473	0.5800	0.4345	
science-fiction	0.3750	0.1800	0.2432	
short-stories	0.1667	0.0238	0.0417	

5.7. Punctuation Importance

To verify our assumption about the inherent importance of punctuation for genre classification, we conducted a series of additional experiments. In this experiment, we removed punctuation marks from the texts. This act decreased slightly the sizes of BOW representations (details are provided in the Appendix A). We applied the experimental setup of Section 4.2 to the simplified SONATA dataset and used the same sample size of $N = 100$ book chunks per genre.

The results of the binary genre classification for this setup for traditional classifiers appear in Table 15. We show the best results for every genre and indicate the representation and classifier that achieve them. The arrows indicate an increase or decrease in accuracy in comparison to the best traditional model results shown on texts with punctuation. We see that in most cases, the results are inferior to those of Section 5.5.1, and stylometric features have a less prominent role. However, for 3 genres out of 11, the results are improved—fantasy, romance, and non-fiction. In these genres, the content words play a crucial role in conveying the theme and style of the text. It is possible that by removing punctuation, these content words are emphasized and can become more indicative of genre characteristics.

Table 15. Binary classification with traditional classifiers with sample size $N = 100$.

Genre	Representation	Classifier	P	R	F1	Acc	vs. Best Punct Acc
action	SE	LR	0.6987	0.6997	0.6989	0.7000	↓
adventure	SE	RF	0.7500	0.7500	0.7500	0.7500	↓
children's	SE	LR	0.8346	0.7989	0.8069	0.8200	↓
classic	char n-grams	XGB	0.5455	0.5197	0.4673	0.5800	↓
contemporary	SE + stylometry	RF	0.7167	0.7093	0.6989	0.7000	↓
detective	char n-grams	LR	0.6575	0.6562	0.6566	0.6600	↓
fantasy	SE + stylometry	RF	0.7788	0.7802	0.7792	0.7800	↑
non-fiction	tfidf + stylometry	RF	0.9600	0.9630	0.9599	0.9600	↑
romance	SE	LR	0.8622	0.8639	0.8599	0.8600	↑
science-fiction	SE + stylometry	LR	0.6502	0.6473	0.6394	0.6400	↓
short-stories	tfidf + stylometry	LR	0.6619	0.6430	0.6380	0.6531	↓

The results of multi-class genre classification for this setup appear in Table 16. We applied single traditional models because they produced the best scores on the original SONATA dataset. In a few cases, an improvement was achieved; however, none of these setups outperformed the best model-representation combo in the original setting that uses the data with punctuation.

Table 16. Multi-class classification with traditional classifiers with sample size $N = 100$, with and without punctuation; grey color indicates improvement.

Representation	Classifier	Acc with Punctuation	Acc without Punctuation
SE	RF	0.3375	0.2926
SE	LR	0.4293	0.3232
SE	XGB	0.2978	0.2850
SE + stylometry	RF	0.3400	0.3181
SE + stylometry	LR	0.4367	0.3282
SE + stylometry	XGB	0.3127	0.2799
char n-grams	RF	0.2333	0.2214
char n-grams	LR	0.2705	0.2366
char n-grams	XGB	0.2432	0.1934
char n-grams + stylometry	RF	0.2531	0.2087
char n-grams + stylometry	LR	0.2506	0.2316
char n-grams + stylometry	XGB	0.2382	0.1985
n-grams	RF	0.2333	0.2341
n-grams	LR	0.2754	0.2748
n-grams	XGB	0.1712	0.1501
n-grams + stylometry	RF	0.2878	0.2316
n-grams + stylometry	LR	0.2779	0.2621
n-grams + stylometry	XGB	0.2233	0.1985
tfidf	RF	0.2432	0.2341
tfidf	LR	0.2903	0.2545
tfidf	XGB	0.1439	0.1908
tfidf + stylometry	RF	0.2804	0.2545
tfidf + stylometry	LR	0.2531	0.2545
tfidf + stylometry	XGB	0.2134	0.1832

To further verify our hypothesis about the importance of punctuation, we also tested the ensemble voting model for the multi-class genre classification; the results are shown in Table 17. No model–representation setup achieved an increase in accuracy over the non-modified dataset. Therefore, we can conclude that preserving punctuation is vital for genre classification.

Table 17. Multi-class classification with the voting model, sample size $N = 100$, with and without punctuation (grey color indicates better result in each case).

Representation	Acc with Punctuation	Acc without Punctuation
SE	0.3970	0.3282
SE + stylometry	0.3921	0.3435
char n-grams	0.2655	0.2392
char n-grams + stylometry	0.2605	0.2494
n-grams	0.2854	0.2468
n-grams + stylometry	0.2953	0.2875
tfidf	0.2382	0.2265
tfidf + stylometry	0.2878	0.2723

6. Conclusions

In this study, we studied the task of genre classification for Russian literature. By introducing a novel dataset comprising Russian books spanning 11 different genres, we facilitate research in genre classification for Russian texts and evaluate multiple classification models to discern their effectiveness in this context. Through extensive experimentation, we explored the impact of different text representations, such as stylometric features, on genre classification accuracy.

Our experimental evaluation confirms that stylometric features improve classification accuracy in most cases, especially in binary genre classification. Therefore, RQ1 is also answered positively. Our evaluation also shows that while there are genres that receive higher accuracy scores, the results depend more on the model being used than on the features. Thus, some genres are 'easier' for traditional models, while other genres are 'easier' for contrastive learning, and so on. It means that RQ2 cannot be answered positively. We have also verified that contrastive learning performs much better than transformer models for both classification tasks, answering RQ3. Finally, we have shown that removing punctuation decreases classification accuracy and thus answered positively to RQ4. We have also found, surprisingly, that the ruBERT model pre-trained on a large Russian corpus performs worse than the multi-lingual BERT model for the multi-class classification task. For the binary classification, ruBERT performs worse than multi-lingual BERT on 8 out of 11 genres, answering negatively to RQ5.

Our study highlights the multi-faceted nature of genre classification in Russian literature and underscores the importance of considering diverse factors, including linguistic characteristics, cultural nuances, and genre-specific features. An accurate model of genre classification is capable of performing literary analysis by automating the identification and categorization of texts. This can help researchers better understand how genres in Russian literature have evolved with social and political changes in Russian-speaking nations by, for example, enabling them to identify trends and changes in literary styles across time. Furthermore, by examining relatively small (300-word) text samples rather than the whole texts, our classification models can be utilized to quickly search and retrieve books by genre given the extensive holdings of Russian literary works in digital libraries and archives.

While our research provides valuable insights into genre classification for Russian texts, it also reveals limitations and areas for future exploration.

7. Limitations and Future Research Directions

While our study provides valuable insights into genre classification for Russian literature, several limitations are worth mentioning. Firstly, the dataset we compiled, though extensive, may not encompass the entirety of Russian literary genres, potentially limiting the generalizability of our findings to all facets of Russian literature.

Furthermore, our evaluation of classification models is based on a subset of the dataset, which may not fully capture the diversity and complexity of genre classification for Russian books. The selection of this subset could introduce sampling bias and impact the robustness of our conclusions.

Another limitation relates to the choice of text representations and classification models evaluated in our study. While we explored a range of traditional machine learning models and state-of-the-art deep learning architectures, there may exist alternative approaches or models that could yield superior performance in genre classification tasks. Moreover, the effectiveness of stylometric features and other text representations may vary across different genres, and our study may not fully capture this variability.

Finally, the research questions addressed in our study provide valuable insights into the genre classification of Russian books; however, they represent only a subset of the broader landscape of research questions in this domain. Future studies may explore additional research questions, such as the impact of personal authorial style on genre classification, the role of narrative structure in genre categorization, transfer learning from other languages, or the influence of reader preferences on genre classification outcomes.

Addressing these limitations through further research could enhance the robustness and applicability of genre classification systems for Russian texts.

Author Contributions: Conceptualization, N.V. and M.L.; methodology, N.V. and M.T.; software, N.V. and G.K.; validation, N.V., M.T. and M.L.; formal analysis, N.V. and M.L.; investigation, M.T.; resources, G.K.; data curation, M.T.; writing—original draft preparation, N.V. and M.T.; writing—review and editing, N.V., M.L. and M.T.; supervision, N.V. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SONATA dataset is residing in repository on GitHub at https://github.com/genakogan/Identification-of-the-genre-of-books. It is freely available to the NLP community.

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

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
RQ	Research Question
RF	Random Forest
XGB	eXtreme Gradient Boost
LR	Logistic Regression
CL	Contrastive Learning
RNN	Recurrent Neural Network
SVM	Support Vector Machine
BERT	Bidirectional Encoder Representations from Transformers
R	Recall
P	Precision
F1	F1 measure

Appendix A

Appendix A.1. The List of Stylometric Features

We use the features provided by the StyloMetrix package [49]. Tables A1–A3 contains the lists of these features according. Unless specified otherwise, the incidence (amount) of appearances of a feature is reported.

Table A1. Lexical features provided by the StyloMetric tool.

Lexical Features
type-token ratio for words lemmas, content words, function words, content words types, function words types, nouns in plural, nouns in singular, proper names, personal names, animate nouns, inanimate nouns, neutral nouns, feminine nouns, masculine nouns, feminine proper nouns, masculine proper nouns, surnames, given names, flat multiwords expressions, direct objects, indirect objects, nouns in Nominative case, nouns in Genitive case, nouns in Dative case, nouns in Accusative case, nouns in Instrumental case, nouns in Locative case, qualitative adj positive, relative adj, qualitative comparative adj, qualitative superlative adj, direct adjective, indirect adjective, punctuation, dots, comma, semicolon, colon, dashes, numerals, relative pronouns, indexical pronouns, reflexive pronoun, posessive pronoun, negative pronoun, positive adverbs, comparative adverbs, superlative adverbs

Table A2. Part-of-speech features provided by the StyloMetric tool.

Part-of-Speech Features
verbs, nouns, adjectives, adverbs, determiners, interjections, conjunctions, particles, numerals, prepositions, pronouns, code-switching, number of words in narrative sentences, number of words in negative sentences, number of words in parataxis sentences, number of words in sentences that do not have any root verbs, words in sentences with quotation marks, number of words in exclamatory sentences, number of words in interrogative sentences, number of words in general questions, number of words in special questions, number of words in alternative questions, number of words in tag questions, number of words in elliptic sentences, number of positionings, number of words in conditional sentences, number of words in imperative sentences, number of words in amplified sentences

Table A3. Grammar features provided by the StyloMetric tool.

Grammar Features
root verbs in imperfect aspect, all verbs in imperfect aspect, active voice, root verbs in perfect form, all verbs in perfect form, verbs in the present tense, indicative mood, imperfect aspect, verbs in the past tense, indicative mood, imperfect aspect, verbs in the past tense, indicative mood, perfect aspect, verbs in the future tense, indicative mood, perfect aspect, verbs in the future tense, indicative mood, imperfect aspect, simple verb forms, verbs in the future tense, indicative mood, complex verb forms, verbs in infinitive, verbs in the passive form, transitive verbs, intransitive verbs, impersonal verbs, passive participles, active participles, adverbial perfect participles, adverbial imperfect participles

Appendix A.2. The List of Russian Stopwords

Below, we show the list of Russian stopwords produced that we used and their translation.

Stopwords (Russian)
и, в, во, не, что, он, на, я, с, со, как, а, то, все, она, так, его, но, да, ты, к, у, же, вы, за, бы, по, только, ее, мне, было, вот, от, меня, еще, нет, о, из, ему, теперь, когда, даже, ну, вдруг, ли, если, уже, или, ни, быть, был, него, до, вас, нибудь, опять, уж, вам, ведь, там, потом, себя, ничего, ей, может, они, тут, где, есть, надо, ней, для, мы, тебя, их, чем, была, сам, чтоб, без, будто, чего, раз, тоже, себе, под, будет, ж, тогда, кто, этот, того, потому, этого, какой, совсем, ним, здесь, этом, один, почти, мой, тем, чтобы, нее, сейчас, были, куда, зачем, всех, никогда, можно, при, наконец, два, об, другой, хоть, после, над, больше, тот, через, эти, нас, про, всего, них, какая, много, разве, три, эту, моя, впрочем, хорошо, свою, этой, перед, иногда, лучше, чуть, том, нельзя, такой, им, более, всегда, конечно, всю, между
Translation
and, in, in the, not, that, he, on, I, with, with, like, and, then, all, she, so, his, but, yes, you, to, at, already, you (plural), behind, would, by, only, her, to me, was, here, from, me, yet, no, about, to him, now, when, even, well, suddenly, whether, if, already, or, neither, to be, was, him, before, to you, ever, again, already, you (plural), after all, there, then, oneself, nothing, to her, can, they, here, where, there is, need, her, for, we, you (singular), them, than, was, oneself, without, as if, of what, time, also, to oneself, under, will be, what, then, who, this, of that, therefore, of this, what kind, completely, him, here, in this, one, almost, my, by, her, now, were, where, why, all, never, can, at, finally, two, about, other, even if, after, above, more, that, through, these, us, about, all, what kind of, many, whether, three, this, my, however, well, her own, this, before, sometimes, better, a bit, that, cannot, such, to them, more, always, of course, whole, between

Appendix A.3. BOW Vector Statistics for the No-Punctuation SONATA Data Sample

The sizes of BOW vectors for the SONATA dataset sample with $N = 100$ are provided in Table A4.

Table A4. BOW vector lengths for samples of size $N = 100$ without punctuation.

Genre	**Char n-Grams Vector Sizes** $n = [1, 2, 3]$	**Word n-Grams Vector Sizes** $n = [1, 2, 3]$	**tf-idf Vector Size**
action	10,593	35,521	17,923
adventure	9085	18,236	11,506
children's	11,647	27,262	14,581
classic	11,907	27,456	15,905
contemporary	11,844	35,396	19,132
detective	11,750	34,963	18,532
fantasy	9334	35,635	18,044
non-fiction	13,900	34,900	17,867
romance	10,867	33,398	16,553
science-fiction	9590	35,824	18,799
short-stories	9298	23,673	13,442
all	23,902	315,812	87,706

Appendix A.4. Changing the Number of Samples

Because of HW limitations, we were unable to run all the experiments on full data, and we used the sampling procedure described in Section 4.2 with the default value of $N = 100$ randomly sampled book chunks per genre. However, we examined the setups with fewer and more sampled chunks ($N = 50$ and $N = 150$). Tables A5 and A6 show the sizes of BOW vectors for both of these setups.

Table A5. BOW vector lengths for samples of size $N = 50$.

Genre	Char n-Grams Vector Sizes $n = [1, 2, 3]$	Word n-Grams Vector Sizes $n = [1, 2, 3]$	tf-idf Vector Size
action	12,104	23,470	13,037
adventure	11,845	18,024	11,108
children's	11,803	23,115	12,414
classic	13,332	23,692	14,159
contemporary	12,956	23,747	14,167
detective	11,680	24,236	13,564
fantasy	10,910	24,235	13,639
non-fiction	14,435	24,656	13,676
romance	10,852	22,598	12,048
science-fiction	11,866	24,234	14,016
short-stories	12,853	20,803	11,874
all	26,192	222,487	69,362

Table A6. BOW vector lengths for samples of size $N = 200$.

Genre	Char n-Grams Vector Sizes $n = [1, 2, 3]$	Word n-Grams Vector Sizes $n = [1, 2, 3]$	tf-idf Vector Size
action	16,721	77,373	31,126
adventure	13,319	29,283	16,241
children's	15,911	58,789	24,921
classic	18,091	62,025	29,327
contemporary	19,068	84,345	36,132
detective	16,964	93,493	35,765
fantasy	15,759	95,349	37,162
non-fiction	23,288	90,602	35,625
romance	15,637	88,178	32,476
science-fiction	17,351	95,400	39,043
short-stories	16,684	46,214	21,788
all	37,228	687,100	139,054

Table A7 shows evaluation results of traditional models for both of these sampling setups and their comparison to the default sampling setup of $N = 100$ for the task of multi-class genre classification. The arrows show a decrease or increase in classification accuracy of the sampling setup $N = 100$ as compared to $N = 50$, and the sampling setup $N = 150$ as compared to $N = 100$. We observe that, with some exceptions, increasing sample size does improve the classification accuracy of traditional models across representations. However, it is worth mentioning that increasing the sample size makes running advanced transformer models in our HW setup much slower and in some cases, impossible.

Table A7. Comparing smaller and larger sample sizes for multi-class genre classification with traditional models; grey color indicates the best result.

Representation	Classifier	Sample Size $N = 50$ Acc	Sample Size $N = 100$ Acc	Sample Size $N = 150$ Acc
SE	RF	0.3645	0.3375 ↓	0.3462 ↑
SE	LR	0.3785	0.4293 ↑	0.4095 ↓
SE	XGB	0.2617	0.2978 ↑	0.3163 ↑
SE + stylometry	RF	0.3458	0.3400 ↓	0.3374 ↓
SE + stylometry	LR	0.3738	0.4367 ↑	0.4130 ↓
SE + stylometry	XGB	0.2710	0.3127 ↑	0.3603 ↑
char n-grams	RF	0.2710	0.2333 ↓	0.2882 ↑
char n-grams	LR	0.2477	0.2705 ↑	0.2953 ↑
char n-grams	XGB	0.1729	0.2432 ↑	0.3005 ↑
char n-grams + stylometry	RF	0.2336	0.2531 ↑	0.2882 ↑
char n-grams + stylometry	LR	0.2477	0.2506 ↑	0.3146 ↑
char n-grams + stylometry	XGB	0.2290	0.2382 ↑	0.3040 ↑
n-grams	RF	0.2570	0.2333 ↓	0.2794 ↑
n-grams	LR	0.3084	0.2754 ↓	0.2988 ↑
n-grams	XGB	0.1168	0.1712 ↑	0.2320 ↑
n-grams + stylometry	RF	0.2757	0.2878 ↑	0.3076 ↑
n-grams + stylometry	LR	0.3037	0.2779 ↓	0.2917 ↑
n-grams + stylometry	XGB	0.2523	0.2233 ↓	0.2812 ↑
tfidf	RF	0.2383	0.2432 ↑	0.2865 ↑
tfidf	LR	0.2991	0.2903 ↓	0.3409 ↑
tfidf	XGB	0.0981	0.1439 ↑	0.2021 ↑
tfidf + stylometry	RF	0.2523	0.2804 ↑	0.2917 ↑
tfidf + stylometry	LR	0.2430	0.2531 ↑	0.3093 ↑
tfidf + stylometry	XGB	0.2664	0.2134 ↓	0.2654 ↑

Appendix A.5. Full Results of Traditional Models for the Multi-Class Genre Classification Task

Table A8 contains the full list of results for the multi-class classification task performed with traditional models.

Table A8. Full results of multi-class genre classification for traditional models; grey color indicates the best result.

Representation	Classifier	P	R	F1	Acc
SE	RF	0.3221	0.3310	0.3113	0.3275
SE	LR	0.4289	0.4332	0.4264	0.4293
SE	XGB	0.3154	0.3019	0.2997	0.3027
SE + stylometry	RF	0.3361	0.3182	0.3003	0.3176
SE + stylometry	LR	0.4415	0.4471	0.4386	0.4367
SE + stylometry	XGB	0.3082	0.3075	0.2961	0.2978
char n-grams	RF	0.2163	0.2315	0.2034	0.2333
char n-grams	LR	0.2865	0.2650	0.2711	0.2705
char n-grams	XGB	0.2180	0.2373	0.2188	0.2357
char n-grams + stylometry	RF	0.2150	0.2436	0.2080	0.2407
char n-grams + stylometry	LR	0.2694	0.2471	0.2550	0.2506
char n-grams + stylometry	XGB	0.2444	0.2480	0.2357	0.2457
n-grams	RF	0.2055	0.2494	0.2062	0.2333
n-grams	LR	0.3004	0.2866	0.2800	0.2754
n-grams	XGB	0.2011	0.1771	0.1600	0.1712
n-grams + stylometry	RF	0.2962	0.2911	0.2617	0.2878
n-grams + stylometry	LR	0.2956	0.2868	0.2784	0.2779
n-grams + stylometry	XGB	0.2373	0.2349	0.2190	0.2233
tfidf	RF	0.2234	0.2332	0.1939	0.2283
tfidf	LR	0.2996	0.2990	0.2372	0.2903
tfidf	XGB	0.2071	0.1964	0.1794	0.1911
tfidf + stylometry	RF	0.2503	0.2617	0.2308	0.2581
tfidf + stylometry	LR	0.2325	0.2549	0.2190	0.2531
tfidf + stylometry	XGB	0.2187	0.2289	0.1975	0.2233

Appendix A.6. The Python Script Used to Download Books from knigogo.net

The script is freely available at https://github.com/genakogan/Identification-of-the-genre-of-books. Due to size limitations, we do not provide it here in full. Its main parts are described below.

- The `url_download_for_each_book` function takes a URL (in this case, a page with links to free books) and retrieves the HTML content. It then parses the HTML to extract URLs that link to book download pages, specifically those that match a pattern (starts with https://knigogo.net/knigi/ (accessed on 1 January 2024) and end with `/#lib_book_download`).
- The `url_text_download_for_each_book` function takes the list of book download URLs obtained in the previous step and retrieves the HTML content of each page. It then parses these pages to extract URLs of the actual text files.
- The `download_url` function attempts to download the content of a given URL and returns the content if successful.
- The `download_book` function receives a text file URL, a book ID, and a save path. It downloads the text file's content and saves it locally as a .txt file in the specified directory.

Appendix A.7. Validation on Texts from a Different Source

To ensure that our best models for genre classification are suitable for texts from different sources, we conducted an additional experiment.

We manually downloaded an additional 200 texts from the https://www.rulit.me (accessed on 15 May 2024) online book repository, and performed the pre-processing described in Section 3.2. We chose the Russian language, .txt format, and the genre of science fiction, which constitutes the most texts in the repository. The positive samples were chosen to be 100 texts belonging to the science fiction genre, and the negative samples were another 100 texts that belong to other genres such as detectives, children's books, romance, documentaries, adventure, thrillers, ancient literature, esoterics, home economics, science, and culture. Then, we split all the texts into 300-word chunks and preserved one chunk from each book.

In the next step, we trained our best-performing models on the training part of the SONATA dataset for the science fiction genre and evaluated them on the rulit.me data. The results of this binary cross-source genre classification are shown in Tables A9 and A10. In Table A9, we can see that traditional models achieve good results on the new data, and the best classifier is RF, which is consistent with our findings in Section 5.5.1. It is also evident that adding stylometric features is beneficial in this case, and the best text representation is n-grams combined with stylometry. Table A10 contains the results of the voting model evaluation and shows that the best representation in this setup is the same as above. However, similarly to the SONATA dataset tests, the voting models, while producing decent results, fail to achieve the same accuracy as the best traditional model in Table A9.

Table A9. Evaluation results for traditional models on the rulit.me data (accessed on 15 May 2024); grey color indicates best results.

Genre	Representation	Classifier	F1	Acc
science-fiction	SE	RF	0.6900	0.6900
science-fiction	SE	LR	0.7097	0.7100
science-fiction	SE	XGB	0.6387	0.6400
science-fiction	SE + stylometry	RF	0.7698	0.7700
science-fiction	SE + stylometry	LR	0.6995	0.7000
science-fiction	SE + stylometry	XGB	0.6297	0.6300
science-fiction	char n-grams	RF	0.6394	0.6400
science-fiction	char n-grams	LR	0.6808	0.6900
science-fiction	char n-grams	XGB	0.6673	0.6700
science-fiction	char n-grams + stylometry	RF	0.7796	0.7800

Table A9. *Cont.*

Genre	Representation	Classifier	F1	Acc
science-fiction	char n-grams + stylometry	LR	0.6784	0.6900
science-fiction	char n-grams + stylometry	XGB	0.5900	0.5900
science-fiction	n-grams	RF	0.6970	0.7000
science-fiction	n-grams	LR	0.7100	0.7100
science-fiction	n-grams	XGB	0.6052	0.6100
science-fiction	n-grams + stylometry	RF	0.7300	0.7300
science-fiction	n-grams + stylometry	LR	0.7100	0.7100
science-fiction	n-grams + stylometry	XGB	0.6096	0.6100
science-fiction	tfidf	RF	0.6532	0.6600
science-fiction	tfidf	LR	0.6800	0.6800
science-fiction	tfidf	XGB	0.5512	0.5600
science-fiction	tfidf + stylometry	RF	0.7093	0.7100
science-fiction	tfidf + stylometry	LR	0.6255	0.6300
science-fiction	tfidf + stylometry	XGB	0.6394	0.6400

Table A10. Evaluation results for the voting model on the rulit.me data (accessed on 15 May 2024); the grey color indicates improvement over individual models.

Genre	Representation	F1	Acc
science-fiction	SE	0.6800	0.6800
science-fiction	SE + stylometry	0.6999	0.7000
science-fiction	char n-grams	0.6862	0.6900
science-fiction	char n-grams + stylometry	0.7383	0.7400
science-fiction	n-grams	0.6768	0.6800
science-fiction	n-grams + stylometry	0.6995	0.7000
science-fiction	tfidf	0.6305	0.6400
science-fiction	tfidf + stylometry	0.6753	0.6800

References

1. Kochetova, L.; Popov, V. Research of Axiological Dominants in Press Release Genre based on Automatic Extraction of Key Words from Corpus. *Nauchnyi Dialog*. **2019**, *6*, 32–49. [CrossRef]
2. Lagutina, K.V. Classification of Russian texts by genres based on modern embeddings and rhythm. *Model. I Anal. Informatsionnykh Sist.* **2022**, *29*, 334–347. [CrossRef]
3. Houssein, E.H.; Ibrahem, N.; Zaki, A.M.; Sayed, A. Semantic protocol and resource description framework query language: A comprehensive review. *Mathematics* **2022**, *10*, 3203. [CrossRef]
4. 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 Int.* **2020**, *13*, 3. [CrossRef]
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 Int.* **2021**, *14*, 4. [CrossRef]
6. Embarcadero-Ruiz, D.; Gómez-Adorno, H.; Embarcadero-Ruiz, A.; Sierra, G. Graph-based siamese network for authorship verification. *Mathematics* **2022**, *10*, 277. [CrossRef]
7. Kessler, B.; Nunberg, G.; Schütze, H. Automatic detection of text genre. *arXiv* **1997**, arXiv:cmp-lg/9707002.
8. Russian language—Wikipedia, The Free Encyclopedia. 2024. Available online: https://en.wikipedia.org/wiki/Russian_language (accessed on 16 May 2024).
9. Shavrina, T. Differential Approach to Webcorpus Construction. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2018*; National Research University Higher School of Economics: Moscow, Russia, 2018.
10. VKontakte. 2024. Available online: https://vk.com (accessed on 1 January 2024).
11. OpenCorpora. 2024. Available online: http://opencorpora.org (accessed on 1 January 2024).
12. Barakhnin, V.; Kozhemyakina, O.; Pastushkov, I. Automated determination of the type of genre and stylistic coloring of Russian texts. In *ITM Web of Conferences*; EDP Sciences: Les Ulis, France, 2017; Volume 10, p. 02001.
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
14. Sun, H.; Liu, J.; Zhang, J. A survey of contrastive learning in NLP. In Proceedings of the 7th International Symposium on Advances in Electrical, Electronics, and Computer Engineering, Xishuangbanna, China, 18–20 March 2022; Volume 12294, pp. 1073–1078.
15. Bulygin, M.; Sharoff, S. Using machine translation for automatic genre classification in Arabic. In Proceedings of the Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, Moscow, Russia, 30 May–2 June 2018; pp. 153–162.

16. Nolazco-Flores, J.A.; Guerrero-Galván, A.V.; Del-Valle-Soto, C.; Garcia-Perera, L.P. Genre Classification of Books on Spanish. *IEEE Access* **2023**, *11*, 132878–132892. [CrossRef]
17. Ozsarfati, E.; Sahin, E.; Saul, C.J.; Yilmaz, A. Book genre classification based on titles with comparative machine learning algorithms. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 14–20.
18. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
19. Saraswat, M.; Srishti. Leveraging genre classification with RNN for Book recommendation. *Int. J. Inf. Technol.* **2022**, *14*, 3751–3756. [CrossRef]
20. Webster, R.; Fonteyne, M.; Tezcan, A.; Macken, L.; Daems, J. Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. *Informatics* **2020**, *7*, 32. [CrossRef]
21. Alfraidi, T.; Abdeen, M.A.; Yatimi, A.; Alluhaibi, R.; Al-Thubaity, A. The Saudi novel corpus: Design and compilation. *Appl. Sci.* **2022**, *12*, 6648. [CrossRef]
22. Mendhakar, A. Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts. *Information* **2022**, *13*, 357. [CrossRef]
23. Williamson, G.; Cao, A.; Chen, Y.; Ji, Y.; Xu, L.; Choi, J.D. Exploring a Multi-Layered Cross-Genre Corpus of Document-Level Semantic Relations. *Information* **2023**, *14*, 431. [CrossRef]
24. Shavrina, T. Genre Classification on Text-Internal Features: A Corpus Study. In Proceedings of the Web Corpora as a Language Training Tool Conference (ARANEA 2018), Univerzita Komenského v Bratislave, Bratislava, Slovakia, 23–24 November 2018; pp. 134–147.
25. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
26. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]
27. Chen, Q.; Zhang, R.; Zheng, Y.; Mao, Y. Dual Contrastive Learning: Text Classification via Label-Aware Data Augmentation. *arXiv* **2022**, arXiv:2201.08702.
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Wright, R.E. Logistic Regression. In *Reading and Understanding Multivariate Statistics*; Grimm, L.G., Yarnold, P.R., Eds.; American Psychological Association: Worcester, MA, USA, 1995; pp. 217–244.
30. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd aCm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
31. Neal, T.; Sundararajan, K.; Fatima, A.; Yan, Y.; Xiang, Y.; Woodard, D. Surveying stylometry techniques and applications. *ACM Comput. Surv.* **2017**, *50*, 86. [CrossRef]
32. Lagutina, K.; Lagutina, N.; Boychuk, E.; Vorontsova, I.; Shliakhtina, E.; Belyaeva, O.; Paramonov, I.; Demidov, P. A survey on stylometric text features. In Proceedings of the 2019 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 5–8 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 184–195.
33. Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. Automatic text categorization in terms of genre and author. *Comput. Linguist.* **2000**, *26*, 471–495. [CrossRef]
34. Sarawgi, R.; Gajulapalli, K.; Choi, Y. Gender attribution: Tracing stylometric evidence beyond topic and genre. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland, OR, USA 23–24 June 2011; pp. 78–86.
35. Eder, M. Rolling stylometry. *Digit. Scholarsh. Humanit.* **2016**, *31*, 457–469. [CrossRef]
36. Eder, M.; Rybicki, J.; Kestemont, M. Stylometry with R: A package for computational text analysis. *R J.* **2016**, *8*, 107–121. [CrossRef]
37. Maciej, P.; Tomasz, W.; Maciej, E. Open stylometric system WebSty: Integrated language processing, analysis and visualisation. *CMST* **2018**, *24*, 43–58.
38. McNamara, D.S.; Graesser, A.C.; McCarthy, P.M.; Cai, Z. Cohesive Features in Expository Texts: A Large-scale Study of Expert and Novice Writing. *Writ. Commun.* **2014**, *31*, 151–183.
39. Okulska, I.; Stetsenko, D.; Kołos, A.; Karlińska, A.; Głąbińska, K.; Nowakowski, A. StyloMetrix: An Open-Source Multilingual Tool for Representing Stylometric Vectors. *arXiv* **2023**, arXiv:2309.12810.
40. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.* **2021**, *54*, 62. [CrossRef]
41. Cunha, W.; Viegas, F.; França, C.; Rosa, T.; Rocha, L.; Gonçalves, M.A. A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification. *ACM Comput. Surv.* **2023**, *55*, 265. [CrossRef]
42. Face, H. Hugging Face. 2016. Available online: https://huggingface.co/ (accessed on 26 April 2024).
43. Kuratov, Y.; Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv* **2019**, arXiv:1905.07213.
44. LitRes. LitRes: Digital Library and E-Book Retailer. 2024. Available online: https://www.litres.ru (accessed on 1 January 2024).
45. Royallib: Free Online Library. 2024. Available online: https://royallib.com/ (accessed on 1 January 2024).
46. knigogo.net. Knigogo. 2013. Available online: https://knigogo.net/zhanryi/ (accessed on 1 January 2024).
47. Belkina, A.C.; Ciccolella, C.O.; Anno, R.; Halpert, R.; Spidlen, J.; Snyder-Cappione, J.E. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat. Commun.* **2019**, *10*, 5415. [CrossRef]

48. Bird, S.; Loper, E.; Klein, E. NLTK: The Natural Language Toolkit. *arXiv* **2009**, arXiv:cs/0205028. [CrossRef]
49. ZILiAT-NASK. StyloMetrix: An Open-Source Multilingual Tool for Representing Stylometric Vectors (Code Repository). 2023. Available online: https://github.com/ZILiAT-NASK/StyloMetrix (accessed on 26 April 2024).
50. Okulska, I.; Stetsenko, D.; Kołos, A.; Karlińska, A.; Głąbińska, K.; Nowakowski, A. StyloMetrix Metrics List (Russian). 2023. Available online: https://github.com/ZILiAT-NASK/StyloMetrix/blob/main/resources/metrics_list_ru.md (accessed on 26 April 2024).
51. Schapire, R.E. Improving Regressors using Boosting Techniques. In Proceedings of the International Conference on Machine Learning (ICML), Austin, TX, USA, 21–23 June 1990.
52. Hiyouga. Dual Contrastive Learning. 2022. Available online: https://github.com/hiyouga/Dual-Contrastive-Learning (accessed on 26 March 2024).
53. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
54. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
55. Google Research. BERT: Multilingual (Uncased). 2018. Available online: https://huggingface.co/google-bert/bert-base-multilingual-uncased (accessed on 26 April 2024).
56. DeepPavlov. RuBERT: Russian (Cased). 2021. Available online: https://huggingface.co/DeepPavlov/rubert-base-cased (accessed on 26 April 2024).
57. Makridakis, S. Accuracy measures: Theoretical and practical concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [CrossRef]
58. Streiner, D.L.; Norman, G.R. "Precision" and "accuracy": Two terms that are neither. *J. Clin. Epidemiol.* **2006**, *59*, 327–330. [CrossRef]