https://doi.org/10.17323/jle.2024.24030

A BERT-BASED CLASSIFICATION MODEL: THE CASE OF RUSSIAN FAIRY TALES

Valery Solovyev [1], **Marina Solnyshkina** [1], **Andrey Ten** [2], **Nikolai Prokopyev** [3]

[1] Kazan Federal University, Kazan, Russia

[2] Nobilis.Team, Kazan, Russia

[3] TAS Institute of Applied Semiotics, Kazan, Russia

ABSTRACT

Introduction: Automatic profiling and genre classification are crucial for text suitability assessment and as such have been in high demand in education, information retrieval, sentiment analysis, and machine translation for over a decade. Of all kinds of genres, fairy tales make one of the most challenging and valuable objects of study due to its heterogeneity and a wide range of implicit idiosyncrasies. Traditional classification methods including stylometric and parametric algorithms, however, are not only labour-intensive and time-consuming, but they are also struggling with identifying corresponding classifying discriminants. The research in the area is scarce, their findings are still controversial and debatable.

Purpose: To fill this crucial void and offers an algorithm to range Russian fairy-tales into classes based on the pre-set parameters. We present the latest BERT-based classification model for Russian fairy tales, test the hypothesis of BERT potential for classifying Russian texts and verify it on a representative corpus of 743 Russian fairy tales.

Method: We pre-train BERT using a collection of three classes of documents and fine-tune it for implementation of a specific application task. Focused on the mechanism of tokenization and embeddings design as the key components in BERT's text processing, the research also evaluates the standard benchmarks used to train classification models and analyze complex cases, possible errors and improvement algorithms thus raising the classification models accuracy. Evaluation of the models performance is conducted based on the loss function, prediction accuracy, precision and recall.

Results: We validated BERT's potential for Russian text classification and ability to enhance the performance and quality of the existing NLP models. Our experiments with cointegrated/rubert-tiny, ai forever/ruBert-base, and DeepPavlov/rubert-base-cased-sentence on different classification tasks demonstrate that our models achieve state-of-the-art results with the best accuracy of 95.9% in cointegrated/rubert-tiny thus outperforming the other two models by a good margin. Thus, the achieved by AI classification accuracy is so high that it can compete with that of human expertise.

Conclusion: The findings highlight the importance of fine-tuning for classifying models. BERT demonstrates great potential for improving NLP technologies and contributing to the quality of automatic text analysis and offering new opportunities for research and application in a wide range of areas including identification and arrangement of all types of content-relevant texts thus contributing to decision making. The designed and validated algorithm can be scaled for classification of as complex and ambiguous discourse as fiction thus improving our understanding of text specific categories. Considerably bigger datasets are required for these purposes.

KEYWORDS:

Machine learning, Bert model, fairy tales, Text classification, Neural networks

Citation: Solovyev, V., Solnyshkina, M., Ten, A., & Prokopyev, N. (2024). A BERT-Based classification model: The case of Russian fairy tales. *Journal of Language and Education, 10*(4), 98-111. https://doi.org/10.17323/jle.2024.24030

Correspondence:
Valery Solovyev,
maki.solovyev@mail.ru

Received: November 21, 2024
Accepted: December 16, 2024
Published: December 30, 2024

CC BY

INTRODUCTION

Natural language processing (NLP) is an important field of research that plays a key role in the development of artificial intelligence. Text understanding and text generation as constituents of NLP have a wide range of applications, including information retrieval, sentiment analysis, machine translation, etc. However,

the existing NLP methods still fail to process context, logical links, lexical chains and detect relationships between parts in a text. The latter refers to both implicit and explicit discourse relations and scholars admit that even hybrid approaches, which combine deep learning and traditional methods, struggle at tasks that heavily involve an understanding of the ways in which entities are connected (Santoro et al., 2018).

Neural network models, especially those based on the Transformer architecture (see Gerasimenko, 2022), have been significantly improving NLP models since the first BERT-based model was designed and developed. Among them BERT, i.e. Bidirectional Encoder Representations from Transformers, presented by Google researchers, stands out among others due to its being conceptually simple and empirically powerful. Designed and developed to pre-train deep bidirectional representations, in most cases BERT models are fine-tuned with only one additional output layer and as such function as state-of-the-art models (see Devlin et al., 2018). The range of applications of BERT models is eminently wide including sentiment analysis, fraud and fake news detection, question-answering systems, document and text classification, information extraction etc. (Rasmy et al. 2021, Atagün et al. 2021, Wang et al. 2020, Jwa et al. 2019, Sun et al., 2019).

A group of widely used BERT models are Masked Language Models, or MLMs, trained to reconstruct the missing tokens which were "masked out", i.e. missed, from the subset of the input text. The training process implies restoring the missed (masked) tokens/words, during which the model learns to generate words in the text taking context into account (Fu et al., 2022). One of reasons BERT, as a pre-trained masked language model, is currently widely used is due to its ability to learn contextualized word representations from large unannotated corpora and restore the masked out fragments (Lai et al., 2020). The success of those models is often attributed to their ability to capture complex syntactic and semantic characteristics of words (Peters et al., 2018).

BERT is viewed as the gold standard for text processing. BERT-based models vary markedly in the number of neurons and parameters. Cointegrated/rubert-tiny is a small model with only 11.8 million parameters included in the well-known HuggingFace's Transformers library (github.com/huggingface/transformers). The full credit of identifying advantages of cointegrated/rubert-tiny over other 10 BERT-based models goes to Bolshakov, V., Kolobov, R., Borisov, E., Mikhaylovskiy, N., and Mukhtarova, G. (2023) who argue that it demonstrated a good balance of accuracy and speed of calculations while processing sentences. The model is highly recommended for quick calculations of small datasets (Tomilov et al., 2024).

We hypothesize that (1) classification of overlapping classes of texts such as Russian fairy tales is a cognitively complex task and (2) its automated classification could be performed using BERT with its enhanced categorizing abilities.

Although the latter were demonstrated on the datasets of well-resourced languages such as English (Tangherlini & Chen 2024), French (see Martin et al., 2019, Bayer et al., 2021), German (Chan et al*.,* 2020, Labusch et al., 2019, Leitner et al., 2020), automation of Russian fairy tales classification, to the best of our knowledge, presents a research problem.

Thus aim of this paper is to demonstrate BERT's potential in the task of classifying Russian folk tales and verify it on a representative corpus of 743 Russian fairy tales.

LITERATURE REVIEW

Fairy Tales as a Genre

Fairy tales make up a unique genre of literature with specific schemata and style. Nevertheless researchers note that fairy tales often contain recurring motifs, archetypes, and plots, which make them a mysterious black box to investigate. Classifications of fairy tales are numerous and based on various features: "leading conflict", motif, main characters, etc. The generally accepted ATU or Aarne–Thompson–Uther Index (Aarne, 1910, Uther, 2004) ranges tales into 5 sections ((1)Animal Tales, (2) Ordinary Folk Tales including Fairy Tales, Religious Tales, Realistic Tales or Novellas, Tales of the stupid Ogre, Giant or Devil, (3) Anecdotes and Jokes, (4) Formula Tales, (5) Unclassified Tales) with an AT number for each entry. The definition of a tale type although published by Thompson in 1928, i.e. after releasing the first AT catalogue in 1910, lacks the main classifying principle. Later fairy tales were classified based on the basis of narrative plots, characters, motifs, etc, but in all cases the catalogues contain numerous exceptions, overlaps of the identified classes, and intersections. Even the generally accepted classification of fairy tales of A. Aarne when revised by N. P. Andreev was downsized to three, i.e. Animal Tales, Magic Tales and Household or Realistic Tales (see Tudorovskaya, 1961). Nevertheless, in the preface to his "Index", Andreev (1929) notes that the accepted classification has a number of shortcomings as the division is always relative and not plausible and the principles of division applied are diverse.

In a bid to overcome the challenges faced, researchers point out the so-called "hard core" and "soft shell" of the fairy tale genre. While the former comprises "classic animal tales" or "tales of magic", the latter is made up of the fairy-tales which may be categorized differently based on one selected parameter. Besides, a narrative, i. e. a plot, may move across genres, acquiring features from a variety of narratives it encounters along its route. All the above indicates that fairy-tale classification is an interesting though extremely laborious and demanding object for automatic classification and analysis (Pompeu, 2019). The above is probably the reason why classification studies of fairy tale texts are relatively rare, although its number has been growing lately (Tangherlini & Chen, 2024).

Text Classification Analysis

Text classification is one of the classic tasks in computational linguistics with an important practical applications including recommender systems which analyze and categorize texts, scan them for user's specific interests, etc. (Kupriavov et al., 2023, Solnyshkina et al., 2024, Reusens et al., 2024). As early as 1997, Kessler, Numberg and Schütze proposed to classify "genres as bundles of facets, which correlate with various surface cues, and argued that genre detection based on surface cues is as successful as detection based on deeper structural properties" (Kessler et al., 1997, p.32). Samothrakis & Fasli (2015) applied machine learning methods to classify fiction from Project Gutenberg collection into six genres, i.e. "science fiction", "horror", "western", "fantasy", "crime fiction", "mystery". The algorithm comprised extraction of relevant information with the help of Natural Language Toolkit and measurement emotional content in each sentence with Wordnet-Affect. The research emphasis was on the analysis of emotive vocabulary: the authors come to the conclusion that the most distinctive feature discriminating the above-mentioned genres is fear. Three years later Worsham & Kalita (2018) implemented a set of different neural network models and classifiers to identify six genres. i.e. Science fiction, Adventure stories, Historical fiction, Love stories, Detective and mystery stories and Western stories. The authors also used multiple strategies to compensate for the extreme lengths of the documents in the dataset and argued that when trained on the BOW form of the Gutenberg Dataset, XGBoost proved to be "a highly optimized, award winning Gradient Boosting solution" (Worsham & Kalita, 2018: p. 1969). Nowadays data extraction, managing and structuring unstructured data envision using a variety of machine learning techniques (Parida et al., 2021) and deep learning neural networks. BERT marked a new level of research and demonstrated significant improvements over previous models on a variety of NLP tasks including text classification. In 2020 (Batraeva et al., 2020) concluded that convolutional neural networks (CNN) and recurrent neural networks (RNN), have gained the greatest popularity for solving classification problems and are rightly recognized as the most effective. The detailed reviews of implementing neural networks to classification tasks published by (Minaee et al., 2021) and (Reusens et al., 2024) came up with revolutionary findings. Reusens et al. (2024) argue that BiLSTM is the overall best-ranked method which significantly outperforms all other methods except LR TF-IDF, and RoBERTa with a confidence level of 95%.

English has always been the most widely studied and resourced language, however scholars worldwide set themselves the task of conducting in-depth genre classification studies into under-resourced languages such as Russian, Arabic (El-Halees, 2017), Hebrew (Devlin, et al. 2018, Liebeskind et al., 2023) and even non-alphabetic languages, i.e. Chinese (Jin et al., 2020), Korean (Liu et al., 2022) and Japanese (Lippert et al., 2022). As for the choice of text collections, the research shows that the most studied is news, including fake news. The range of classes comprises topic, emotion, polarity and even sarcasm detection. Although, there is ample research into other text types and discourses, e,g. Barros, Rodriguez and Ortigosa (2013) focus on automatic classification of Spanish poetry by Francisco de Quevedo utilizing emotional content and sentiment categorization.

The past several years, i.e. 2019-2024, witnessed significant advancements in classification of Russian texts that were largely driven by deep learning techniques and transformer-based models (Solovyev et al., 2023, Tomin et al., 2023). BERT is now widely implemented in numerous applications based on Russian datasets such as fiction (detective stories, children's literature, poetry, fantasy, and science fiction), academic discourse (History, Natural Sciences, Medicine and Health, Culture), business, news, research and political discourse, advertisement, tweets, reviews etc. The text collections, tools, and algorithms used for experiments with Russian text classifications differ greatly. For instance, experiments conducted by Dubovik (2017) on texts of four functional styles, i.e. scientific, fiction, business and media with stylometric methods proved extremely successful with F1-measure ranging from 0.7 in media texts to 1.0 in business. Batraeva, Nartsev and Lezgyan (2020) implemented convolutional neural networks (CNN) on the collection of five genres, i.e. history, detective stories, children's literature, poetry, and science fiction, reaching 73.12%.of classification accuracy for all 5 classes. Lagutina et al. (2021) report that implementing "rhythmic patterns" to range research articles, advertisement, tweets, novels, reviews and political articles into classes resulted in the highest accuracy (F1=98%) for fiction. Two years later the same group of researchers, using a similar algorithm, accomplished an even more ambitious task classifying novels, articles, reviews, VKontakte posts and OpenCorpora news with even higher accuracy (F1=99%) (Lagutina, 2023). A more challenging task, i. e. taxonomy of ten genres, including Fiction, Fantasy, Detectives, Prose, History. Historical Sciences, Information Technology, Natural Sciences, Medicine and Health, Cooking, Culture. Art was undertaken by (Nikolaev, 2022). The best accuracy of results (F1=71.11%) was obtained after only three epochs of training the neural network.

Fairy Tales Classification

Enabled by available datasets and advances in technology, modern scholars accomplish extremely ambitious tasks of fairytales classification. Among the first in the area were (Nguyen et al., 2012, 2013), who trained classification models, i.e. SVMs (2012) and Learning to Rank methods (2013), for Dutch fairytales. The authors reported a macro-average F_1 score of 0.62 for classifying fairy tales and indicated a high impact of character n-grams. Though the implemented models demonstrated relatively moderate success, they were followed by others. In 2013, Nguyen, Trieschnigg, Meder & Theune designed and developed a fairy tale classifi-

er using Learning to Rank and BM25 queries. The features employed in the study were lexical and story similarity, information retrieval measures, as well as subject-verb-object triplets. The results indicated the highest mean reciprocal rank accuracy of 0.82. In the same year, 2013, Karsdorp and Van den Bosch published "Identifying Motifs in Folktales using Topic Models" in which they argued that that Labeled LDA and Big Document Model produce representations that match relatively well to a manually constructed motif classification system used in folktale research.

Six years later, in 2019, based on the Hierarchical Attention Network (HAN), Pompeu successfully evaluated a cross-language neural network approach on the biggest collection in his dataset, i.e. English subset of the folktales. In 2022 Ostrow reports on the unique model with an overall F1 score of 0.77 able to parse fairy tale characters into Proppian archetypes by tracking their probabilistic association with linguistic occurrences such as adjectives and verbs. The researcher argues that the classification schema enables a broader classification of fairy tales into the types identified by Propp (1984). Thus, the research performed on fairy tales automatic classification fail to develop a reliable and accurate taxonomy achieved in various other tasks of computer linguistics. Besides, the studies conducted in the area use different philological classifications of fairy tales and as such lack a unifying foundation theory. As for Russian fairytales, to the best of our knowledge, they were never utilized for type or genre classification. All the above holds great promises for going beyond the unhelpful traditional approaches to the study of fairy-tales as a genre.

METHOD

Dataset

We selected a dataset of Russian fairytales from the collection of Afanasyev (1982) and sourced it from nukadeti.ru and www.rodon.org/other/rnsoj.htm. The collection comprises three main types of tales, i.e. Magic Tales, Realistic (Household) Tales and Animal Tales, which constitute the core of Afanasyev's Russian fairy tales collection. The types of tales differ in their plots, themes, and styles thus being suitable for classification tasks. The Dataset information is provided in Table 1 below.

Table 1

Dataset Information

Fairy tales	Words	Sentences	Number of fairy tales
Realistic tales	10,766	1,179	203
Animal Tales	10,754	1,018	342
Tales of Magic	9,371	874	198
TOTAL	**30,891**	**3,071**	**743**

Method

The method of training a neural network implemented in the current research is standard and includes (1) training the network with the texts classified and labeled by experts; (2) reorganization of the network parameters as a result of multiple stages of training and (3) evaluation of accuracy and efficiency on validated datasets.

BERT training involves tuning hyper-parameters, i. e. mini-batch size, number of epochs, learning rate, etc. The loss function (Loss) is viewed as a significant parameter which measures effectiveness of the model to predict the target values compared to the true one. The loss function calculates the model error and is used to update the model parameters during training with the help of gradient descent or other optimization algorithms.

To train the model, the data was split into two sets: the training (df_train) set and the validation (df_val) set in an 80/20 ratio, where 80% of the data is used for training and 20% for validation. The latter enables to test the quality of model generalization functions based on the data which was not used in training.

Below we provide description of parameters, training and testing procedures.

BERT was trained on two simultaneous tasks: generation of missing tokens and prediction of next sentences. BERT also receives tokenized pairs of sentences with masked, i.e. missed, tokens. Thanks to the MLM technique, the network learns a deep bidirectional representation of the language, thus taking into account context of a sentence. The task of predicting the next sentence presents a binary classification task which implies identifying if the second sentence follows the first one. Thanks to this binary classification, network trains to identify relationships between sentences in a text.

Although BERT is generally a bidirectional transformer, in this research we used only the input encoder. The main idea of transformers is to apply an attention mechanism that allows the model to weigh the importance of different parts of the input text for each token processed.

The overall architecture of BERT is illustrated in Figure 1 with a sentence fragment fed as input.

BERT architecture utilizes several types of vector representations (embeddings) that transform text data into numeric vectors.

1. Token Embedding

Each word or sub-word is represented as a unique vector in the embedding space, which is standard practice in modern NLP models. BERT-based models function similarly and imply:

(1) Tokenization Model: BERT implements tokenization using WordPiece algorithm, which splits words into sub-words. This approach helps the model effectively handle rare words and morphological variations. For example, the word "unbelievable" might be split into "un," "believ," and "able," each of which has its own embedding.

(2) Token Vectors: Each token, whether it is a complete word or part of a word, receives a numerical representation—a fixed-length vector (e.g., 768 dimensions for BERT base and 1024 for BERT large). This vector includes semantic information, helping the model understand the meaning and context of words.

(3) Characteristics: The token embedding vector allows the model to understand relationships between words, even if they do not follow each other in a sentence. This is crucial for transformers, which rely on self-attention. Token embeddings allow the model to map out the meaning of words and sub-words, building semantic connections.

Thus, token embedding provides the model with information about the meaning and semantic context of individual tokens in a sentence.

2. Segment Embedding

For BERT, it is of utmost importance to distinguish tokens from different sentences, especially when handling tasks that require understanding of two sentences or text parts. Its works implies the following:

(1) Purpose of Segments: BERT is trained on tasks that require distinguishing the contexts of two sentences, similar to Natural Language Inference (NLI), where it is necessary to determine if two sentences are contradictoty, neutral, or entail each other.

(2) Encoding Segments: Each token receives a special segment embedding that indicates which sentence it belongs to:

- Segment A comprises tokens from the first sentence.
- Segment B comprises tokens from the second sentence (if there is one).

(3) Single-Sentence Input: When the text contains only one sentence, all tokens are assigned to the same segment embedding, meaning they all belong to one sentence. This does not hinder the model's ability to understand meaning and structure, as segment embeddings simply help in tasks with more than one part of text.

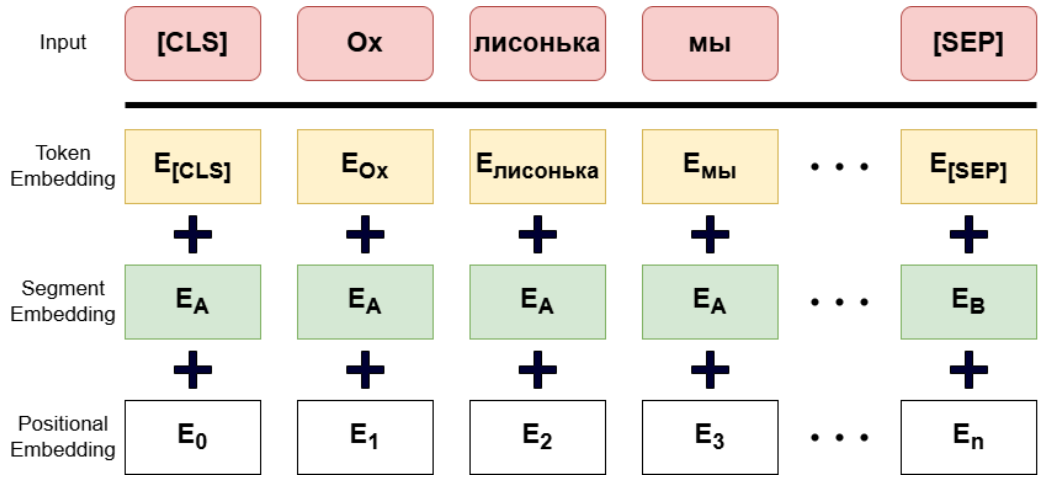
Segment embeddings allow the model to understand not only text-based features of words but also contextual features in two-part tasks, such as question answering and inference tasks.

3. Positional Embedding

Transformers, including BERT, are built on self-attention, where the model can view all tokens at once but does not

Figure 1

BERT architecture



know their positions. Positional embeddings are added to tokens to account for word order.

(1) Lack of Order Information: Transformers cannot recognize token sequence on their own as they see all words simultaneously and lack built-in information on sequencing words. This differs BERT from RNNs, which process information sequentially taking order into account.

(2) Positional Embeddings: For the purpose of helping the model to differentiate token positions, each token is assigned to a positional vector. Each position is unique and complies with the first, second, etc., positions of tokens in the sentence. These vectors help the model understand the relative position of tokens, which is necessary for accurately capturing structure and order.

(3) Mathematical Formula: BERT creates positional embeddings using sinusoidal functions of different frequencies, which allow the model to map out positions at both short and long distances. Each token position has a unique vector based on these sinusoidal functions.

Positional embeddings provide the model with information on each token's position, which is important for maintaining text structure, especially in long sentences.

Input

BERT input is the sum of three embeddings:

Input Embedding = Token Embedding + Segment Embedding + Position Embedding

Each token is represented in the model as the sum of its token, position, and segment embeddings, which together form a fixed-length vector (usually 768 or 1024, depending on the model configuration). Thus, the input vector for each token contains information not only about what the token itself is (Token Embedding), but also about its position in the sentence (Position Embedding) and what segment it belongs to (Segment Embedding). Data classification presupposes proper preparation of the dataset, when each sequence in a text receives a corresponding class label. Hypothetically each document in the collection may belong to more than one class, and as such receives the corresponding number of labels: two in a binary classification or more in a multi-label classification problem.

On the next stage, each word is tokenized with PyTorch (pytorch.org/get-started/locally/) (Fig. 2 below) and each sentence is converted into an identifier. Figures 2 and 3 illustrate representation of sentences.

These actions turn the dataset into a list (or Series/DataFrame object from pandas) of lists. Before BERT processes the dataset, vectors lengths are balanced by padding shorter vectors with an ID of 0.

BERT utilizes a special vocabulary compiled on the pre-training stage. It contains thousands of tokens mapped to a unique identifier. Compiling vocabulary involves using WordPiece algorithm to breaks words into sub-tokens. The latter allows the model to effectively deal with rare and unknown words by breaking them into smaller parts. BERT also requires specific tokenization which breaks texts into tokens and adds special tokens [CLS] at the beginning and [SEP] at the end of each sequence.

Figure 2
Tokenization

Raw dataset

category	text
animal_fairy_tale	Идет волк по лесу. Видит, дятел долбит
animal_fairy_tale	А дятел волку и говорит: А ты, волк, все
animal_fairy_tale	А я тебе принесу овечек!
animal_fairy_tale	Согласилась лиса.
animal_fairy_tale	Вот волк приносит лисе овец: одну, дру
animal_fairy_tale	Ты зарежь ее и принеси хвост и гриву на
animal_fairy_tale	Пошел волк и видит лошадь . Подкрался
animal_fairy_tale	И сейчас по снегу волка косточки блестя
animal_fairy_tale	Бежала лисица по лесу, увидала на дере
animal_fairy_tale	Лиса с журавлем подружилась, даже по
animal_fairy_tale	– Приходи, куманек, приходи, дорогой!
animal_fairy_tale	Журавль хлоп-хлоп носом, стучал, стуча
animal_fairy_tale	Лиса начала вертеться вокруг кувшина,
animal_fairy_tale	Жили курочка с кочетком, и пошли они
animal_fairy_tale	Вот кинул кочеток орешек, и попал кур
animal_fairy_tale	Жил кот с кочетком. Кот идет за лыками

Sequences of Token IDs

```
{
    'input_ids': tensor([
        [101, 1186, 10325, 29436, 10325, 1011, 1181, 29113, 29436, 10325,
         1196, 22919, 10260, 16856, 10325, 23925, 1196, 14150, 1196, 22919,
         10260, 16856, 29748, 29750, 14150, 10325, 1012, 1195, 14150, 29747,
         29436, 10325, 1198, 1192, 10325, 29750, 1184, 14150, 29752, 28598,
         1188, 1196, 29113, 18947, 14150, 29752, 15290, 23925, 1012, 102]
    ]),
    'token_type_ids': tensor([
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
    ]),
    'attention_mask': tensor([
        [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
    ])
}
```

Figure 3

Matrix/Tensor for the Neural Network Input

		Tokens in each sequence			
		0	1	…	66
Input sequences (reviews)	0	101	1037	…	0
	1	101	2027	…	0
	…	…	…	…	
	1,999	101	1996	…	0

BERT input parameters include the following:

1. Text data**:** Russian fairy tales.
2. Fairy tales categories**:** Animal, Realistic (Household), Magical.
3. Tokens**:** text is tokenized using a pre-trained BERT tokenizer.
4. Token identifiers (input_ids): Numeric representations of words in the text.
5. Attention masks: Specification of tokens to be taken into account.
6. Category labels: Converted to numeric labels for classification.

We operationalize cross-entropy loss or cross-entropy loss function widely used for classification purposes, specifically in neural networks. For multi-class classification we implement the following formula:

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

During the training process, the model goes through all the data sets in each epoch. The main steps include the following:

1. *Feeding data to the model***:** The data from the training set is fed to the model to make predictions.
2. *Calculating loss and updating model weights:* the loss is measured with the loss function (CrossEntropyLoss), then backpropagation of the error is performed and the model weights are updated using the optimizer.
3. *Printing average loss and accuracy on the training set*: After each epoch, with the purpose to monitor the training process we calculate the average loss and accuracy on the training dataset.
4. *A screenshot of a fragment of the learning process is presented below in Figure* 4.

As a result of all the above steps, we developed an effective system to train the model with BERT neural network aimed at classifying fairy tales. This process includes careful data preparation, tuning the optimizer and scheduler, as well as sequential training and evaluation of the model to achieve high accuracy and performance.

Figure 4

Model Training

```
100%|██████████| 8233/8233 [04:17<00:00, 31.98it/s]
100%|██████████| 969/969 [00:09<00:00, 104.03it/s]
Epochs: 1 | Train Loss:  0.301            | Train Accuracy:  0.633            | Val Loss:  0.204            | Val Accuracy:  0.756
100%|██████████| 8233/8233 [03:53<00:00, 35.25it/s]
100%|██████████| 969/969 [00:09<00:00, 104.61it/s]
Epochs: 2 | Train Loss:  0.151            | Train Accuracy:  0.820            | Val Loss:  0.169            | Val Accuracy:  0.796
100%|██████████| 8233/8233 [03:53<00:00, 35.24it/s]
100%|██████████| 969/969 [00:09<00:00, 104.49it/s]
Epochs: 3 | Train Loss:  0.098            | Train Accuracy:  0.887            | Val Loss:  0.162            | Val Accuracy:  0.810
100%|██████████| 8233/8233 [03:53<00:00, 35.24it/s]
100%|██████████| 969/969 [00:09<00:00, 104.61it/s]
Epochs: 4 | Train Loss:  0.075            | Train Accuracy:  0.916            | Val Loss:  0.162            | Val Accuracy:  0.813
100%|██████████| 6565/6565 [01:00<00:00, 108.86it/s]
```

Results of the Work and Evaluation of the Model

The models comparison and evaluation were carried out using several generally accepted metrics, i.e. Loss, Accuracy, Precision, Recall. The research algorithm includes comparison of the following models: rubert -tiny, ai -forever/ ruBert -base, DeepPavlov/rubert - base - cased - sentence (cf. Tables 2 – 5). Constituting a characteristic, feature and a parameter, *accuracy* evaluates how correctly the model classifies objects in a set and is measured as the ratio of correct predictions to the total number of predictions:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

where TP (True Positive): Correctly predicted positive classes; TN (True Negative): Correctly predicted negative classes; FP (False Positive): Incorrectly predicted positive classes and FN (False Negative): Incorrectly predicted negative classes.

Precision measures how many of all positive classes predicted by the model are actually positive and is calculated with the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

where TP (True Positive) is correctly predicted positive classes; FP (False Positive) is incorrectly predicted positive classes.

The results demonstrate, that the optimal number of epochs for the models under study is 5, as it reaches its pick with the batch size of 5 (see Tables 2-5) and when the number of batches increases to 6, the accuracy on the validation set begins to decrease (see Table 5).

In our experiment, the best result received was that by cointegrated / rubert-tiny model with accuracy of 0.875 and minimal losses. However, the model is far from being perfect, and about 12% of the data was classified incorrectly. The erroneous examples from the validation sample are presented in Table 6 below.

Table 2

Models Evaluation with Batch Size 3

Model	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.098	0.810	0.820	0.815	3
ai-forever / ruBert-base	0.114	0.804	0.810	0.805	3
DeepPavlov / rubert -base-cased-sentence	0.189	0.727	0.735	0.730	3

Table 3

Models Evaluation with Batch Size 4

Model name	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.075	0.813	0.810	0.815	4
ai-forever / ruBert-base	0.114	0.804	0.800	0.805	4
DeepPavlov / rubert -base-cased-sentence	0.189	0.727	0.730	0.725	4

Table 4

Models Evaluation with Batch Size 5

Model name	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.054	0.875	0.870	0.870	5
ai-forever / ruBert-base	0.114	0.804	0.800	0.805	5
DeepPavlov / rubert -base-cased-sentence	0.189	0.727	0.730	0.725	5

Table 5

Learning Results in Batch 5 and 6

Batch size	Train Loss	Train Accuracy	Val Loss	Val Accuracy
5	0.008	0.989	0.054	0.875
6	0.008	0.983	0.054	0.863

The model classified the sentences “In the old days there lived a peasant. The peasant had a bee” from the fairy tale “How the Deacon was Treated with Honey” as components of the Class “Fairy Tale about Animals”, although according to the test data set they belong to the Class “Realistic Fairy Tale”. Similarly, the fragment “A passerby says to them: “You, good fellows, instead of nagging and tugging, should get off the cart. Here the horse will ride up the mountain!” from the fairy tale “Seven Stupid Agathons” as a constituent of the Class “Household Fairy Tale” was assigned by the model to the class of “ Fairy Tale of Magic“.

We suggest solving the problem by changing the process of training, since, in the tested version, the model perceives and processes key words and the local not the global context only. It may be also caused either by the class imbalance problem or class overlapping problem in the original dataset. The latter refers to the cases when (in our case) fairy tales from different classes exhibit similar features. The problem is viewed as “one of the toughest problems in machine learning and data mining communities” (see Xiong et al., 2010: p. 491). In situations when texts classification is hampered, it is important and recommended to increase the size of the input data (e.g. for the class of Realistic fairy tales) and retrain the model. The suggested minimum of the size increase is one paragraph.

As it was stated above, we selected 5 batches, but the tensor structure of the input data on everyday fairy tales became 4 – 5 times larger. The results for the best cointegrated / rubert - tiny model with an increase in the input data size to one paragraph only improved Accuracy, precision and Recall significantly (see Table 7).

Table 8 exemplifies class probability of classifying four fairy tales and as we can see, the classification accuracy though high is still below 100%.

DISCUSSION

Classification as one of the main scientific methods applied ubiquitously requires special carefulness when used on works of art. It is caused predominantly by their nature, i.e. the ability to reflect the whole world and encapsulate myriads of ideas. The latter makes the works of art difficult to classify. The current study aims at demonstrating potential of the latest generation of neural networks to solve the abovementioned problem. And the Russian fairy tales embody a model problem as the resulting classification is easy to validate against the indexed catalogue compiled by professional linguists which researchers have at their disposal. The research indicates that BERT-based classification model

Table 6

Classification Errors Generated by Cointegrated / Rubert-tiny Model

oid	text	category
729	V staroye vremya zhil da byl muzhichok. U muzhichka byla pchela	animal_fairy_tale
	Once upon a time there lived a peasant. The peasant had a bee	
594	Vidit, chto ovtsy razbrelis’ po polyu, davay ikh lovit’ da glaza vydirat’. Vsekh perelovil, vsem glaza vydolbil, sobral stalo v odnu kuchu i sidit sebe radokhonek	animal_fairy_tale
	He saw sheep had wandered off across the field and began catching them and gouging out their eyes. He caught them all, gouged out their eyes, gathered them all into one flock and sat there happily	
497	Da smotri bol’shogo vozu ne nakladyvay, a vpered na menya ne nadeysya: segodnya day da zavtra day, a potom	animal_fairy_tale
	But watch out, don’t load a big cart, and don’t rely on me in the future: give me today, give me tomorrow, and then	
407	Ne spal, vse barskuyu zagadku otgadyval. Razdumayet, tak i malo li chego na svete ne byvayet, a i to v um pridet: “Mozhet, eto i byvayet, tol’ko ya ne	animal_fairy_tale
	He did not sleep, but tried to guess the master’s riddle. Things do happen in the world. And it occurs to him: «Maybe this does happen, but I don’t	

Table 7

Rubert-tiny Results for batch size 5

Model name	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.0 34	0.959	0.915	0.920	5

Table 8
Probability of Belonging to Class

Genre	Fairy-tale	Class Probability		
		Realistic tales	Animal Tales	Tales of Magic
Realistic Tales	«Porridge from an Axe» by Alexander Afanasyev	0.9281	0.0349	0.0368
Realistic Tales	"Soldier's Overcoat" by Sergey Saptsov	0.8563	0.0174	0.1260
Animal Tales	"The Crow and the Crawfish" by Konstantin Ushinsky	0.0913	0.7042	0.2045
Tales of Magic	"Geese-Swans" Alexey Tolstoy	0.0667	0.0932	0.8398

demonstrates high accuracy of classifying fairy tales into the three main categories. Below we provide our views of the findings received and research prospects.

We achieved a significantly higher accuracy, i.e. 95%, than Nguyen et al. (2012; 2013) with 82 %. Even though this fact alone does not signify a breakthrough, it marks sustainability and competitiveness of our algorithms. Earlier classifications of other text types resulted in different outcomes largely dependent on the groups categorized. E.g. Lagutina et al. (2021), when ranging into classes such various types of texts as research articles, advertisements, tweets, novels, reviews and political articles, achieved 98% of accuracy. In this regard, classification of fairy tales as one genre into subclasses is a much more difficult task and 98% of accuracy is viewed at the moment as increadable and unattainable. In addition to the main result, i.e. a fairly high percentage of classification accuracy achieved by the neural network on Russian fairy tales, we also obtained a number of auxiliary results potentially useful for further research. Namely: (a) while contrasting three modifications of BERT, cointegrated / rubert -tiny variant performed the best results, b) the optimal number of training epochs proves to be 5 only, c) input data increase, equal to one paragraph only, is prone to higher levels of accuracy. All the above may be viewed as mandatory conditions of elaborated algorithm.

While presenting text classification experimental failures, researchers point out a number of reasons. The first one usually refers either to lack of representativity, insufficience or misbalance of the training collection, categories or sub-categories of the texts under study (Pompeu, 2019). Similarly, in our case the model performance tends to scale with the number of samples for each category in the collection, which suggests that results may improve if the size of the training data increases. Another reason for misclassification of fairy tales is the above-mentioned "class overlapping problem" (Xiong et al., 2010) when constituents of sub-classes within the class possess very similar characteristics. The latter is very true about fairy-tales "since it is difficult to determine which of the features [in a fairy-tale – *authors' insert*] is the main one, the task is reduced to assigning the same fairy-tale to two or more classes (groups) "(Andreev, 1929). What we have managed to accomplish in this research is to state a new problem and a baseline which are open up to further studies.

Prospects of the designed fairy tale classifier lie in the three main directions. First, with a representative dataset and using tales and stories from around the world we plan to pursue comparative classification studies. The latter are of great interest to linguists, historians, cultural scientists and anthropologists and open a vista for further cognitive studies. Second, as fairy tales are viewed in the modern research paradigm as a genre manifesting and transmitting cultural values and as such able of targeting diverse broad audiences, we also plan to implement the designed algorithms into developing a fairy tale profiler with the function of assigning fairy tales for target age and cultural groups. A fairy tale profiler of the kind will provide possibilities to conduct stylometric and multidimensional analysis of fairy tales for specific age groups thus enabling findings and discoveries of cultural and cognitive (dis)similarities of peoples. Third, snce there are a number of overlapping genres manifesting features similar to those in fairy tales, i.e. fables, myths, fantasy, etc., experiments with the neural network trained on the three abovementioned types of fairy tales are of the authors' particular research interest.

Limitations

A standard limitation of neural networks utilization is its dataset or, more specifically, its amount and quality. A relatively small collection of fairy tales used in the current study probably affected accuracy of its classification. Another problem is ambiguity of classification parameters accepted (or ignored) by human experts, but causing fundamental questions: which of the proposed classifications is "correct" (if any), which of them should be used to train the neural network and whether any classification of neural network may be qualified "correct". The results we obtained are not absolute, though positively relative to the selected classification.

CONCLUSION

Our study highlights significant feasibility of the automatic classification of fairy tales and confirms that further explo-

ration of BERT-based classification model is necessary. BERT represents a substantial advancement in natural language processing due to its ability to provide deep analysis and process context. The present study highlights BERT's significant classification power and effectiveness in developing a taxonomy of Russian folk tales. While pre-trained on representative corpus and fine-tuned for specific tasks BERT is able to accurately classify texts, identifying subtle relationships and contextual features characteristic of Russian folk tales. In particular, models such as cointegrated / rubert-tiny, ai forever / ruBert-base, and DeepPavlov / rubert-base-cased-sentence, demonstrated high levels of accuracy, with the best accuracy of 95.9% for the cointegrated / rubert-tiny model.

BERT classification power opens up broad prospects for further research and applications, however, despite the progress made, there are still open questions and directions for future research, including quality improvement of tokenization and embeddings, as well as adapting the model to different languages and specific tasks. Overall, BERT demonstrates great potential for enhancing NLP technologies and advancing the time of much more sophisticated and intelligent NLP systems. It is a powerful tool that can significantly improve the quality of automated text analysis and offer new opportunities for research and application in a wide range of areas.

Classification problems regarding fairy tales are caused by numerous factors including topical similarity of classification objects, miscellaneousness of their constituents and the lack of universally accepted genre classification. Two more contributions to the above are fuzzy boundaries of fairy tale as the concept and their ability to be incorporated into bigger genres, e.g. "Master and Margarita", "Monday Begins on Saturday", "The Lord of the Rings". Further research using increasingly powerful AI systems may result in better understanding and conceptualization of fiction. Our findings signify both challenges and prospects in the area.

ACKNOWLEDGEMENT

The research was supported by the RSF grant 24-28-01355 "Genre-discourse characteristics of the text as a function of lexical range".

DECLARATION OF COMPETITING INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Valery Solovyev: conceptualization; investigation; methodology; project administration.

Marina Solnyshkina: formal analysis; writing – original draft; funding acquisition.

Andrey Ten: resources; software; visualization.

Nikolai Prokopyev: writing – review & editing.

REFERENCE

Aarne, A. (1910). *Verzeichnis der Märchentypen* [List of fairy tale types]. *Folklore Fellows' Communications,* (3). Suomalaisen Tiedeakatemian Toimituksia.

Andreev, N. P. (1929). *Index of fairy-tale plots according to the Aarne System*. Russian Geographical Society.

Atagün, E., Hartoka, B. & Albayrak A. (2021). Topic modeling using LDA and BERT Techniques: Teknofest example.*6th International Conference on Computer Science and Engineering* (pp. 660–664)*.* Akdeniz University Publisher. https://doi.org/10.1109/UBMK52708.2021.9558988

Barros, L., Rodriguez, P., & Ortigosa, A. (2013). Automatic classification of literature pieces by emotion detection: A study on Quevedo's poetry. *Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 141–146)*.* IEEE. https://doi.org/10.1109/ACII.2013.30

Batraeva, I. A., Nartsev, A. D., & Lezgyan, A.S. (2020). Using the analysis of semantic proximity of words in solving the problem of determining the genre of texts within deep learning", *Tomsk State University Journal of Control and Computer Science*, *50*, 14–22. https://doi.org/10.17223/19988605/50/2

Bayer, M., Kaufhold, M.-A., & Reuter, Ch. (2021). *A survey on data augmentation for text classification*. arXiv preprint. arXiv:2107.03158. https://doi.org/10.48550/arXiv.2107.03158

Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6788–6796). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.598

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint.* arXiv:1810.04805, https://doi.org/10.48550/arXiv.1810.04805

Dubovik, A.R. (2017). Automatic text style identification in terms of statistical parameters. *Komp'yuternaya lingvistika i vychislitel'nye ontologii*, *1,* 29–45. https://doi.org/10.17586/2541-9781-2017-1-29-45

Fu, Z., Zhou W., Xu J., Zhou H., & Li L. (2022). Contextual representation learning beyond Masked Language Modeling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 2701-2714). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.193

El-Halees, A. M. (2017). Arabic text genre classification. *Journal of Engineering Research and Technology, 4*(3), 105–109.

Gerasimenko, N.A., Chernyavsky, A.S. & Nikiforova, M.A. (2022) ruSciBERT: A transformer language model for obtaining semantic embeddings of scientific texts in Russian. *Doklady Mathematics, 106* (Suppl. 1), 95–96. https://doi.org/10.1134/S1064562422060072

Jin, Q., Xue, X., Peng, W., Cai, W., Zhang, Y., Zhang, L. (2020). TBLC-rAttention: A deep neural network model for recognizing the emotional tendency of Chinese medical comment. *IEEE Access, 8*, 96811–96828. https://doi.org/10.1109/ACCESS.2020.2994252

Jwa, H. D. Oh, K. Park, J. M. Kang, & H. Lim (2019). exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences*, *9*(19), 4062. https://doi.org/10.3390/app9194062

Karsdorp, F. & Bosch, Van den A. (2013). Identifying motifs in folktales using topic models. *Proceedings of BENELEARN 2013* (pp. 41–49). Radboud University. https://hdl.handle.net/2066/112943

Kelodjoue, E., Goulian, J., & Schwab D. (2022). Performance of two French BERT models for French language on verbatim transcripts and online posts. *Proceedings of the 5th International Conference on Natural Language and Speech Processing* (pp. 88–94). Association for Computational Linguistics. https://aclanthology.org/2022.icnlsp-1.10

Kessler B., Numberg G. & Schütze H. (1997). Automatic detection of text genre. *Proceedings of the Eighth Conference on European chapters of the Association for Computational Linguistics.* (pp. 32–38). Association for Computational Linguistics. https://doi.org/10.3115/976909.979622

Kupriyanov, R.V., Solnyshkina, M.I. & Lekhnitskaya, P.A. (2023). Parametric taxonomy of educational texts. *Science Journal of VolSU. Linguistics, 22*(6), 80–94. https://doi.org/10.15688/jvolsu2.2023.6.6

Labusch, K., Kulturbesitz, P., Neudecker, C., & Zellhofer, D. (2019). BERT for named entity recognition in contemporary and historical German. *Proceedings of the 15th Conference on Natural Language Processing* (pp. 9–11). Erlangen.

Lagutina, K. V., Lagutina, N. S., & Boychuk, E. I. (2021). Text classification by genre based on rhythm features. *Modeling and Analysis of Information Systems*, *28*(3), 280–291. https://doi.org/10.18255/1818-1015-2021-3-280-291

Lagutina, K. V. (2023). Genre classification of Russian texts based on Modern Embeddings and Rhythm. *Automatic Control and Computer Sciences, 57*(7), 817–827. https://doi.org/10.3103/S0146411623070076

Lai, Y. A., Lalwani, G. & Zhang, Y. (2020). context analysis for pre-trained masked language models. *Findings of the Association for Computational Linguistics* (pp. 3789–3804). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.338

Liebeskind, Ch., Liebeskind, Sh., & Bouhnik, D. (2023) Machine translation for historical research: A case study of Aramaic-Ancient Hebrew translations. *Journal on Computing and Cultural Heritage, 17*(2), 1–23. https://doi.org/10.1145/3627168

Leitner, E., Rehm, G., & Moreno-Schneider, J. (2020). A dataset of German legal documents for named entity recognition. *arXiv preprint*. arXiv:2003.13016. https://doi.org/10.48550/arXiv.2003.13016

Lippert, Ch., Junger, A., Golam R., Md., Mohammad Ya., Hasan Sh., Md, & Chowdhury, Md. (2022). *Kuzushiji (Japanese Text) classification*. Technical Report. https://doi.org 10.13140/RG.2.2.22416.07680

Liu, C., Zhao, Y., Cui X. & Zhao, Y. (2022) A comparative research of different granularities in Korean text classification. In *IEEE International Conference on Advances in Electrical Engineering and Computer Applications* (pp. 486–489). CONF-CDS. Publisher. https://doi.org/10.1109/AEECA55500.2022.9919047

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Villemonte de La Clergerie, É., Seddah, D., & Sagot, B. (2019). Camembert: A tasty French language model. *arXiv preprint.* arXiv:1911.03894. https://doi.org/10.18653/v1/2020.acl-main.645

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys, 54*(3), 1–40. https://doi.org/10.1145/3439726

Nikolaev, P.L. (2022) Classification of books into genres based on text descriptions via deep learning. *International Journal of Open Information Technologies, 10*(1), 36–40.

Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2012). Automatic classification of folk narrative genres. *Proceedings of the KONVENS 2012* (pp. 378–382). ASAI. http://www.oegai.at/konvens2012/proceedings/56_nguyen12w/

Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2013) Folktale classification using learning to rank. *Proceedings of the European Conference on Information Retrieval*. *Lecture Notes in Computer Science* (vol. 7814, pp. 195–206). Springer. https://doi.org/10.1007/978-3-642-36973-5_17

Ostrow, R. A., (2022). Heroes, villains, and the in-between: A Natural Language Processing approach to fairy tales. *Senior Projects Spring,* 275.

Parida, U., Nayak, M., Nayak, A.K., (2021) News text categorization using random forest and naive bayes. In *1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology* (pp. 1–4). IEEE. https://doi.org/ 10.1109/ODICON50556.2021.9428925

Peters, M., E., Neumann, M., Iyyer, M., Gardner, M., Clark, Ch., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv*, abs/1802.05365. https://doi.org/ 10.18653/v1/N18-1202

Pompeu, D. P. (2019). *Interpretable deep learning methods for classifying folktales according to the Aarne-Thompson-Uther Scheme* [Master's Thesis]. Instituto Superior Técnico.

Propp, V. (1984). *The Russian fairy tale*. Izd. LSU.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. (2021) Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine, 4*(1)*, 86.* https://doi.org/10.1038/s41746-021-00455-y

Reusens, M., Stevens, A., Tonglet, J., De Smedt, J., Verbeke, W., Vanden Broucke, S., & Baesens, B. (2024). Evaluating text classification: A benchmark study. *Expert Systems with Applications, 254,* 124302. 10.1016/j.eswa.2024.124302

Sabharwal, N. & Agrawal, A. (2021). *BERT model applications: Question answering system in hands-on question answering systems with BERT*. Apress eBooks. https://doi.org/10.1007/978-1-4842-6664-9

Samothrakis, B. S., & Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS One, 10*(11), e0141922. https://doi.org/10.1371/journal.pone.0141922

Santoro, A. & Faulkner, R. & Raposo, D. & Rae, J. & Chrzanowski, M. & Weber, Th. & Wierstra, D. & Vinyals, O. & Pascanu, R. & Lillicrap, T. (2018). *Relational recurrent neural networks*. *arXiv.* https://doi.org/10.48550/arXiv.1806.01822

Solnyshkina, M.I., Kupriyanov, R.V. & Shoeva, G.N. (2024). Linguistic profiling of text: Adventure story vs. Textbook. In *Scientific Result. Questions of Theoretical and Applied Linguistics*, *10*(1), 115-132. https://doi.org/10.18413/2313-8912-2024-10-1-0-7 (In Rus).

Solovyev, V., Solnyshkina, M., & Tutubalina, E. (2023). Topic modeling for text structure assessment: The case of Russian academic texts. *Journal of Language and Education*, *9*(3), 143-158. https://doi.org/10.17323/jle.2023.16604

Sun, F., Liu, J., Wu, J., Pei, Ch., Lin, X., Ou, W. & Jiang P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1441–1450). Association for Computing Machinery. https://doi.org/10.1145/3357384.3357895

Tangherlini, T. & Chen, R. (2024). Travels with BERT: Surfacing the intertextuality in Hans Christian Andersen's travel writing and fairy tales through the network lens of large language model based topic modeling. *Orbis Litterarum, 79*(6), 519–562. https://doi.org/10.1111/oli.12458

Tianqi, Ch. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Tomin, E., Solnyshkina, M., Gafiyatova, E. & Galiakhmetova, A. (2023). Automatic text classification as relevance measure for Russian school physics texts. In 2023 *16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip* (pp. 366-370). IEEE. https://doi.org/10.1109/MCSoC60832.2023.00061

Tudorovskaya, E.A. (1961). On classification of Russian folk fairy tales. Specifics of Russian folklore genres. *Specificity of genres of Russian folklore: Theses of the report.* Institute of Russian Literature (Pushkin House).

Uther, H.-J. (2004). The types of international folktales: A classification and bibliography, based on the system of Antti Aarne and Stith Thompson. *Folklore Fellows' Communications* (vol. 3, pp. 284–286). Suomalainen Tiedeakatemia.

Thompson, S. (1928). The types of the folk-tale: A classification and bibliography. *Folklore Fellows' Communications,* (74). Suomalainen Tiedeakatemia.

Thompson, S. (1977). *The folktale*. University of California Press.

Wang, Z., Wu, H. Liu, H.& Cai, Q.-H. (2020). BertPair-networks for sentiment classification. *2020 International Conference on Machine Learning and Cybernetics* (pp. 273–278). IEEE Xplore. https://doi.org/10.1109/ICMLC51923.2020.9469534

Worsham, B, J., & Kalita, J. (2018). Genre identification and the compositional effect of genre in literature. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1963–1973). Association for Computational Linguistics. https://aclanthology.org/C18-1167

Xiong, H. & Wu, J. & Liu, L. (2010). Classification with ClassOverlapping: A systematic study. *1st International Conference on E-Business Intelligence* (pp. 303–309). Atlantis Press. https://doi.org/10.2991/icebi.2010.43

DATASET SOURCES

Narodnye Russkie Skazki (Engl. Russian Folklore Tales) from the collection by A. N. Afanasyev, Moscow, Pravda Publishing House, 1982.

Russian folk tales, Moscow, "Fiction", 1965.

Nukadeti.ru

www.rodon.org/other/rnsoj.htm