Вестник РУДН. Серия: Инженерные исследования
RUDN Journal of Engineering Research

2025;26(3):298–309

ISSN 2312-8143 (Print); ISSN 2312-8151 (Online)
journals.rudn.ru/engineering-researches

DOI: 10.22363/2312-8143-2025-26-3-298-309
EDN: ZZNVXS

Научная статья / Research article

Сравнение моделей и методов классификации текста

А.В. Захарова✉, **А.Ю. Вишнякова**, **А.А. Детков**

Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, *Екатеринбург, Российская Федерация*
✉ zakharova.linusha@mail.ru

История статьи

Поступила в редакцию: 17 апреля 2025 г.
Доработана: 13 июня 2025 г.
Принята к публикации: 7 июля 2025 г.

Заявление о конфликте интересов

Авторы заявляют об отсутствии конфликта интересов.

Аннотация. Рассмотрен процесс автоматической классификации текста и его составляющие. Актуальность данной темы обусловлена стремительным ростом объема данных и развитием технологий машинного обучения. Цель исследования — определение наилучших методов и моделей автоматической классификации текста. В качестве материалов для анализа были выбраны научные статьи, написанные в течение последних четырех лет, наиболее подходящие по теме. В результате определено, что эффективная предобработка текстовых данных должна состоять из нормализации, токенизации, удаления стоп-слов и стемминга или же лемматизации. Для представления текста целесообразно использовать модель BERT. Однако следует отталкиваться от условий конкретной задачи, в которых альтернативные подходы могут быть предпочтительнее. Наилучшими методами непосредственно классификации текста можно считать метод логистической регрессии, сверточные нейронные сети и RoBERTa. Выбор среди этих моделей зависит от назначения и технических возможностей.

Ключевые слова: обработка естественного языка, NLP, предобработка текста, представление текста, машинное обучение, нейронные сети

Вклад авторов

Вишнякова А.Ю., Детков А.А. — концепция и дизайн исследования; *Захарова А.В.* — обработка материалов, написание текста. Все авторы ознакомлены с окончательной версией статьи и одобрили ее.

Для цитирования

Захарова А.В., Вишнякова А.Ю., Детков А.А. Сравнение моделей и методов классификации текста // Вестник Российского университета дружбы народов. Серия: Инженерные исследования. 2025. Т. 26. № 3. С. 298–309. http://doi.org/10.22363/2312-8143-2025-26-3-298-309

© Захарова А.В., Вишнякова А.Ю., Детков А.А., 2025

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License
https://creativecommons.org/licenses/by-nc/4.0/legalcode

Comparison of Text Classification Models and Methods

Angelina V. Zakharova✉, **Alina Yu. Vishnyakova**, **Alexander A. Detkov**

Ural Federal University named after the first President of Russia B.N. Yeltsin, *Ekaterinburg, Russian Federation*
✉ zakharova.linusha@mail.ru

Article history
Received: April 17, 2025
Revised: April 17, 2025
Accepted: July 7, 2025

Conflicts of interest
The authors declare that there is no conflict of interest.

Abstract. The study considers the process of automatic text classification and its components. The relevance of this topic is due to the rapid growth of data and the development of machine learning technologies. The purpose of the study is to determine the best methods and models for automatic text classification. The scientific articles written over the past four years that are most suitable for the topic were selected as material for analysis. Consequently, it was determined that effective preprocessing of text data should consist of normalization, tokenization, removal of stop words and stemming or lemmatization. The BERT model is recommended to be used to represent the text. However, it is worth starting from the conditions of a specific task, in which alternative approaches may be preferable. The most effective methods of direct text classification are the logistic regression method, convolutional neural networks, and RoBERTa. The selection of a particular model is determined by the intended application and the technological capabilities available.

Keywords: natural language processing, NLP, text preprocessing, text representation, machine learning, neural networks

Authors' contribution
Vishnyakova A.Yu., Detkov A.A. — the concept and design of the study; *Zakharova A.V.* — processing materials, writing. All authors read and approved the final version of the article.

For citation
Zakharova AV, Vishnyakova AYu, Detkov AA. Comparison of text classification models and methods. *RUDN Journal of Engineering Research.* 2025;26(3):298–309. (In Russ.) http://doi.org/10.22363/2312-8143-2025-26-3-298-309

Введение

В современном мире объем текстовой информации растет с каждым днем, и эффективная обработка данных становится одной из ключевых задач в различных областях, от бизнеса до науки. Инструменты, позволяющие извлекать полезную информацию из неструктурированных данных, таких как текстовые документы, статьи, новости и отзывы, становятся все более востребованными. Одним из ключевых процессов, используемых для решения таких задач, является классификация текстовых данных.

Классификация текста — это процесс автоматического распределения текстовых данных по заранее определенным категориям или классам на основе их содержания. Этот процесс позволяет систематизировать информацию, облегчая ее поиск и анализ.

Процесс классификации текста обычно включает в себя три ключевых шага, каждый из которых в той или иной степени влияет на конечный результат. Подробнее эти этапы рассмотрены в теоретическом разделе данной работы.

Задачи исследования:

– формулировать, из каких шагов должен состоять оптимальный процесс предварительной обработки текстовых данных;

– определить модели представления текста, наиболее подходящие для решения задачи классификации текстовых данных;

– сравнить и выбрать наилучшие методы классификации текста.

1. Теоретические аспекты

1.1. Основные шаги предварительной обработки текстовых данных

Для работы с текстами на естественном языке (Natural Language Processing, NLP) необходима предварительная обработка, поскольку алгоритмы машинного обучения не способны работать с такими данными.

Предобработка текста, позволяющая компьютерам понимать и интерпретировать человеческий язык в полезный и значимый формат, включает в себя несколько этапов, которые могут отличаться в зависимости от задачи. Далее перечислим основные из них:

- *Нормализация* — этап, позволяющий исключить из текста шумовую информацию, в результате которого тексты приводятся к нужному регистру (зачастую к нижнему), удаляются знаки препинания, цифры и прочие небуквенные символы.
- *Токенизация* – процесс разбиения текста на отдельные элементы, называемые токенами и представляющие собой слова/фразы/символы/предложения (в зависимости от конкретной задачи и подхода к токенизации) [1].
- Удаление стоп-слов, т.е. слов, не несущих смысловой нагрузки. К ним можно отнести частицы, предлоги, союзы и т.п. Такие слова часто встречаются в текстах и обеспечивают связность предложений, однако при машинной обработке естественного языка являются шумом [2].
- *Стемминг* — метод нормализации токенов, в ходе которого от слов отсекаются префиксы, суффиксы и окончания, в результате чего выделяется основа слова [2]. Цель стемминга заключается в том, чтобы свести различные формы одного и того же слова к единой и базовой, что позволяет упростить анализ текста и уменьшить размер словаря.
- *Лемматизация* (альтернатива стемминга) — метод, основная идея которого заключатся в том, чтобы привести слова к словарной форме, называемой «лемма» [3]. Лемматизация более тонкий процесс, использующий словарь и морфологический анализ [1]. В отличие от стемминга, который просто удаляет окончания и суффиксы, лемматизация учитывает грамматические правила и контекст, чтобы определить правильную лемму для данного слова.

1.2. Ключевые модели представления данных

Текст нельзя напрямую подать на вход модели машинного обучения, его надо сначала перевести в цифровой формат. Этот шаг называют представлением текста. Современные модели опираются на эмбеддинги (или векторные представления) — способ представления слов, фраз или других объектов в виде числовых векторов в многомерном пространстве. Такой подход позволяет захватывать семантические и синтаксические свойства языка[1].

К наиболее известным моделям векторизации текста с использованием эмбеддингов, которые чаще всего встречаются в научных исследованиях по классификации текста, можно отнести следующие:

- Word2Vec (Word to Vector) — подход, основанный на численном преставлении слов, сохраняющих контекстную близость (семантическую связь). Согласно алгоритму: слова, часто встречающиеся в тексте с одинаковыми словами, имеют близкие (по косинусному расстоянию) векторы [4].
- GloVe (Global Vectors for Word Representation) — подход, опирающийся на глобальные статистики, в качестве которых берется вероятность совместного появления слов в документах. Основная его идея состоит в том, чтобы извлечь семантические отношения между словами используя матрицу совместного использования [5].

Оба подхода относят к группе методов, которые используют статические эмбеддинги, представляющие каждое слово одним и тем же

[1] Классификация документов: 7 практических подходов для небольших наборов данных // Habr: [сайт]. URL: https://habr.com/ru/articles/504744/ (дата обращения: 19.01.2025).

вектором независимо от окружения. Значительным достижением в области NLP является возможность кодирования многозначных (или полисемичных) слов разными векторами в зависимости от контекста, в котором они используются. Это достигается благодаря контекстуальным эмбеддингам, которые создаются трансформерными моделями. К таким моделям относится, например, BERT (Bidirectional Encoder Representations from Transformers) и GPT (Generative Pre-trained Transformer) [6].

1.3. Методы, используемые при классификации текстов

Методы, которые используются для решения задачи классификации текста, можно разделить на три основные группы:

1) машинные методы;

2) нейросетевые методы;

3) методы на основе трансформеров.

В табл. 1 отображены часто встречающиеся в научных исследованиях методы автоматической классификации текстовых данных.

Таблица 1

Обзор методов, используемых при классификации текста

№	Метод	Обозначение	Группа	Описание метода
1	Наивный байесовский метод	NB	1	Основывается на теореме Байеса с предположением о независимости признаков и позволяет определить класс объекта, опираясь на предшествующее распределение вероятности [4]. Хорошо работает с текстовыми данными и задачами, где требуется быстрая классификация
2	Дерево решений	DT	1	Использует древовидную структуру для принятия решений на основе значений признаков. Каждый узел дерева представляет собой вопрос о значении признака, а ветви — возможные ответы [7]. Деревья легко интерпретируемы, но могут быть подвержены переобучению
3	Случайный лес	RF	1	Строит множество деревьев решений и объединяет их результаты для улучшения точности и устойчивости модели. Использует случайную выборку подмножеств данных и признаков для создания каждого дерева, что помогает избежать переобучения и повышает обобщающую способность [8]
4	Логистическая регрессия	LR	1	Моделирует вероятность принадлежности объекта к классу, используя логистическую функцию [1]. Подходит для задач, где зависимая переменная является категориальной и позволяет интерпретировать коэффициенты как влияние предикторов на вероятность события
5	Метод k-ближайших соседей	KNN	1	Определяет класс объекта на основе классов его k-ближайших соседей в пространстве признаков [4]. Прост в реализации и не требует обучения, но может быть чувствителен к шуму и требует значительных вычислительных ресурсов при больших объемах данных
6	Метод опорных векторов	SVM	1	Ищет гиперплоскость, максимально разделяющую классы в пространстве признаков [8]. Может использовать различные ядра для обработки нелинейных данных и хорошо работает в высокоразмерных пространствах, но может быть чувствителен к выбору параметров и требует тщательной настройки
7	Многослойные перцептроны	MLP	2	Состоит из одного или нескольких скрытых слоев, которые обрабатывают входные данные и передают их на выходной слой. Каждый нейрон в слое связан с нейронами следующего слоя, и обучение происходит с использованием алгоритма обратного распространения ошибки [1]
8	Рекуррентные нейронные сети	RNN	2	Предназначен для обработки последовательных данных, таких как текст или временные ряды. RNN имеют циклические связи, что позволяет сохранять информацию о предыдущих состояниях и учитывать контекст [1]. Однако, могут страдать от проблемы затухающего градиента, что затрудняет обучение на длинных последовательностях
9	Сверточные нейронные сети	CNN	2	Обучаются распознавать признаки во входных данных с использованием сверточных слоев, после чего вычисленные признаки передаются в полносвязную сеть. Для решения задач, которые связаны с обработкой последовательностей, вместо двумерных сверток используются одномерные [1]

Окончание табл. 1

10	BERT	BERT	3	Обучается на задаче Masked Language Model (MLM), где некоторые слова в предложении замещаются масками, и модель предсказывает их [1]. Учитывает контекст как слева, так и справа от слова, что позволяет лучше понимать значение слов в зависимости от их окружения
11	SciBERT	SciBERT	3	Использует корпус научных публикаций для улучшения производительности в задачах, связанных с научной терминологией и контекстом [9]. Применяет тот же подход, что и BERT, но с учетом специфики научного языка, что делает ее особенно полезной для задач, таких как извлечение информации из научных статей, классификация и анализ текстов в области науки
12	RoBERTa	RoBERTa	3	Оптимизирует процесс предобучения BERT, используя более крупные объемы данных, более длительное время обучения и различные изменения в архитектуре. Демонстрирует лучшие результаты на многих задачах NLP по сравнению с оригинальным BERT, благодаря более эффективному обучению и более глубокому пониманию контекста [10]
13	T5	T5	3	Преобразует все задачи NLP в формат «текст в текст» — это означает, что как входные данные, так и выходные данные представлены в виде текста, что позволяет использовать одну и ту же архитектуру для различных задач, таких как перевод, суммирование, классификация и генерация текста [11]

И с т о ч н и к: выполнено А.В. Захаровой

Table 1

Overview of the methods used in text classification

№	Method	Designation	Group	Method description
1	Naive Bayes classifier	NB	1	This approach is founded on Bayes' theorem, operating under the assumption of feature independence. It facilitates the determination of an object's class based on its prior probability distribution [4]. It is well-suited for text data and tasks that require rapid classification
2	Decision tree	DT	1	It employs a tree structure to make decisions based on feature values. Each node of the tree represents a question regarding the meaning of a feature, and the branches represent possible answers [7]. Trees are easily interpretable but may be susceptible to overfitting
3	Random Forest	RF	1	It creates multiple decision trees and combines their results to improve the accuracy and stability of the model. The approach involves a random selection of subsets of data and features to create each tree, which helps to avoid overfitting and increases generalizing ability [8]
4	Logistic regression	LR	1	This approach utilizes a logistic function to simulate the probability of an object belonging to a given class [1]. This is an appropriate solution for situations in which the dependent variable is categorical. It allows for the interpretation of coefficients as the influence of predictors on the probability of an event
5	K-nearest neighbor algorithm	KNN	1	This method utilizes the classes of its k-nearest neighbors in the feature space to define the class of an object [4]. The implementation of this system is straightforward and does not require specialized training. However, it is sensitive to noise and requires substantial computing resources for effective operation, particularly when dealing with large volumes of data
6	Support vector machine	SVM	1	It looks for a hyperplane that maximizes the separation of classes in the feature space [8]. The method utilizes multiple cores to process non-linear data and functions effectively in high-dimensional spaces. However, it is susceptible to parameter selection and requires thorough configuration
7	Multilayer perceptron	MLP	2	It consists of one or more hidden layers that process the input data and transmit it to the output layer. Each neuron in a layer is connected to the neurons of the subsequent layer, and learning is enabled by the error backpropagation algorithm [1]
8	Recurrent neural network	RNN	2	It is designed to process sequential data, such as text or time series. These systems have a cyclical relationship, which allows for the storage of information about previous states and the consideration of context [1]. They may experience a decaying gradient issue, which can hinder learning effectiveness on long sequences

Ending of the Table 1

9	Convolutional neural network	CNN	2	The model is trained to recognize features in the input data using convolutional layers. The calculated features are then transmitted to a fully connected network. To solve problems related to sequence processing, one-dimensional convolutions are used instead of two-dimensional convolutions [1]
10	BERT	BERT	3	It is trained on the Masked Language Model (XML) task, where some words in a sentence are replaced by masks, and the model predicts them [1]. It considers the context both to the left and to the right of the word, which allows you to better understand the meaning of words depending on their environment
11	SciBERT	SciBERT	3	This tool utilizes the extensive collection of scientific publications to enhance productivity in tasks related to scientific terminology and context [9]. The model employs a similar approach to BERT, but it is tailored to the nuances of scientific language. This makes it particularly well-suited for tasks such as extracting information from scientific articles, classifying and analyzing texts in the scientific field
12	RoBERTa	RoBERTa	3	It optimizes the BERT retraining process by leveraging larger data sets, extended training times, and diverse architectural modifications. This model has been shown to achieve superior outcomes in numerous NLP applications when compared to the original BERT model. This enhancement can be attributed to its more effective training methods and a more profound understanding of the context [10]
13	T5	T5	3	It converts all NLP tasks into a text-to-text format, ensuring that both input and output data are presented as text. This allows for the use of a common architecture for various tasks, such as translation, summation, classification, and text generation [11]

S o u r c e: by A.V. Zakharova

2. Результаты и обсуждение

Для определения наилучшего метода были изучены научные статьи по теме автоматической классификации текста за последние 4 года. На основании результатов исследований, а именно значений показателя F1, были сформированы сводные таблицы для машинных (табл. 2), нейросетевых (табл. 3) и трансформерных методов (табл. 4).

Из табл. 2 видно, что наивысшее значение показателя F1 соответствует LR и примерно равняется 82,2 %.

Из табл. 3 видно, что наивысшее значение показателя F1 соответствует CNN и равняется примерно 85,6 %.

Таблица 2

Результаты исследований машинных методов классификации текста, %

Исследование	Метод					
	KNN	DT	NB	RF	SVM	LR
Бобина Т.С [12]	–	–	53,7	72	–	69,8
Гальченко Ю.В. [13]	–	–	80,6	–	–	89,8
Иномов Б.Б. [14]	70,3	65,6	–	63,8	–	81,8
Кусакин И.К. [15]	–	–	–	80,6	80,8	79,8
Минаев В.А. [16]	–	–	85,4	–	87,5	87,4
Мотовских Л.В. [17]	–	–	–	86	–	–
Мотовских Л.В. [18]	–	–	–	–	60	–
Плешакова Е.С., Гатауллин С.Т., Осипов А.В., Романова Е.В., Самбуров Н.С. [19]	2	70	65	75	88	77 %
Рашитов Т.Ф. [20]	–	–	–	72,7	–	–
Челышев Э.А., Оцоков Ш.А., Раскатова М.В., Щёголев П. [21]	–	–	75 %	88	–	90
Среднее значение	49,7	67,8	71,9	76,9	79,1	82,2

И с т о ч н и к: выполнено А.В. Захаровой

Table 2

Results of research on machine methods of text classification, %

Research	Method					
	KNN	DT	NB	RF	SVM	LR
Bobina T.S. [12]	–	–	53.7	72	–	69.8
Galchenko Y.V. [13]	–	–	80.6	–	–	89.8
Inomov B.B. [14]	70.3	65.6	–	63.8	–	81.8
Kusakin I.K. [15]	–	–	–	80.6	80.8	79.8
Minaev V.A. [16]	–	–	85.4	–	87.5	87.4
Motovskikh L.V. [17]	–	–	–	86	–	–
Motovskikh L.V. [18]	–	–	–	-	60	–
Pleshakova E.S., Gataullin S.T., Osipov A.V., Romanova E.V., Samburov N.S. [19]	29	70	65	75	88	77
Rashitov T.F. [20]	–	–	–	72.7	–	–
Chelyshev E.A., Otsokov S.A., Raskatova M.V., Shchegolev P. [21]	–	–	75	88	–	90
The average value	49.7	67.8	71.9	76.9	79.1	82.2

S o u r c e: by A.V. Zakharova

Таблица 3

Результаты исследований нейросетевых методов классификации текста, %

Исследование	Метод		
	MLP	RNN	CNN
Бобина Т.С [12]	70	68,2	–
Внуков И.А. [22]	78,4	75,8	82,3
Гальченко Ю.В. [13]	–	89,2	–
Иномов Б.Б. [14]	81,5	–	–
Куликов А.А. [23]	–	92,5	–
Кусакин И.К. [15]	75	84,5	–
Минаев В.А. [16]	–	89,7	89,6
Нежников Р.И. [10]	–	83,5	85
Плешакова Е.С., Гатауллин С.Т., Осипов А.В., Романова Е.В., Самбуров Н.С. [19]	79	–	–
Среднее значение	76,8	83,3	85,6

И с т о ч н и к: выполнено А.В. Захаровой

Table 3

Results of research on neural network methods of text classification, %

Research	Method		
	MLP	RNN	CNN
Bobina T.S. [12]	70	68.2	–
Nepotes I.A. [22]	78.4	75.8	82.3
Galchenko Yu.V. [13]	–	89.2	–
Inomov B B. [14]	81.5	–	–
Kulikov A.A. [23]	–	92.5	–
Kusakin I K. [15]	75	84.5	–
Minaev V.A. [16]	–	89.7	89.6
Nezhnikov R.I. [10]	–	83.5	85
Pleshakova E.S., Gataullin S.T., Osipov A.V., Romanova E.V., Samburov N.S. [19]	79	–	–
The average value	76.8	83.3	85.6

S o u r c e: by A.V. Zakharova

Таблица 4

Результаты исследований методов классификации текста на основе трансформеров, %

Исследование	Метод			
	SciBERT	BERT	T5	RoBERTa
Бондаренко В.И. [9]	71,6	64,8	–	–
Внуков И.А., Филиппов Ф.В. [22]	–	75,7	–	–
Кусакин И.К. [15]	–	86,5	–	–
Нежников Р.И. [10]	–	90	–	92,5
Прошина М.В. [11]	–	69	78	90
Среднее значение	71,6	77,2	78	91,3

И с т о ч н и к: выполнено А.В. Захаровой

Table 4

Results of research on text classification methods based on transformers, %

Research	Method			
	SciBERT	BERT	T5	RoBERTa
Bondarenko V.I. [9]	71.6	64.8	–	–
Vnukov I.A., Philippov F.V. [22]	–	75.7	–	–
Kusakin I.K. [15]	–	86.5	–	–
Nezhnikov R.I. [10]	–	90	–	92.5
Proshina M.V. [11]	–	69	78	90
The average value	71.6	77.2	78	91.3

S o u r c e: by A.V. Zakharova

Из табл. 4 видно, что наивысшее значение показателя F1 соответствует улучшенной версии модели BERT — RoBERTa и примерно равно 91,3. Однако следует отметить, что по методам на основе трансформеров было рассмотрено меньше исследований, что свидетельствует о менее точных средних значениях показателя F1.

Таким образом, если рассматривать по отдельности группы методов автоматической классификации текстовых данных, можно выделить наиболее точные:

– среди машинных методов наилучшим является логистическая регрессия ($F1_{LR} \approx 82{,}2$);

– среди нейросетевых методов наилучший — это сверточные нейронные сети ($F1_{CNN} \approx 85{,}6$);

– среди методов на основе трансформеров наилучшим является RoBERTa ($F1_{\text{RoBERTa}} \approx 91{,}3$).

Заключение

На основании изученных научных исследований по теме автоматической классификации текста можно сделать следующие выводы:

1. Оптимальный процесс предобработки данных для решения задачи классификации текста состоит из нескольких шагов и включает в себя как минимум нормализацию, токенизацию, удаление стоп-слов и стемминг/лемматизацию. Однако в зависимости от конкретной задачи эти шаги могут быть скорректированы.

2. Для перевода текста в цифровой формат лучше выбирать BERT, так как эта модель использует двунаправленный контекст и демонстрирует высокую точность в различных задачах NLP. Если ресурсы ограничены, можно рассматривать GPT как альтернативу, но эта модель больше ориентирована на генерацию текста, поэтому менее эффективна для задач, требующих глубокого понимания контекста. Word2Vec и GloVe могут быть использованы в менее сложных задачах, где контекст не так критичен.

3. Наилучшими методами автоматической классификации текста являются метод логистической регрессии, сверточные нейронные сети и RoBERTa. Точность каждого из них превышает 82, что является достаточно хорошим

результатом, поэтому все эти методы могут рассматриваться при решении уже конкретных задач:

– LR больше подходит для простых задач и небольших наборов данных;

– CNN и RoBERTa могут использоваться для сложных текстов и больших наборов данных, выбор зависит от доступности ресурсов и требований к точности.

Список литературы

1. *Логунова Т.В., Щербакова Л.В., Васюков В.М., Шимкун В.В.* Анализ алгоритмов классификации текстов // Universum: технические науки. 2023. № 2 (107). С. 4–20. https://doi.org/10.32743/UniTech.2023.107.2.15064 EDN: MYDAJG

2. *Челышев Э.А., Оцоков Ш.А., Раскатова М.В.* Автоматическая рубрикация текстов с использованием алгоритмов машинного обучения // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2021. № 4. С. 175–182. https://doi.org/10.18137/RNU.V9187.21.04.P.175 EDN: SBCVLA

3. *Акжолов Р.К., Верига А.В.* Предобработка текста для решения задач NLP // НИЦ Вестник науки. 2020. № 3 (24) Т. 1. С. 66–68. EDN: KCGMUZ

4. *Максютин П.А., Шульженко С.Н.* Обзор методов классификации текстов с помощью машинного обучения // Инженерный вестник Дона. 2022. № 12 (96). С. 1–9. EDN: USWOAI

5. *Pennington J., Socher R., Manning D.* Christopher. GloVe: Global Vectors for Word Representation. URL: https://nlp.stanford.edu/pubs/glove.pdf с.3 (Дата обращения: 20.01.2025)

6. *Жусип М.Н., Жаксыбаев Д.О.* Сравнение чатботов с использованием трансформеров и нейросетей: исследование применения архитектур GPT и BERT // НИЦ Вестник науки. 2024. № 9 (78) Т. 2. С. 287–290. EDN: DEXNMS

7. *Батура Т.В.* Методы автоматической классификации текстов // Международный журнал «Программные продукты и системы». 2017. Т. 30. № 1. С. 85–99. EDN: ZDUXCL

8. *Буйлова Н.Н.* Классификация текстов по жанрам с помощью алгоритмов машинного обучения // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2018. № 8. С. 34–38. EDN: XYBWQP

9. *Бондаренко В.И.* Классификация научных текстов с помощью методов глубокого машинного обучения // Вестник Донецкого национального университета. Серия Г: Технические науки. 2021. № 3. С. 69–77. EDN: FJPQFE

10. *Нежников Р.И., Марьенков А.Н.* Сравнительный анализ моделей трансформера для классификации неструктурированной текстовой информации // Прикаспийский журнал: управление и высокие технологии. 2024. № 2 (66). С. 32–38. EDN: LREEXX

11. *Прошина М.В., Виноградов А.Н.* Анализ эффективности трансформеров для решения некоторых задач NLP // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем : материалы Всероссийской конференции с международным участием, Москва, 17–21 апреля 2023 года. Москва: Российский университет дружбы народов (РУДН), 2023. С. 153–157. EDN: RXMCCJ

12. *Бобина Т.С.* Автоматическая классификация текста при помощи методов машинного обучения и нейронных сетей // Современные информационные технологии в образовании, науке и промышленности : сборник трудов. XXVIII Международная конференция. XXVI Международный конкурс научных и научно-методических работ. Всероссийский конкурс проектов «Научное творческое сообщество», Мытищи, Москва, 25–26 апреля 2024 года. Москва : Экон-Информ, 2024. С. 253–258. EDN: PMVIHF

13. *Гальченко Ю.В., Нестеров С.А.* Классификация текстов по тональности методами машинного обучения // Системный анализ в проектировании и управлении : сборник научных трудов XXVI Международной научно-практической конференции : в 3 частях. Санкт-Петербург, 13–14 октября 2022 года. Т. Часть 3. Санкт-Петербург : Санкт-Петербургский политехнический университет Петра Великого, 2023. С. 369–378. https://doi.org/10.18720/SPBPU/2/id23-501 EDN: YURQCU

14. *Иномов Б.Б., Тропманн-Фрик М.* Классификация научных текстов по специальностям методами машинного обучения // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. Т. 20. № 2. С. 27–36. https://doi.org/10.25205/1818-7900-2022-20-2-27-36 EDN: ORMRCL

15. *Кусакин И.К., Федорец О.В., Романов А.Ю.* Исследование методов машинного обучения для классификации научных текстов на русском языке // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2022. № 12. С. 6–9. https://doi.org/10.36535/0548-0027-2022-12-2 EDN: EPASJQ

16. *Минаев В.А. Поликарпов Е.С., Симонов А.В.* Применение глубинных нейронных сетей для выявления деструктивного контента в социальных медиа // Информация и безопасность. 2021. Т. 24. № 3. С. 361–372 https://doi.org/10.36622/VSTU.2021.24.3.004 EDN: IMHBIG

17. *Мотовских Л.В.* Классификация медиатекстов с использованием машинного обучения // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2020. № 12 (841). С. 124–130. EDN: YZFGJN

18. *Мотовских Л.В.* Автоматическая классификация текстов различных СМИ // Collegium Linguisticum–2021 : сборник научных статей ежегодной конференции Студенческого научного общества МГЛУ, Москва, 17–19 марта 2021 года. Москва : Московский государственный лингвистический университет, 2021. С. 83–88. EDN: LGMFHK

19. *Плешакова Е.С., Гатауллин С.Т., Осипов А.В., Романова Е.В., Самбуров Н.С.* Эффективная классификация текстов на естественном языке и определение тональности речи с использованием выбранных методов машинного обучения // Вопросы безопасности. 2022. № 4. С. 1–14. https://doi.org/10.25136/2409-7543.2022.4.38658 EDN: UPWMCV

20. *Рашитов Т.Ф., Квасов М.Н.* Использование метода машинного обучения «Случайный лес» для классификации текстов по рубрикам // Состояние и перспективы развития современной науки по направлению «АСУ, информационно-телекоммуникационные системы» : сборник статей III Всероссийской научно-технической конференции, Анапа, 22–23 апреля 2021 года. Том 2. Анапа : Федеральное государственное автономное учреждение «Военный инновационный технополис "ЭРА"», 2021. С. 76–78. EDN: QTEYUB

21. *Челышев Э.А., Оцоков Ш.А., Раскатова М.В., Щёголев П.* Сравнение методов классификации русскоязычных новостных текстов с использованием алгоритмов машинного обучения // Вестник кибернетики. 2022. № 1 (45). С. 63–71. https://doi.org/10.34822/1999-7604-2022-1-63-71 EDN: VHTYBB

22. *Внуков И.А., Филиппов Ф.В.* Средства глубокого обучения для классификации новостных текстов в интеллектуальных рекомендательных системах // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2024) : материалы XIII Международной научно-технической и научно-методической конференции, Санкт-Петербург, 27–28 февраля 2024 года. Санкт-Петербург : Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, 2024. С. 190–194. EDN: EWPVOP

23. *Куликов А.А., Маилян Э.К.* Сравнение архитектур рекуррентных нейронных сетей в задаче бинарной классификации текстов // Инновационное развитие техники и технологий в промышленности (ИНТЕКС-2021) : сборник материалов Всероссийской научной конференции молодых исследователей с международным участием, Москва, 12–15 апреля 2021 года. Часть 3. Москва : Российский государственный университет имени А.Н. Косыгина (Технологии. Дизайн. Искусство), 2021. С. 223–226. EDN: XQKUHP

References

1. Logunova TV, Shcherbakova LV, Vasyukov VM, Shimkun VV. Analysis textus classificationis algorithmorum. *Universi: scientiarum technicarum: electronic scientiae acta.* 2023;(2):4–20. (In Russ.) https://doi.org/10.32743/UniTech.2023.107.2.15064 EDN: MYDAJG

2. Chelyshev EA, Otsokov SA, Raskatova MV. Automatic textus rubricationis utens machina algorithms discendi. *Bulletin Novae universitatis russicae. Series: Systemata Complexa: exempla, analysis et administratio.* 2021;(4):175–182. (In Russ.) https://doi.org/10.18137/RNU.V9187.21.04.P.175 EDN: SBCVLA

3. Akzholov RK, Veriga AV. Textus praeprocessing ad SOLVENDAS NLP difficultates. *Sic Bulletin Scientiae*. 2020;(3):66–68. (In Russ.) EDN: KCGMUZ

4. Maksyutin PA, Shulzhenko SN. Recensio textuum methodorum classificationis utens machina discendi. *Ipsum Bulletin De Don*. 2022;(12):1–9. (In Russ.) EDN: USWOAI

5. Pennington J, Socher R, Manning D. *Christopher. GloVe: Global Vectors for Word Representation*. Available from: https://nlp.stanford.edu/pubs/glove.pdf c.3 (accessed: 20.01.2025)

6. Zhusip MN, Zhaksybaev DO. Comparatio chatbotorum utens transformatoribus et reticulis neuralis: studium applicationis GPT et BERT architecturae. *Sic Bulletin Scientiae*. 2024;(9):287–290. (In Russ.) EDN: DEXNMS

7. Batura TV. Methodi textus classificationis latae. *Acta Internationalis Productorum Et Systematum Programmatum.* 2017;30(1):85–99. (In Russ.) EDN: ZDUXCL

8. Bulova NN. Classificatio textuum per genus machinae algorithmorum discendi utens. *Scientifica et technica notitia 2: Processiones Et dispositiones*. 2018;(8):34–38. (In Russ.) EDN: XYBWQP

9. Bondarenko VI. Classificatio textuum scientificorum utens machinae altae methodi discendi. *Bulletin Universitatis Nationalis Donetsk. Series G: Scientiarum Technicarum*. 2021;(3):69–77. EDN: FJPQFE

10. Nezhnikov RI, Marienkov AN. Analysis Comparativa transformatoris exemplorum pro classificatione informationis textualis. *Acta Caspiae: Administratio et Technologiae Altae*. 2024;(2):32–38. (In Russ.) EDN: LREEXX

11. Proshina MV, Vinogradov AN. Analysis efficaciae transformatorum ad solvendas QUASDAM DIFFICULTATES NLP. *Informationes et technologiae telecommunicationis et mathematicae exemplaris systematum summus technicorum: acta Colloquii Omnium russorum Cum Participatione Internationali, Moscow,*

17–21 aprilis 2023. Moscow: RUDN University; 2023; 153–157. (In Russ.) EDN: RXMCCJ

12. Bobina TS. Automatic textus Classificationis utens machinae methodi discendi et retiacula neuralis. *Modernae informationes technologiae in educatione, scientia et industria: Acta. Colloquium Internationale 28th. 26th Competition internationalis operum scientificorum et emendatiorum. Omnes-Russian Project Competition "Communitas Creatrix Scientifica" Mytishchi, Moscow, 25–26 Aprilis, 2024*. Moscow*:* Limitata Rusticis Company "Ekon-Certiorem Libellorum Domus;" 2024:253–258. (In Russ.) EDN: PMVIHF

13. Galchenko YV, Nesterov SA. Classificatio textuum per tonalitatem machinae methodi discendi. *Systematis analysi in consilio et administratione: acta 26th Conferentiae Scientificae Et Practicae Internationalis. Ad 3 a. m., Saint Petersburg, 13–14 octobris, 2022. Part 3.* Saint Petersburg: Petrus Magnus S. Petersburg Universitas Polytechnica; 2023. P. 369–378. (In Russ.) https://doi.org/10.18720/SPBPU/2/id23-501 EDN: YURQCU

14. Inomov BB, Tropmann-Frick M. Classificatio textuum scientificorum a propriis utens machinae methodi discendi. *Bulletin Novosibirsk Universitatis Publicae. Series: Informationis Technicae*. 2022;(2):27–36. (In Russ.) https://doi.org/10.25205/1818-7900-2022-20-2-27-36 EDN: ORMRCL

15. Kusakin IK, Fedorets OV, Romanov AY. Investigatio machinae methodi discendi ad digerendos textus scientificos in Notitia russica. *Scientifica et Technica. 2: Processiones Et dispositiones*. 2022;(12):6–9. (In Russ.) https://doi.org/10.36535/0548-0027-2022-12-2 EDN: EPASJQ

16. Minaev VA, Polikarpov ES, Simonov AV. Usus reticulorum neuralium profundorum ad cognoscendum contentum perniciosum in instrumentis socialibus. *Informationibus et Securitate*. 2021;(3):361–372. (In Russ.) https://doi.org/10.36622/VSTU.2021.24.3.004 EDN: IMHBIG

17. Motovskikh LV. Classificatio textuum instrumentorum utens machina discendi. *Bulletin Universitatis Linguisticae Civitatis Moscuae. Humanas*. 2020;(12): 124–130. (In Russ.) EDN: YZFGJN

18. Motovskikh LV. Classificatio Latae textuum variarum instrumentorum. *Collegium Linguisticum-2021: Collectio articulorum scientificorum annui conferentiae Mglu Studentium Societatis Scientificae, Moscow, martii 17–19, 2021*. Moscow: State Linguistic University; 2021: 83–88. (In Russ.) EDN: LGMFHK

19. Pleshakova ES, Gataullin ST, Osipov AV, Romanova EV, Samburov NS. Efficax classificatio textuum in lingua naturali et determinatio loquelae tonality utens delectae machinae discendi methodos. *Quaestiones Securitatis*. 2022;(4):1–14. (In Russ.) https://doi.org/10.25136/2409-7543.2022.4.38658 EDN: UPWMCV

20. Rashitov TF, Kvasov MN. Usus machinae "Temere Silvae" methodus discendi ad textus digerendos per capita. *Statum et spem evolutionis scientiae modernae in agro automated systemata moderandi, informationes et systemata telecommunicationis: Collectio articulorum III Conferentiae Scientificae et Technicae Omnes-russicae, Anapa, 22–23 aprilis 2021. 2 volumine*. Anapa: Status Foederalis Institutio Sui Iuris "Innovatio Militaris Technopolis ERA." 2021;76–78. (In Russ.) EDN: QTEYUB

21. Chelyshev EA, Otsokov SA, Raskatova MV, Shchegolev P. Comparatio methodorum classificationis de textibus nuntiorum russorum linguarum utentes machinae algorithmorum discendi. *Bulletin Cyberneticorum*. 2022; (1):63–71. (In Russ.) https://doi.org/10.34822/1999-7604-2022-1-63-71 EDN: VHTYBB

22. Vnukov IA, Philippov FV. Alta discendi instrumenta ad digerendos nuntios textus in intelligentibus commendatione systemata. *Actualia problemata communicationum infotelec in scientia et educatione (APINO 2024): Acta Xiii Conferentiae Scientificae Internationalis, Technicae Et Scientificae Methodologicae, Saint Petersburg, 27–28 februarii, 2024*. Saint Petersburg: Universitas Civitatis S. Petersburg Telecommunicationum ex nomine nuncupatur Professor M.A. Bonch-Bruevich. 2024: 190–194. (In Russ.) EDN: EWPVOP

23. Kulikov AA, Mailyan E.K. Comparatio architecturae retis neuralis recurrentis in problemate textus classificationis binarii. *Innovative evolutionis machinationis et technologiae in industria (INTEX-2021): Acta Omnium-russorum Conferentiae Scientificae Inquisitorum Iuvenum Cum Participatione Internationali, Moscow, 12–15 aprilis 2021. Pars 3 Volumine*. Moscow: A.N. Kosygin Universitas Civitatis russicae (Technologia. Consilio. Ars). (In Russ.) EDN: XQKUHP

Сведения об авторах

Захарова Ангелина Валерьевна, магистрант кафедры анализа систем и принятия решений, Уральский федеральный университет им. Первого Президента России Б.Н. Ельцина (УрФУ), Российская Федерация, 620062, г. Екатеринбург, ул. Мира, д. 19; eLIBRARY SPIN-код: 6278-8518, ORCID: 0009-0007-9651-4530; e-mail: zakharova.linusha@mail.ru

Вишнякова Алина Юрьевна, старший преподаватель, аспирант кафедры анализа систем и принятия решений, Уральский федеральный университет им. Первого Президента России Б.Н. Ельцина (УрФУ), Российская Федерация, 620062, г. Екатеринбург, ул. Мира, д. 19; eLIBRARY SPIN-код: 5641-6945, ORCID: 0000-0003-1649-4167; e-mail: alina.vishniakova@urfu.ru

Детков Александр Александрович, кандидат экономических наук, доцент кафедры анализа систем и принятия решений, Уральский федеральный университет им. Первого Президента России Б.Н. Ельцина (УрФУ), Российская Федерация, 620062, г. Екатеринбург, ул. Мира, д.19; eLIBRARY SPIN-код: 5310-3027, ORCID: 0009-0003-3958-3549; e-mail: a.a.detkov@urfu.ru

About the authors

Angelina V. Zakharova, Master's student of the Department of Systems Analysis and Decision-making, Ural Federal University named after the first President of Russia B.N. Yeltsin, 19 Mira St, Yekaterinburg, 620062, Russian Federation; eLIBRARY SPIN-code: 6278-8518, ORCID: 0009-0007-9651-4530; e-mail: zakharova.linusha@mail.ru

Alina Yu. Vishnyakova, Senior Lecturer, Postgraduate Student of the Department of Systems Analysis and Decision-making, Ural Federal University named after the first President of Russia B.N. Yeltsin, 19 Mira St, Yekaterinburg, 620062, Russian Federation; eLIBRARY SPIN-code: 5641-6945, ORCID: 0000-0003-1649-4167; e-mail: alina.vishniakova@urfu.ru

Alexander A. Detkov, PhD in Economics, Associate Professor of the Department of Systems Analysis and Decision-Making, Ural Federal University named after the first President of Russia B.N. Yeltsin, 19 Mira St, Yekaterinburg, 620062, Russian Federation; eLIBRARY SPIN-code: 5310-3027, ORCID: 0009-0003-3958-3549; e-mail: a.a.detkov@urfu.ru