2025;10(1):32–49 Научный редактор и издатель / Science Editor and Publisher

РЕЦЕНЗИРОВАНИЕ / PEER REVIEW

Оригинальная статья / Original paper

https://doi.org/10.24069/SEP-25-35

Определение рецензента методами машинного обучения

Д.Ю. Большаков

Концерн воздушно-космической обороны «Алмаз – Антей», г. Москва, Российская Федерация

✉ *press@almaz-antey.ru*

Резюме. Рассматривается задача автоматического назначения рецензентов на основе исторических данных о ранее поступивших и прорецензированных рукописях. В традиционной редакционной практике подбор экспертов опирается на субъективные решения редактора, что может приводить к задержкам и снижению качества экспертизы. Цель исследования – продемонстрировать, что использование простых моделей обработки естественного языка позволяет эффективно и прозрачно автоматизировать этот процесс. В качестве исходных данных использованы тексты опубликованных и отклоненных рукописей научно-технического журнала «Вестник Концерна ВКО «Алмаз – Антей» (с 2011 по 2024 г.), сопровожденные информацией о назначенных рецензентах. Методологически подход основан на предварительной лемматизации текстов, удалении стоп-слов и знаков пунктуации, а также последующей векторизации с использованием моделей *bag-of-words* (*BoW*) и *Term Frequency-Inverse Document Frequency* (*TF-IDF*). Близость текстов оценивалось путем вычисления максимального косинусного расстояния между их векторными представлениями. Предполагается, что статья, прорецензированная ранее и демонстрирующая наибольшую близость к поступившей, была рассмотрена рецензентами, которых система может рекомендовать для оценки новой рукописи. Результаты показывают, что простые частотные модели (*BoW*, *TF-IDF*) демонстрируют более высокую точность назначения рецензентов (до 99 %) по сравнению с нейросетевыми подходами (например, моделью *Doc2Vec*), особенно при дополнении графом связей между экспертами. При этом модель остается интерпретируемой, не требует значительных вычислительных ресурсов и может быть реализована на компьютере офисного уровня. Показано, что модель эффективно работает в условиях дисбаланса классов и применима даже к относительно небольшим корпусам, начиная от 30 статей. Однако ее обобщение на мультижурнальные редакции требует локальной адаптации, а для решения задачи прогнозирования вероятности принятия к публикации необходимо существенно увеличить объем выборки и привлечь модели глубокого обучения. Предложенный подход может быть легко интегрирован в цифровые редакционные системы для сокращения времени принятия решений, повышения прозрачности экспертизы и снижения нагрузки на сотрудников журнала.

Ключевые слова: компьютерная лингвистика, косинусное расстояние, мешок слов, *TF-IDF*, машинное обучение, лемматизация, регулярные выражения, обработка естественного языка

Для цитирования: Большаков Д.Ю. Определение рецензента методами машинного обучения. *Научный редактор и издатель*. 2025;10(1):32–49. https://doi.org/10.24069/SEP-25-35

A reviewer identification using machine learning methods

D.Yu. Bolshakov

Joint Stock Company "Almaz – Antey" Air and Space Defence Corporation, Moscow, Russian Federation

✉ *press@almaz-antey.ru*

Abstract. This article addresses the task of automatically assigning peer reviewers based on historical data from previously submitted and reviewed manuscripts. In conventional editorial practice, reviewer selection relies heavily on the subjective judgment of editors, which can lead to delays and inconsistencies in the quality of expert evaluation. The purpose of this study is to demonstrate that simple natural language

processing (*NLP*) models can be used to automate this process in an efficient and transparent manner. The dataset used in this research consists of both published and rejected articles submitted to the Almaz-Antey Air and Space Defense Corporation Journal, enriched with information about the reviewers assigned to each manuscript. Methodologically, the approach relies on basic text preprocessing, including lemmatization, removal of stop words and punctuation, followed by vectorization using *bag-of-words* (*BoW*) and *Term Frequency-Inverse Document Frequency* (*TF-IDF*) models. Text similarity is calculated via cosine distance between vectorized representations. The core assumption is that a newly submitted manuscript is most similar to an already reviewed one and, therefore, can be assigned to the same reviewers. The results indicate that simple frequency-based models (*BoW, TF-IDF*) achieve higher accuracy in reviewer assignment (up to 99%) compared to neural network approaches such as Doc2Vec, especially when enhanced with a reviewer co-review graph. The proposed method remains interpretable, requires minimal computational resources, and is fully compatible with office-level computing environments. The model has been shown to perform reliably under class imbalance and is applicable even to relatively small datasets, starting from around 30 manuscripts. However, its generalization to multi-journal editorial systems would require local adaptation, and the task of predicting publication outcomes calls for significantly larger corpora and the use of deep learning architectures. This approach can be seamlessly integrated into digital editorial platforms, contributing to faster decision-making, increased transparency in peer review, and reduced workload for journal staff.

Keywords: computational linguistics, cosine distance, bag-of-words model, TF-IDF model, machine learning, lemmatization, regular expressions, natural language processing

For citation: Bolshakov D. Yu. A reviewer identification using machine learning methods. *Science Editor and Publisher*. 2025;10(1):32–49. (In Russ.) https://doi.org/10.24069/SEP-25-35

ВВЕДЕНИЕ

Вопрос о способности машины к интеллектуальному поведению впервые был системно сформулирован в 1950 г. в статье Алана Тьюринга *Computing Machinery and Intelligence*, где автор предложил мысленный эксперимент, вошедший в историю как тест Тьюринга [1]. В этой работе основное внимание уделено исследованию возможности машинной обработки текстовой информации с целью приближения результатов к человеческому мышлению. Такой подход был сосредоточен исключительно на тексте, поскольку в середине XX в. ни синтез естественной речи, ни тем более генерация видео не были технически осуществимы.

Современное направление обработки естественного языка (*Natural Language Processing*, *NLP*) сформировалось на пересечении математической лингвистики, машинного обучения и искусственного интеллекта. Активное развитие обработки естественного языка в последние два десятилетия обусловлено прежде всего ростом вычислительных мощностей, позволившим использовать ресурсоемкие алгоритмы машинного обучения. Так, работы Й. Голдберга (Y. Goldberg) [2] и Д. Девлина с соавт. (J. Devlin et al.) [3] продемонстрировали потенциал нейросетевых моделей, включая трансформеры, в решении задач языкового моделирования и текстовой классификации. Особенно значимый вклад в повышение точности обработки текстов внесла предложенная Д. Девлином в работе [3] архитектура BERT, которая положила начало широкому внедрению предобученных моделей в различные приложения *NLP*. Параллельно развивались методы, ориентированные на интерпретируемость, как, например, в монографии С. Чжу с соавт. (X. Zhu et al.), где рассматриваются компромиссы между точностью и прозрачностью алгоритмов [4]. Систематический обзор, представленный Ц. Цзя и соавт. (J. Jia et al.) [5], подтвердил эффективность гибридных подходов, сочетающих глубокое обучение и традиционные методы текстовой векторизации в задачах классификации и извлечения информации. Фундаментальные основы *NLP* были заложены в работе Д. Журавски и Дж. Мартина (D. Martin & J. Jurafsky) [6], которая остается актуальной благодаря регулярным переизданиям. Несмотря на преобладание переводной литературы в области NLP, имеются и значимые российские исследования, в том числе коллективная монография сотрудников НИУ ВШЭ [7].

Одной из узкоспециализированных, но практически значимых задач в области обработки естественного языка является автоматическое назначение рецензента на основе анализа текста поступившей рукописи. В традиционной редакционной практике это решение принимает человек, что при его загрузке, отсутствии средств связи или их ограниченной доступности может привести к задержкам в публикации и неравномерному распределению экспертной нагрузки. Автомати-

зация подобного выбора может существенно повысить скорость, объективность и прозрачность процесса рецензирования, особенно в технических журналах с высоким объемом подачи.

Несмотря на то что ведущие международные издательства уже внедрили цифровые инструменты поддержки редактора, точные алгоритмические механизмы их работы не раскрываются. Так, в системе Springer Nature с января 2025 г. функционирует комплексный модуль экспертной проверки, охватывающий 14 параметров оценки рукописи, включая выбор рецензента. Association for Computing Machinery (ACM) Submission System предлагает рекомендации на основе профиля исследователя и ключевых слов, IEEE Xplore сочетает анализ данных с проверкой текстовых совпадений (CrossCheck), а Elsevier использует платформу Editorial Manager, автоматически сопоставляющую текст рукописи с профилем потенциального эксперта. Однако данные о том, используют ли эти решения, например, метод *k*-ближайших соседей, метод опорных векторов или модели глубокого обучения, отсутствуют.

Анализ недавних научных публикаций подтверждает растущий интерес к задаче автоматизированного подбора рецензентов. В исследовании С. Бхаттачарья и соавт. (S. Bhattacharya et al.) [8] рассматривается применение пяти методов машинного обучения, включая логистическую регрессию и метод опорных векторов (*Support Vector Machine*, *SVM)*. Однако в их работе используется корпус[1] статей из разных журналов и не учитывается, что к одному классу может быть отнесено разное количество статей (так называемый классовый дисбаланс, характерный для реальных редакций). Работа Ш. Тан и соавт. (S. Tan et al.) [9] предлагает семантический подход с попыткой реконструкции связей между рецензентами и статьями, но без учета специфики отдельных изданий. В исследовании А. Адебийи с соавт. (A. Adebiyi et al.) [10] делается вывод о преимуществе метода *TF-IDF* в сравнении с *LSI (Latent Semantic Indexing)* и *LDA* (*Latent Dirichlet Allocation)*, однако оно основано исключительно на данных конференций и игнорирует внутренние редакционные массивы отклоненных рукописей. В то же время О. Анжум (O. Anjum) с соавт. [11] критически оценивают модель *bag-of-words* (*BoW*) при работе с широкими корпусами, однако эта оценка может быть необоснованной в случае специализированных тематических журналов. Работа Х. Пэна и соавт. (H. Peng et al.) [12] учитывает изменение исследовательского профиля рецензента во времени, но не задействует внутренние данные редакций и графовые модели взаимодействий. Вопросам доверия к большим языковым моделям при проведении рецензирования посвящены недавние работы китайских и корейских коллег Ч. Ли и соавт. (C. Li et al.) [13], В. Ляна и соавт. (W. Liang et al.) [14], Д. Ли и соавт. (J. Lee et al.) [15], которые делают попытки переложить труд рецензента на большие нейросетевые модели.

Обобщение этих подходов показывает, что существующие решения либо требуют больших и неоднородных корпусов, либо игнорируют локальные условия редакций, либо исключают внутреннюю информацию о ходе рецензирования. В табл. 1 систематизированы основные различия между существующими подходами и требованиями, вытекающими из реальной практики.

[1] Тексты, собранные для анализа методами *NLP*, называются корпусом.

Таблица 1. Сравнительный анализ подходов разных исследователей к автоматизированному выбору рецензентов

Table 1. Comparison of the studied methods and the proposed approaches of different researchers to the automated selection of reviewers

Критерии сравнения	С. Бхаттачарья [8]	Ш. Тан [9]	А. Адебийи [10]	О. Анжум [11]	Автор данной работы
Минимальное количество статей	> 100	> 100	> 100	> 1000	> 2
Проблема дисбаланса классов	Да	Нет	Нет	Нет	Нет
Учет связей между рецензентами	Нет	Нет	Нет	Нет	Да
Интерпретируемость результатов	Да	Нет	Да	Да	Да
Автоматический подбор рецензента	Да	Нет	Да	Да	Да
Наличие опубликованных статей рецензента	Да	Да	Да	Да	Нет
Ограничение использования только одним журналом	Нет	Нет	Нет	Нет	Да

Цель исследования – разработка прозрачной, интерпретируемой и вычислительно доступной модели автоматического подбора рецензента, адаптированной для работы в рамках отдельной редакции и обеспечивающей эффективное использование как опубликованных, так и отклоненных рукописей, а также данных о связях между экспертами. Описание архитектуры предложенной модели и ее характеристик приводится в следующих разделах.

МАТЕРИАЛЫ И МЕТОДЫ

Исходные данные

Алгоритмы NLP применяются к исходному массиву данных, представленному рукописями, поступившими в редакцию научно-технического журнала «Вестник Концерна ВКО «Алмаз – Антей» с 2011 по 2024 г. (табл. 2). Для анализа сформировано два корпуса текстов, репрезентирующих основные научные области журнала, – «Радиолокация» и «Газодинамика». Всего отобрано 260 уникальных опубликованных статей и неопубликованных рукописей (206 по радиолокации и 54 по газодинамике), при этом в журнал с 2011 по 2024 г. поступило 1079 рукописей, из которых 485 опубликовано (~45 %).

Статьи для исследования собирались вручную посредством включения в корпус всех поступивших в журнал статей. Все статьи формата .doc, .docx, indd (Adobe InDesign) переводили в формат TXT для ввода в модель. Дисбаланс классов в разделе «Радиолокация» составил 58 неопубликованных статей против 148 опубликованных, для раздела «Газодинамика» – 9 неопубликованных против 45 опубликованных статей.

Для проверки работоспособности алгоритмов дополнительно использовался небольшой корпус рукописей по наукометрии (12 статей, из них 9 опубликованных). Тексты этого корпуса являются рукописями автора по тематике издания и продвижения научных журналов, опубликованными и не опубликованными в разных изданиях с 2020 по 2024 г.

Теоретическое обоснование

Для сравнения текстов их следует привести к виду, который подразумевает математическую оценку каждого текста и их сопоставление между собой [16]. В *NLP* такая оценка проводится векторизацией, то есть приведением текстов к числовому виду [17]. Однако процедуру векторизации предваряют несколько процедур приведения текстов к виду, пригодному для векторизации: лемматизация, удаление стоп-слов, удаление знаков пунктуации, применение регулярных выражений [18]. Рассмотрим кратко каждый из этих этапов.

Лемматизация

Первым этапом обработки всех текстов в настоящем исследовании выступает лемматизация – приведение каждого слова текста к нормальной форме [17]. Нормальной формой для существительного и прилагательного считается форма именительного падежа, единственного числа, для прилагательного – мужского рода, для глаголов, причастий и деепричастий – неопределенная форма глагола (инфинитив) несовершенного вида. Например, фраза

«Ехал Ваня на коне, попевая весёлую песенку»

преобразуется в

«**е**ха**ть в**аня на кон**ь,** попева**ть** весёл**ый** песенк**а**».

Следует отметить, что важным аспектом лемматизации является приведение всех букв к единому регистру. Для буквенных языков стандартная практика – использование нижнего регистра, поэтому все прописные буквы преобразуются в вышеприведенном примере в строчные.

Лемматизация позволяет унифицировать лексические формы для последующего сравнения. Например, сведенное к единой форме в результате лемматизации словосочетание «веселый песенка» в различных грамматических формах (единственное и множественное число, шесть падежей, два рода) может иметь до 24 вариантов написания.

Таблица 2. Характеристики исследуемых корпусов текстов
Table 2. Text corpora used and their characteristics

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем (в текстах)	206	54	12
Объем (в символах)	3 704 604	825 349	344 462
Средняя длина текста (в символах)	17 984	15 573	28 704

Удаление стоп-слов

Стоп-слова – это часто встречающиеся слова в языке, которые обычно не несут полезной информации (предлоги, частицы, междометия и т.д.) [17]. Список стоп-слов есть в любом ориентированном на *NLP* инструменте обработки текстов. Следует также отметить, что стоп-слова в целом по значению похожи в разных языках, но по количеству значительно различаются. Например, в русском языке их около 150 («на», «что», «как» и т.д.), в английском около 200 («as», «at», «of» и т.д.), а в китайском более 800 («个», «上下», «之一» и т.д.)[2].

В вышеприведенном в качестве примера предложении предлог «на» должен быть удален, чтобы оно было приведено к форме

«**е**ха**ть в**аня кон**ь,** попева**ть** весел**ый** песенк**а**».

Как видно из полученной фразы, смысловая ясность конструкции утрачивается, поскольку второе и третье слово могут быть неоднозначно интерпретированы из-за отсутствия предлога.

Удаление пунктуации

Для выделения специальной лексики в тексте следует исключить также знаки пунктуации [17], так как очевидно, что их частотность в тексте существенно превышает частоту употребления узкоспециализированных терминов.

Применение регулярных выражений

Регулярные выражения представляют собой формальный язык, который используется для поиска и обработки текстовых данных [17]. Другими словами, это способ поиска и замены, более эффективный, чем стандартный. Например, во всем тексте одной командой можно удалить все пунктуационные знаки. В настоящем исследовании регулярные выражения использовались для очистки текстов рукописей от специальных символов, таких как неразрывные пробелы или тире, а также для удаления списков литературы и служебной информации (например, DOI и/или УДК).

Векторизация

После предварительной обработки (лемматизации, удаления стоп-слов и пунктуационных знаков) текст целиком преобразуется в числовое представление методом векторизации, когда каждое слово из текста заменяется числовым вектором в заданном векторном пространстве, где каждому слову соответствует уникальный цифровой код.

Векторизацию проще всего рассмотреть на примерах. Пусть будет три предложения:

«Мама мыла раму»,
«Рама мыта мамой»,
«Мама мамы мыла раму».

Количество уникальных слов с учетом морфологических признаков (падеж, род и форма глагола) равно семи (мама, мамой, мамы, мыла, мыта, рама, раму). Однако после лемматизации остается только три леммы (мама, мыть, рама):

«мама мыть рама»,
«рама мыть мама»,
«мама мама мыть рама».

Если представить, что каждое слово из трех (мама, рама, мыть) – это ось, а частота этих слов – числовые значения на осях, координаты в трехмерном пространстве, то вышеуказанные предложения трансформируются в числовые векторы, показанные в табл. 3.

Предложение «мама мыть рама» преобразуется в числовой вектор (1, 1, 1), а предложение «мама мама мыть рама» – в числовой вектор (2, 1, 1). При векторизации очень важно, чтобы все векторы имели одинаковую размерность, иначе их сравнение будет некорректным. Делать это вручную чрезвычайно долго, поэтому во всех инструментах, связанных с *NLP*, применяются соответствующие функции для осуществления лемматизации и подсчета во всем корпусе текстов частоты появления каждой леммы.

Рассмотренный выше простой метод (примененный к данным табл. 3) основан на подсчете количества появлений каждого уникального слова в тексте и расстановки полученных значений на соответствующих позициях вектора. Такой метод называется «мешок слов» (*BoW*) [17].

Таблица 3. Координаты лемматизированных предложений, представленных в виде векторов, в трехмерном пространстве (мама, мыть, рама)
Table 3. Lemmatized sentences and their coordinates in a three-word vector space (“мама”, “мыть”, “рама”)

Предложение	Координаты (частота слов)		
	мама	мыть	рама
мама мыть рама	1	1	1
рама мыть мама	1	1	1
мама мама мыть рама	2	1	1

[2] Более полный список стоп-слов для различных языков можно найти в документации NLTK (Natural Language Toolkit). Режим доступа: https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/stopwords.zip (дата обращения: 25.07.2025).

Косинусное расстояние

Единая размерность полученных векторов дает возможность сравнивать их по так называемому косинусному расстоянию [9; 10; 19]. Очевидно, что если конкретное слово присутствует в предложении, то его координата в пространстве будет больше единицы, если отсутствует, то координата равна нулю. Таким образом, все векторы будут находиться в первом квадранте векторного пространства (где координаты либо положительны, либо равны нулю), угол, аналогичный прямому углу (90°) в дву- и трехмерном пространстве, сохраняется и здесь, т.е. расстояние между всеми векторами по величине косинуса будет не менее 0 и не более 1. Следует также отметить, что направление измерения угла (по часовой стрелке или против нее) не имеет значения, поскольку косинус – функция четная и при любом направлении получаем значения в диапазоне от 0 до 1.

При расчете расстояний между векторами на практике тригонометрические функции не используются, вместо этого вычисляется отношение суммы произведений соответствующих координат векторов к произведению их норм, где норма – это квадратный корень из суммы квадратов его координат [19]:

$$\cos(x,y)=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}}.$$

Пример расчета косинусного расстояния для предложений 1 и 2 и представляемых ими векторов из табл. 3:

«мама мыть рама» $x = (1, 1, 1)$,

«рама мыть мама» $y = (1, 1, 1)$,

$$\cos(x,y)=\frac{1\cdot1+1\cdot1+1\cdot1}{\sqrt{1^2+1^2+1^2}\cdot\sqrt{1^2+1^2+1^2}}=$$
$$=\frac{3}{\sqrt{3}\cdot\sqrt{3}}=\frac{3}{\sqrt{3}^2}=\frac{3}{3}=1.$$

Косинусное расстояние равно единице, то есть векторы математически одинаковые.

А вот расчет косинусного расстояния для предложений 1 и 3 и представляемых ими векторов из табл. 3 дает другой результат:

«мама мыть рама» $x = (1, 1, 1)$,

«мама мама мыть рама» $y = (2, 1, 1)$,

$$\cos(x,y)=\frac{1\cdot2+1\cdot1+1\cdot1}{\sqrt{1^2+1^2+1^2}\cdot\sqrt{2^2+1^2+1^2}}=$$
$$=\frac{4}{\sqrt{3}\cdot\sqrt{6}}=\frac{4}{\sqrt{18}}=\frac{4}{3\sqrt{2}}\approx0,94.$$

Косинусное расстояние меньше единицы, то есть между векторами есть угол и векторы не одинаковые.

Метод *Term Frequency-Inverse Document Frequency* (*TF-IDF*)

Второй использованный в данной работе метод называется *TF-IDF* [17]. Он использовался в дополнение к простейшему методу «мешок слов», так как эти два метода являются наиболее используемыми при анализе частотности текстов [17].

Величина *TF-IDF* состоит из двух сомножителей *TF* и *IDF*. Величина *TF* отображает частоту появления слова в тексте, показывая относительную значимость этого слова внутри конкретного документа (статьи):

$$tf(t,d)=\frac{n_t}{\sum_k n_k},$$

где n_t – количество вхождений слова t в документ; $\sum_k n_k$ – общее число слов в данном документе.

Например, для трех предложений (статей) из табл. 3 подсчитать *TF* не составляет труда (см. табл. 4), так как слов в предложении не более четырех и используются только три разных леммы.

Величина *IDF* показывает инверсную частоту встречаемости слова в документах корпуса (множества собранных статей). Эта величина вычисляется как логарифм (с постоянным основанием

Таблица 4. Значения *TF* в трехмерном пространстве слов
Table 4. *TF* values in a three-dimensional word space

Предложение	Количество слов в предложении	Значение *TF* (координаты слов в пространстве *TF*)		
		мама	мыть	рама
мама мыть рама	3	1/3	1/3	1/3
рама мыть мама	3	1/3	1/3	1/3
мама мама мыть рама	4	2/4	1/4	1/4

для всех расчетов) отношения общего числа документов к количеству документов, содержащих данное слово:

$$IDF(t,D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|},$$

где $|D|$ – количество документов в коллекции; $|\{d_i \in D \mid t \in d_i\}|$ – число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

Существуют разные формулы для вычисления *TF-IDF*, связанные с особенностями расчета в каждой среде. Например, настоящая работа выполнена с помощью языка *Python* и библиотеки *scikit-learn*, где формула *TF-IDF* модифицирована для предотвращения вырожденных случаев либо путем корректировки логарифма, либо добавлением единицы после вычисления логарифма[3]:

$$IDF(t,D) = \log\left(\frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}\right) + 1 \qquad (1)$$

или

$$IDF(t,D) = \log\left(\frac{|D|}{1+|\{d_i \in D \mid t \in d_i\}|}\right).$$

При расчетах ниже будем пользоваться формулой (1).

Вычисление *IDF* тоже не представляет труда (табл. 5), так как документов (предложений) и входящих в них слов не очень много.

В данном случае получился тривиальный случай, так как все слова есть во всех трех документах, поэтому при перемножении значений *TF* и *IDF* получаем произведение, равное *TF* (см. табл. 4).

В работе использовались следующие параметры метода *TF-IDF*:

– обрабатываемые фрагменты текста представлены униграммами (параметр Ngram_range установлен на значения 1–1);

– применен сглаживающий эффект;

– использовано логарифмическое масштабирование веса терминов;

– ограничение по минимальной или максимальной частотности слов отсутствует.

Сравнение методов *BoW* и *TF-IDF*

Первое сравнение проведено на основе корреляционного анализа. Коэффициент корреляции между косинусными расстояниями, подсчитанными методами *BoW* и *TF-IDF*, составил 85 %, что свидетельствует о сильной взаимосвязи косинусных расстояний, подсчитанных разными способами.

Для наглядного сравнения результатов расчетов с использованием двух методов на рис. 1 приведена гистограмма распределения косинусных расстояний, полученных при обработке статей из раздела «Газодинамика» научно-технического журнала «Вестник Концерна ВКО «Алмаз – Антей». Из графика видно, что распределение косинусных расстояний, рассчитанных по методу *TF-IDF*, имеет меньший разброс (меньшую дисперсию) по сравнению с *BoW*, поэтому можно утверждать, что оценка по методу *TF-IDF* точнее ($7 \cdot 10^{-3}$ методом *BoW* против $3 \cdot 10^{-3}$ методом *TF-IDF*).

В ходе исследования было выявлено, что распределение косинусных расстояний для обоих методов (*BoW* и *TF-IDF*) соответствует гамма-распределению [20]:

$$f_X(x) = \begin{cases} x^{k-1}\dfrac{e^{-x/\theta}}{\theta^k \Gamma(k)}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

где

$$k = \frac{(\mathbb{E}[X])^2}{\mathrm{Var}(X)}; \quad \theta = \frac{\mathrm{Var}(X)}{\mathbb{E}[X]},$$

$\mathbb{E}[X]$, $\mathrm{Var}(X)$ – математическое ожидание и дисперсия случайной величины X (в данном случае совокупности величин косинусных расстояний).

Для приведения частотных гистограмм к одинаковому масштабу с теоретическим гамма-распределением каждое значение гистограммы было разделено на ширину интервала гистограммы (0,03) и сумму всех частот.

[3] Scikit Learn. User Guide. 7. Dataset transformations. 7.2. Feature extraction. Available from: https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction (accessed: 25.07.2025).

Таблица 5. Значение *IDF* в трехмерном пространстве слов
Table 5. IDF values in a three-dimensional word space

Предложение	Общее количество документов (предложений)	Число документов (предложений), где встречается каждое из слов (мама – 3, мыть – 3, рама – 3)	*IDF*		
			мама	мыть	рама
мама мыть рама	3	3	log(3/3) + 1 = 1	1	1
рама мыть мама	3	3	1	1	1
мама мама мыть рама	3	3	1	1	1

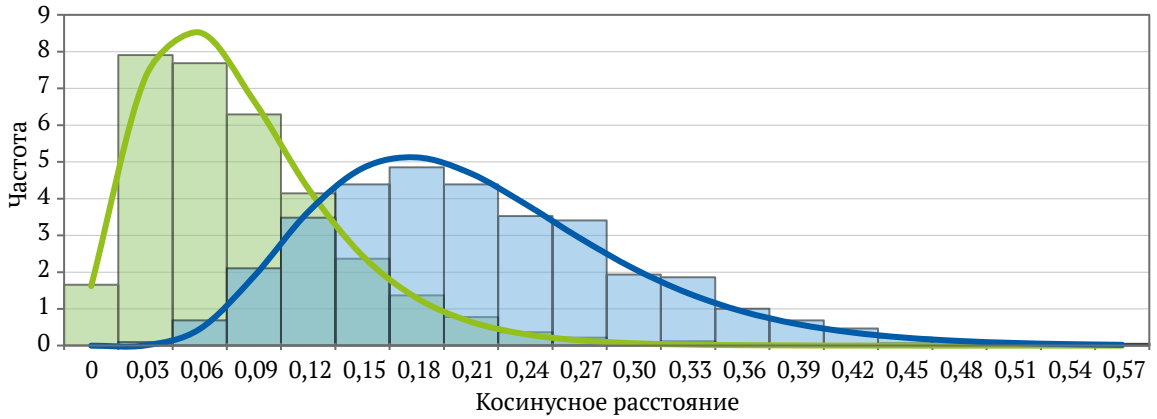

Рис. 1. Нормированные гистограммы распределения косинусных расстояний, рассчитанных методом *BoW* (голубой цвет) и *TF-IDF* (зеленый цвет), и теоретические кривые гамма-распределения, параметры которых оценены по данным гистограммам
Fig. 1. Normalized histograms of cosine distance distributions for *BoW* and *TF-IDF* models, along with theoretical gamma distribution values calculated by estimating parameters from the histograms (x-axis represents cosine distance value, y-axis shows density or normalized histogram)

Следует отметить, что закономерность распределения косинусных расстояний сохраняется и при сравнении разных корпусов текстов (например, «Радиолокации» и «Газодинамики»). В этом случае среднее значение гамма-распределения косинусных расстояний и по методу *BoW*, и по методу *TF-IDF* уменьшается (что очевидно, так как тексты содержат разную узкоспециализированную лексику и менее сходны, чем тексты из одного корпуса).

Для корпуса «Наукометрия», содержащего всего 12 статей, ввиду малого объема выборки не удалось получить достоверные статистические результаты о соответствии полученного распределения косинусных расстояний гамма-распределению и общему количеству парных комбинаций статей, вычисляемому по формуле [20]

$$C_n^m = \frac{n!}{(n-m)!m!}. \quad (2)$$

Для $n = 12$ и $m = 2$ по формуле (2) получаем количество значений для построения гистограммы:

$$C_{12}^2 = \frac{12!}{(12-2)!2!} = \frac{12!}{(10)!2!} = \frac{12 \cdot 11}{2} = 6 \cdot 11 = 66.$$

Этого количества достаточно для построения гистограммы, но недостаточно для проверки критерия о соответствии распределения гамма-распределению. Следует отметить, что значимые результаты получаются при количестве статей в корпусе больше 30 (что соответствует количеству сочетаний – 435).

Найти какое-то практическое применение выводу о соответствии распределения косинусного расстояния гамма-распределению пока не удалось. Изначально выдвигалась гипотеза, предполагающая использование порогового значения гамма-распределения для фильтрации статей с низкими показателями косинусного расстояния, однако она не подтвердилась.

РЕЗУЛЬТАТЫ

Полный исходный код исследования на языке *Python* с полученными результатами (но без исходных данных) опубликован на *GitHub*[4] под лицензией *CC BY-NC*[5].

Построение и валидация модели

Эксперимент проводился в среде *Google Collab* на языке *Python* с использованием библиотеки для проведения лемматизации[6] *pymorphy3*.

Гипотеза эксперимента заключалась в том, что для каждой рукописи в наборе может быть найдена наиболее семантически близкая к ней по величине косинусного расстояния. Для поиска каждая анализируемая рукопись корпуса последовательно временно исключается из общего корпуса статей, а между исключенной и оставшимися статьями корпуса рассчитывается косинусное

[4] Available from: https://github.com/denisbolshakoff/analysis_text_corpus_project (accessed: 25.07.2025).

[5] Available from: https://creativecommons.org/licenses/by-nc/4.0/ (accessed: 25.07.2025).

[6] Available from: https://pypi.org/project/pymorphy3/ (accessed: 25.07.2025).

расстояние. Статья с максимальным косинусным расстоянием признается наиболее близкой к анализируемой. И, следовательно, для анализируемой рукописи может быть назначен тот же рецензент, который ранее рецензировал наиболее близкую к ней статью.

Переменные эксперимента – косинусные расстояния между лемматизированными и векторизованными текстами рукописей, целевая метрика – достижение максимального косинусного расстояния. Порядок расположения рукописей при расчете был всегда задан один и тот же – от первой в наборе до последней. При этом из набора исключалась каждая исследуемая рукопись, так как очевидно, что ее косинусное расстояние с самой собой будет максимально и равно 1.

Каждая рукопись корпуса была лемматизирована, очищена от стоп-слов и знаков пунктуации, переведена в векторное представление. Затем обрабатывались все рукописи указанных корпусов («Радиолокация», «Газодинамика», «Наукометрия») и выполнялся попарный расчет косинусного расстояния между всеми рукописями с целью определения максимального косинусного расстояния среди всех возможных пар. Рукопись А считается ближайшей к рукописи *B*, если значение косинусного расстояния между ними наибольшее среди всех пар *A* и B_i. Максимальное косинусное расстояние служит критерием близости одной рукописи к другой. Соответственно, если одна из них рассматривается как новая, для нее могут быть назначены те же рецензенты, которые ранее работали со статьей, наиболее близкой к вновь поступившей.

Результаты моделирования выдавались в виде пары чисел: первое число – номер анализируемой статьи, второе – номер статьи, наиболее близкой к первой по значению косинусного расстояния. Результаты были представлены в виде, приведенном в табл. 6 (показаны только первые восемь записей набора, так как вывод всех 274 результатов занял бы пять печатных страниц).

Таблица 6. Результаты применения двух методов для определения рукописи, наиболее близкой к исследуемой

Table 6. Results of applying two methods to determine the manuscript closest to the investigated one

Номер рукописи	*BoW*	*TF-IDF*
1	0*, 1	0, 135
2	1, 180	1, 180
3	2, 161	2, 161
4	3, 87	3, 74
5	4, 36	4, 36
6	5, 93	5, 93
7	6, 21	6, 2
8	7, 151	7, 145

* «0» означает первый текст, так как в Python нумерация начинается с нуля.

Таблица 7. Результаты подбора рецензентов по модели *BoW*

Table 7. Partial experiment results (reviewer surnames in the table are fictional but their ratios in the manuscripts under study are correct)

Номер рукописи	Результаты расчета по модели *BoW*	Рецензенты исследуемой рукописи	Рецензенты рукописи, наиболее близкой к исследуемой по косинусному расстоянию	Совпадающие рецензенты
1	0, 1	Каплин, Снежинский	Верхоглядов, Каплин, Снежинский	Каплин, Снежинский
2	1, 180	Верхоглядов, Каплин, Снежинский	Каплин, Крыловатых, Синяков	Каплин
3	2, 161	Толоконников, Каплин, Косинец	Работин, Каплин, Воронин	Каплин
4	3, 87	Остроженков, Елахин, Ножнин, Порошин	Дубов, Высотин, Остроженков	Остроженков
5	4, 36	Верхоглядов, Снежинский	Сажин, Ивушкин	–
6	5, 93	Каплин, Корабельников	Гамов, Остроженков, Елахин	–
7	6, 21	Гамов, Елахин, Остроженков, Сапельников	Толоконников, Косинец	–
8	7, 151	Каплин, Снежинский	Верхоглядов, Каплин, Снежинский	Каплин, Снежинский

Примечание. Фамилии рецензентов в таблице вымышленные, но их соотношения в исследуемых рукописях правильные.

Как видно из табл. 6, результаты обработки не всегда бывают одинаковыми, поскольку методы преобразования текстов в численные векторы различаются (см. рукописи 1, 4, 7 и 8). Однако полное совпадение результатов расчетов по методам *BoW* и *TF-IDF* (см. рукописи 2, 3, 5, 6) наблюдается в 63 % случаев. Данный факт связан с особенностями способов векторизации текста, а также сходством словарного состава статей в разделах «Радиолокация» и «Газодинамика», где отнесение отдельных слов к разным рукописям может рассматриваться как статистическая погрешность расчета косинусного расстояния. Необходимо учитывать, что абсолютные значения чисел в табл. 6 не подлежат интерпретации сами по себе, потому что представляют собой порядковые номера статей с максимальным косинусным расстоянием. При этом разница в сроках публикации близких по содержанию статей может доходить до нескольких лет, поэтому их порядковые номера могут различаться на 100 единиц и более.

Часть результатов эксперимента (для метода векторизации *BoW*) с фамилиями рецензентов по найденным в табл. 6 соответствиям приведена в табл. 7.

Автор как редактор исследуемого научного журнала проверил все результаты обработки – и по методу *BoW*, и по методу *TF-IDF* – исследуемых рукописей и рукописей, которые максимально близки к исследуемым по косинусным расстояниям. В качестве редактора исследуемого научного журнала автор мог бы поручить (в некоторых случаях поручил) найденным по косинусным расстояниям рецензентам оценить пришедшую рукопись, которая была обработана моделью, предложившей соответствующих рецензентов.

Оценка точности модели

Пофамильное сравнение

Если в третьей и четвертой колонке табл. 7 совпадает хотя бы одна фамилия рецензента, считаем, что модель корректно указала рецензента. Рассчитанная таким образом точность модели приведена в табл. 8. Общее количество парных комбинаций для 273 статей составило 37 128. Для сравнения в табл. 8 приведена нейросетевая модель *Doc2Vec* [7].

[7] Bolshakov D. Yu. Classification, clustering of a corpus of texts in scientific style for radar, gas dynamics, and scientometrics. GitHub Repository. 2025. Available from: https://github.com/denisbolshakoff/classification_of_texts_in_scientific_style/ (accessed: 25.07.2025).

Таблица 8. Точность (%) выбора рецензента моделью
Table 8. Predictor accuracy of reviewer model

Корпус	*BoW*	*TF-IDF*	*Doc2Vec*
Радиолокация	55	56	53
Газодинамика	85	76	55

В табл. 8 отсутствуют данные о корпусе «Наукометрия», поскольку автору неизвестны рецензенты, которые принимали решения о публикации или отклонении его рукописей. При этом ни одна рукопись из этого корпуса не была идентифицирована как семантически близкая статьям из корпусов «Радиолокация» или «Газодинамика» при использовании как модели *BoW*, так и *TF-IDF*.

Пофамильное сравнение с учетом фактора времени

В данных табл. 8 не учтен фактор времени, то есть в результаты расчета включены все рецензенты, в том числе и прекратившие сотрудничество с журналом, а значит, вновь пришедшая рукопись не может быть направлена к ним на рецензирование. Аналогично рукопись, которая поступила несколько лет назад, была направлена другим рецензентам, потому что рекомендуемые моделью рецензенты на тот момент не сотрудничали с журналом. За 15 лет существования журнала сменилось более трехсот рецензентов, и историческая информация, которая входит в расчетные данные модели, может приводить к некорректным результатам, так как назначенного эксперта уже может не быть в списке рецензентов [12; 21; 22]. Если ограничить выборку только теми специалистами, которые до сих пор сотрудничают с журналом, точность рекомендаций рецензента моделью возрастает (табл. 9). Для сравнения в табл. 9 приведены результаты, полученные с применением нейросетевой модели *Doc2Vec*.

Большой разброс показателей точности между корпусами в табл. 8 и 9 обусловлен тем, что «Газодинамика» – относительная новая рубрика для журнала (примерно семь лет), в то время как «Радиолокация» существует с момента основания журнала в 2011 г.

Таблица 9. Точность (%) предсказания рецензента моделью с учетом фактора времени
Table 9. Predictor accuracy of reviewer model considering time factor

Корпус	*BoW*	*TF-IDF*	*Doc2Vec*
Радиолокация	72	66	68
Газодинамика	92	89	63

Пофамильное сравнение с учетом фактора времени и связей между рецензентами

Как показало исследование автора [21; 22], в редакционной коллегии научного журнала со временем формируются взаимосвязи между всеми рецензентами благодаря совместному рецензированию статей. Граф на рис. 2 показывает текущие связи между рецензентами в процессе рецензирования статей. Он отличается от приведенного в более ранних работах автора (см., например, [21]) включением данных за прошедшие с момента публикации статьи 4 года (с 2021 по 2024 г.). Граф связей между рецензентами строился с учетом того, что в процессе рецензирования для статьи обычно[8] назначается несколько рецензентов (в среднем 2,2, максимум 8 – для массива, исследуемого в настоящей статье научного журнала). В редакции журнала существует практика взаимного ознакомления рецензентов с экспертными оценками коллег. При этом количество связей определяется числом сочетаний:

[8] Для статей, не соответствующих тематике журнала или показавших процент оригинальности при проверке в системе «Антиплагиат» ниже требуемого, рецензенты не назначались, для откровенно слабых статей назначался один рецензент с целью отправки формальной рецензии с отказом.

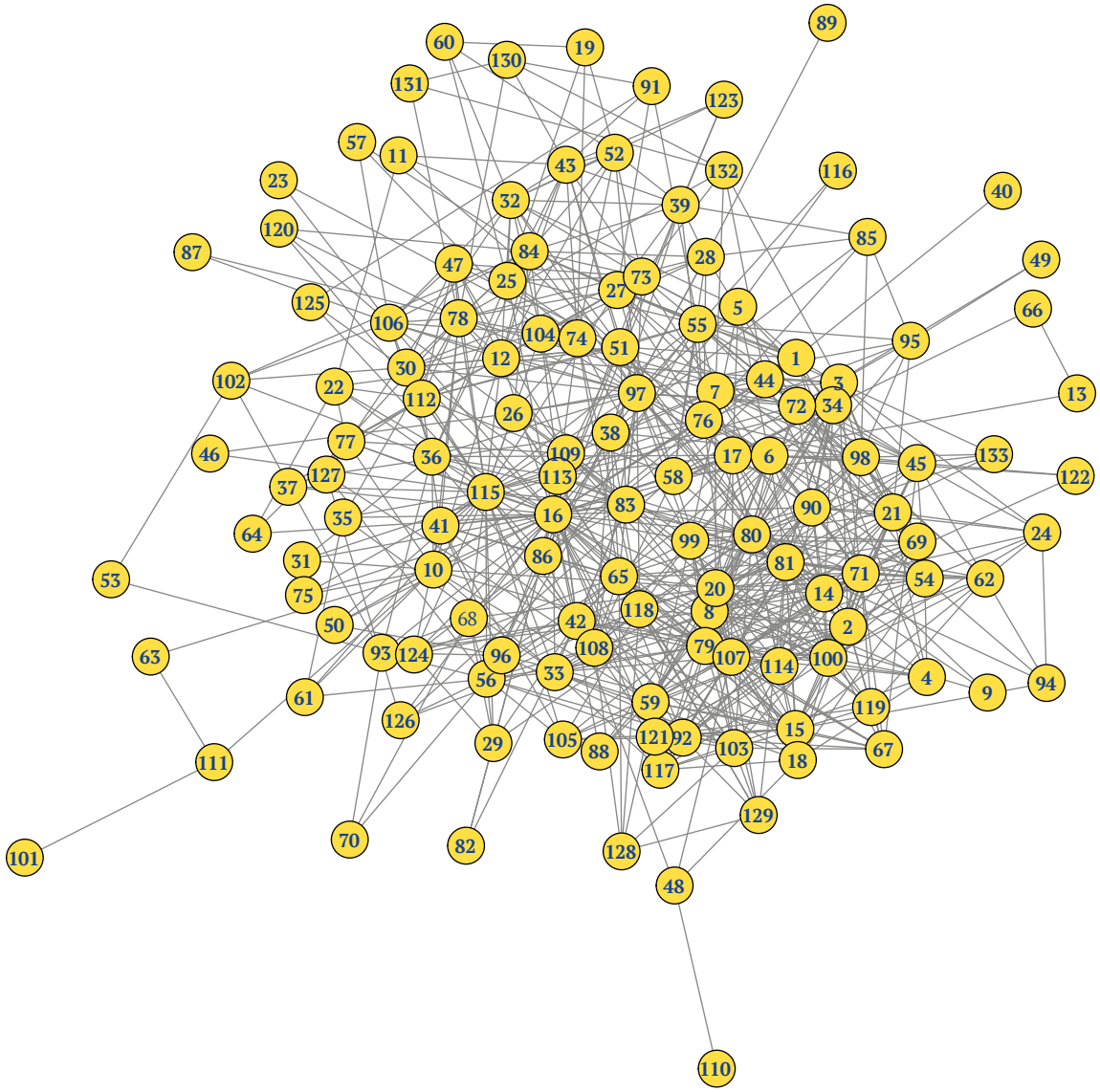

Рис. 2. Граф связей между рецензентами в редакционной коллегии научно-технического журнала «Вестник Концерна ВКО «Алмаз – Антей» с 2015 г.

Fig. 2. Reviewer connection graph in the editorial board of the scientific and technical the Journal of “Almaz – Antey” Air and Space Defence Corporation since 2015

два рецензента, участвующих в оценке, образуют одну связь, три рецензента – три связи, четыре – шесть связей и так далее по формуле числа сочетаний из N по 2 (формула (2), где $n = N$, $m = 2$). За несколько лет образуются связи между всеми рецензентами, причем диаметр графа, то есть максимальное расстояние между любыми двумя рецензентами, не превышает шести [21].

Результат модели считаем успешным, если для поступившей рукописи назначается рецензент, непосредственно связанный (длина связи = 1) с рецензентом, ранее проводившим экспертную оценку статьи, наиболее близкой к новой по косинусному расстоянию. Например, на рис. 3 представлены остовные деревья графа, построенные с помощью алгоритма ближайшего соседа [23], для трех вершин графа на рис. 2 (рецензенты 10 и 31 и их связи с другими рецензентами). Если для рукописи, которую оценивали рецензент 10 и 31, модель выбрала как наиболее близкую рукопись, прорецензированную экспертами 16 и 36 (каждый из которых имеет связи и с 10, и с 31), то считаем, что модель корректно назначила рецензентов.

Остовное дерево на рис. 3 отражает все прямые связи данного рецензента с другими специалистами в процессе экспертной оценки рукописи без учета связей других рецензентов друг с другом. То есть остовное дерево можно интерпретировать как потенциальные альтернативные варианты назначения рецензента в случае невозможности привлечения к рецензированию основного специалиста в данный момент (отпуск, командировка, загруженность и т.д.). Такая процедура может быть реализована средствами языка программирования *R*, использованного автором для анализа связей рецензентов друг с другом. Замены определяются автоматически путем пересечения двух множеств соседних вершин остовных деревьев командой

```
intersect(c(neighbors(g, 10)),c(neighbors(g, 31))),
```

которая возвращает список из трех общих смежных вершин (рецензентов 16, 36, 113), связанных одновременно с рецензентами 10 и 31 (*g* означает граф, представленный на рис. 2; c() – массив; функция intersect подсчитывает пересечение множеств).

В редакции научного журнала замены рецензентов происходят довольно часто, и редко все статьи по определенной тематике оценивает только один рецензент: обычно есть пул рецензентов по каждой предметной области. Поэтому замены в пуле рецензентов не следует считать ошибкой модели. В таком случае скорректированная точность модели указана в табл. 10, в которой также приведены показатели точности, полученные с использованием нейросетевой модели *Doc2Vec*.

Точность в 99 % объясняется тем, что рукопись из раздела «Радиолокация» была ошибочно отнесена моделью по косинусному расстоянию в раздел «Газодинамика», а точность в 95 % – тем, что 3 рукописи из 54, составляющих раздел «Газодинамика», неверно классифицированы как относящиеся к разделу «Радиолокация».

Невысокая точность определения рецензента нейросетевой моделью *Doc2Vec* свидетельствует о целесообразности использования более простых моделей вместо сложных нейросетевых. Однако стоит отметить, что тренировка и проведение предсказаний нейросетевой модели *Doc2Vec* занимают значительно меньше времени (минуты), чем расчет и сравнение косинусных расстояний лемматизированных текстов (около часа).

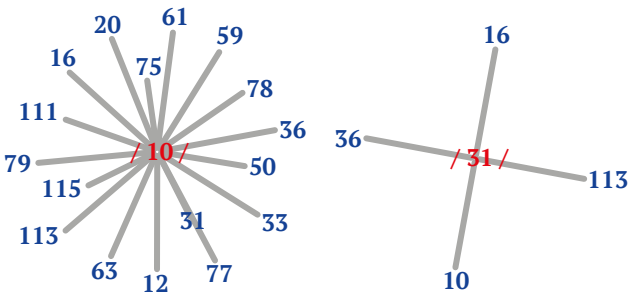

Рис. 3. Построенное с использованием алгоритма ближайшего соседа остовное дерево графа для рецензентов и их связей с коллегами, образованных при совместном рецензировании

Fig. 3. Reviewers and their minimum spanning tree of nearest neighbor graph – co-reviewer colleagues who jointly reviewed a manuscript

Таблица 10. Точность (%) выбора/назначения рецензента моделью с учетом фактора времени и фактора связанности рецензентов

Table 10. Predictor accuracy of reviewer model considering both time factor and reviewer connectivity factor

Корпус	*BoW*	*TF-IDF*	*Doc2Vec*
Радиолокация	99	99	93
Газодинамика	98	95	70

Полный исходный код исследования связей рецензентов в редакции научного журнала на языке *R* с полученными результатами и исходными данными доступен на платформе *GitHub*[9] под открытой лицензией *CC BY-NC*[10].

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Настоящее исследование было направлено на разработку и верификацию простой, интерпретируемой и применимой на практике модели автоматического назначения рецензентов на основе классических методов обработки естественного языка. Результаты экспериментов, представленные в табл. 10, позволяют утверждать, что предложенная модель демонстрирует высокий уровень точности соответствия рукописей ранее прорецензированным публикациям. Однако, как справедливо подчеркивается в работах [8–12], вопрос достоверности таких оценок остается открытым, поскольку в реальной редакционной практике решение о назначении рецензента принимает человек, а не алгоритм. Таким образом, даже высокая степень совпадения не может интерпретироваться как окончательный критерий корректности.

Отдельного внимания заслуживает интеграция графа связей между рецензентами (см. рис. 2), которая позволила значительно повысить точность предсказаний, приближая ее к 100%. Эффект объясняется тем, что в модель включается дополнительный слой данных – исторически сложившиеся экспертные связи между рецензентами, неявно отражающие как тематическую близость, так и повторяющееся взаимодействие в рамках редакционных процессов. Подобный подход, использующий социально-сетевые структуры, редко реализуется в современных системах автоматической обработки текстов, ориентированных преимущественно на контентный анализ. Следовательно, можно утверждать, что проведенное исследование предлагает новое перспективное направление моделирования сети рецензентов.

С практической точки зрения предложенная модель может быть внедрена в редакционную систему любого моно- или мультидисциплинарного журнала, имеющего архив не менее чем из 30 рассмотренных рукописей, при условии системного назначения двух и более рецензентов на статью. В противном случае построение графа связей становится невозможным и эффективность модели снижается. Отметим также, что качество предсказания оказывается особенно высоким при работе с научными текстами, насыщенными терминологией из узкой предметной области, что соответствует выводам, представленным в исследованиях по тематической релевантности векторных моделей [10; 11].

Переходя к оценке потенциала модели для решения более амбициозной задачи – прогнозирования судьбы рукописи (будет ли она принята к публикации или отклонена), следует признать, что текущие результаты носят *предварительный* характер. Построенная модель стохастического вложения соседей с *t*-распределением (*t*-SNE), использованная для визуализации результатов классификации методом *k*-ближайших соседей [5], позволила успешно разделить тексты тестовой выборки по тематическим кластерам (рис. 4). При этом для моделей векторизации *BoW, TF-IDF* и *HashVectorizer* были достигнуты показатели *precision* (доля *верно предсказанных моделью положительных случаев)*, *recall* (доля реальных положительных случаев, *правильно предсказанных моделью*) и *F1-score* (гармоническое среднее *precision и recall*), превышающие 0,98. Сравнимые результаты демонстрирует и полносвязная нейросеть (рис. 5), классифицирующая рукописи по тематике с макрометриками, превышающими 0,97.

В то же время попытки классифицировать рукописи по признаку «публикуемость» показали ограничения метода. Согласно данным рис. 6, использование нейросети привело к снижению показателя *recall* до 0,71, а *F1-score* – до 0,76, что свидетельствует об ошибках модели примерно в трети случаев. Возможной причиной может служить недостаточный объем выборки: как отмечают Л. Ван Дер Маатен и Г. Хинтон (L. van der Maaten & G. Hinton) [24], обучение даже простых моделей для бинарной классификации требует наличия нескольких тысяч размеченных примеров, особенно в условиях жанровой, стилистической и тематической неоднородности текстов.

Рисунки 4–6 демонстрируют, что кластеризация по тематике («Радиолокация», «Газодинамика», «Наукометрия») реализуется достаточно уверенно: соответствующие цветовые группы визуально разделены, пересечения между ними минимальны, несмотря на понижение размерности до двумерного пространства. Однако идентификация признаков публикационной успешности оказывается существенно более затрудненной. Даже при использовании полносвязной нейронной

[9] Bolshakov D. Yu. Application of graph theory to analyze the connectivity of reviewers in scientific journal. GitHub Repository. 2025. Available from: https://github.com/denisbolshakoff/to_analyze_the_connectivity_of_reviewers (accessed: 25.07.2025).

[10] Creative Commons. Attribution Non-Commercial 4.0 International. Available from: https://creativecommons.org/licenses/by-nc/4.0/ (accessed: 25.07.2025).

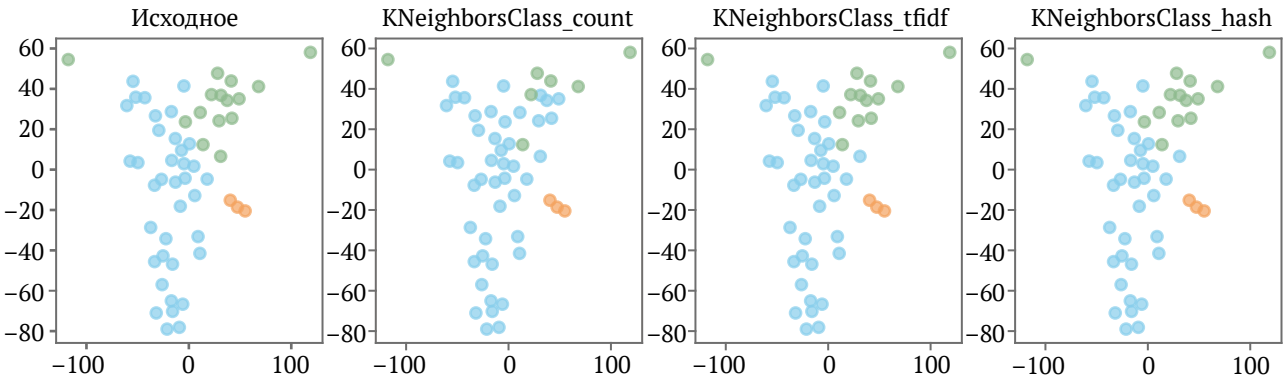

Рис. 4. Результаты разделения статей по тематике методом *k*-ближайших соседей (n_neighbors=3) тремя моделями векторайзерами *BOW*, *TF-IDF* и *HashVectorizer* («Радиолокация» – голубой цвет, «Газодинамика» – зеленый, «Наукометрия» – оранжевый)

Fig. 4. Classification results using KNeighborsClassifier (n_neighbors=3) for three vectorizers: BOW, TF-IDF, and HashVectorizer (blue radar detection, green gas dynamics, orange scientometrics)

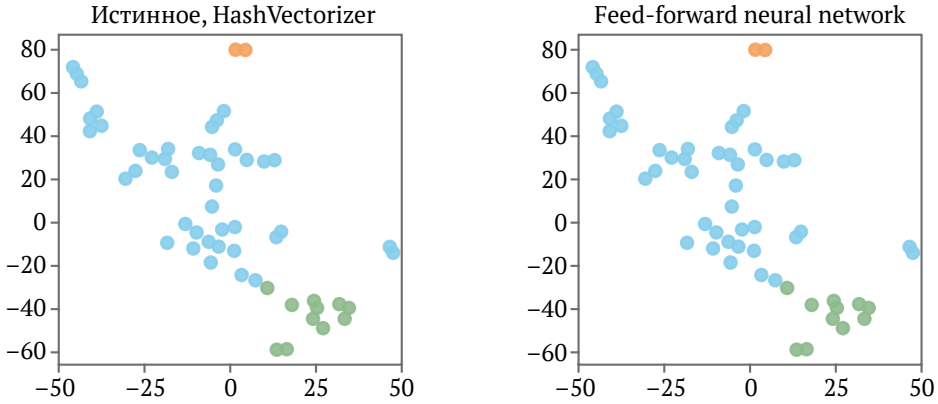

Рис. 5. Результаты разделения полносвязной нейронной сетью статей тестовой выборки на три класса («Радиолокация» – голубой цвет, «Газодинамика» – зеленый, «Наукометрия» – оранжевый) с использованием t-SNE для векторизованных данных, полученных с помощью простейшего метода OneHotEncoding

Fig. 5. Construction of t-SNE for test classification sample of fully connected neural network on three classes of articles (blue radar detection, green gas dynamics, orange scientometrics) built using simple vectorization method OneHotEncoding

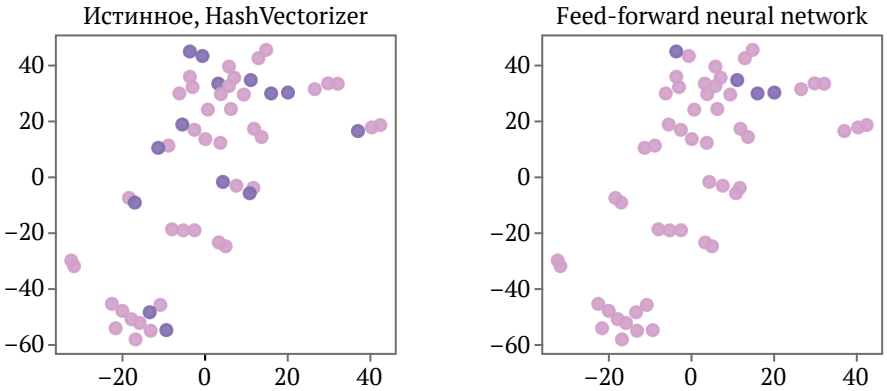

Рис. 6. Результаты разделения нейронной сетью статей тестовой выборки на два класса с использованием *t*-NSE (публикуемая статья – сиреневый цвет; непубликуемая – фиолетовый)

Fig. 6. Construction of t-SNE for test classification sample of neural network on two article classes (published light purple and unpublished dark purple)

сети не удается добиться результатов, делающих модель пригодной для применения в редакции: алгоритм в 30% случаев ошибочно классифицирует публикуемую статью как непубликуемую и наоборот. Этот результат лишь подтверждает необходимость существенного расширения корпуса для решения задач прогностического характера.

Результаты кластеризации без учета классов (рис. 7) также подтверждают ограниченность возможностей модели при отсутствии надежной разметки: метрики точности и полноты оказались неудовлетворительными (*precision* 0,53, *recall* 0,37, *F1-score* 0,32), что свидетельствует о слабой структурности латентного пространства признаков при отсутствии обучающего сигнала.

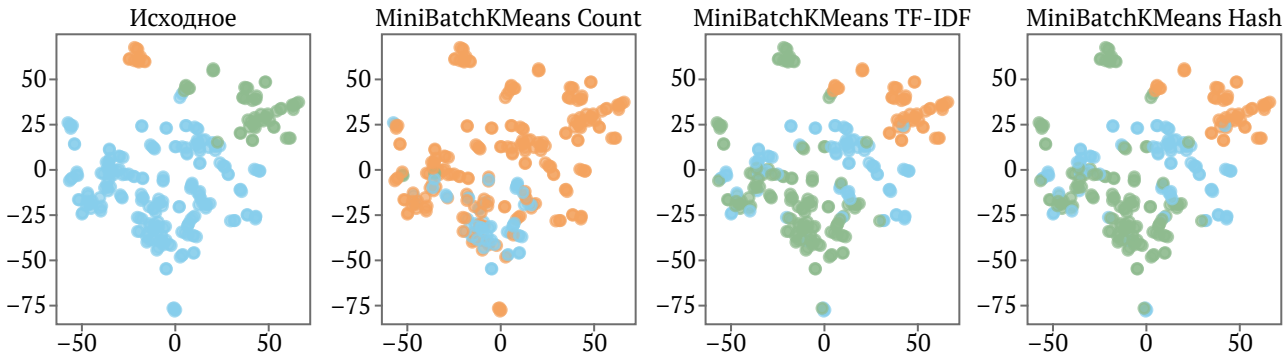

Рис. 7. Результаты кластеризации с применением метода t-SNE
(«Радиолокация» – голубой цвет, «Газодинамика» – зеленый, «Наукометрия» – оранжевый)

Fig. 7. Clustering results using MiniBatchKMeans method
(left: t-SNE for original data – blue radar detection, green gas dynamics, orange scientometrics)

Таблица 11. Частотные зависимости частей речи для языков мира
Table 11. Frequency dependencies of parts of speech for world languages

Язык	Семья	Группа	Доля говорящих в мире, %	Доля в языке, %				
				Сущ.	Гл.	Прил.	Нар.	Ост.
Английский (брит.)	И	г	19,13	52	19	17	7	5
Английский (амер.)	И	г		51	20	17	7	5
Арабский	С	ц	3,75	57	17	18	2	6
Голландский	И	г	0,25	46	25	17	6	6
Испанский	И	р	7,50	49	22	19	3	7
Китайский	СТ	к	16,25	37	23	11	7	22
Корейский	ИЗ	и	0,98	58	21	5	7	9
Немецкий	И	г	1,63	48	24	16	6	6
Португальский	И	р	3,13	52	20	16	3	9
Персидский	И	ир	2,80	63	7	24	3	3
Русский	И	с	3,13	53	19	17	6	5
Турецкий	Т	о	1,06	63	15	16	3	3
Французский	И	р	4,29	45	23	16	6	10
Чешский	И	с	0,13	43	22	18	7	10
Японский	Я	я	1,63	64	17	5	7	7
		Всего:	66,00	–	–	–	–	–

Примечание. Обозначения: [семья] И – индоевропейская, ИЗ – изолированная, С – семитская, СТ – сино-тибетская, Т – тюркская, Я – японская; [группа] г – германская, ц – центральносемитская, р – романская, к – китайская, и – изолированная, ир – иранская, с – славянская, о – огузская, я – японо-рюкюская; сущ. – существительные, гл. – глаголы, прил. – прилагательные, нар. – наречия, ост. – остальные.
Источники: данные таблицы рассчитаны автором по частотным словарям языков мира [25–39].

Эти данные согласуются с выводами о недостаточной информативности векторных представлений без контекстуализации, сделанными Л. Хобсоном с соавт. (L. Hobson et al.) [17], Д. Киелой и С. Кларком (D. Kiela & S. Clark) [19].

Представляет интерес и тестирование модели на языковую универсальность. Эксперимент, в рамках которого использовались только существительные, показал, что в этом случае точность предсказаний снижается лишь на 23 %, оставаясь на уровне 77 %. Напротив, исключение существительных приводит к падению точности до 38 %. Эти результаты интерпретируются в свете лингвистических закономерностей: как видно из табл. 11, существительные являются наиболее информативным классом слов в большинстве языков мира, независимо от их генетической и типологической принадлежности. Таким образом, при необходимости межъязыкового переноса модели возможно использование машинного перевода с фокусом на существительные, что открывает путь к универсализации подхода.

В совокупности полученные результаты позволяют сделать вывод, что предложенная модель является эффективным инструментом автоматизированного подбора рецензентов в пределах отдельно взятого научного журнала. Вместе с тем решение задачи прогнозирования публикационного исхода требует увеличения объема корпуса и привлечения более сложных архитектур глубокого обучения. Потенциал такого подхода представляется значительным, однако требует отдельного масштабного исследования.

ЗАКЛЮЧЕНИЕ

Настоящее исследование предложило простую и в то же время высокоэффективную модель автоматического подбора рецензентов для научных журналов на основе классических методов обработки естественного языка: лемматизации, *BoW* и *TF-IDF*. Модель показала высокую точность (до 99 %) при минимальных вычислительных затратах и доказала свою устойчивость в условиях дисбаланса классов и ограниченного объема данных. Важным преимуществом предложенного подхода является его доступность: модель может быть реализована на обычных офисных компьютерах и не требует использования дорогостоящих вычислительных ресурсов.

В то же время модель имеет определенные ограничения: она ориентирована на один конкретный журнал, требует регулярного обновления базы данных статей, а ее интерпретация основана на частотном анализе лексики без учета более глубоких семантических связей.

Наиболее перспективным направлением дальнейших разработок видится расширение модели за счет включения информации о профессиональных связях и тематической специализации рецензентов (включая граф связей), а также внедрение современных нейросетевых архитектур для повышения гибкости и масштабируемости системы. Учитывая накопленные данные редакций о стабильных рецензентских связках и предпочтениях, такие расширения могут не только повысить точность рекомендаций, но и усилить прозрачность и эффективность системы научной экспертизы в целом.

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликта интересов.

CONFLICT OF INTERESTS

The author declares no relevant conflict of interests.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Turing A. Computing Machinery and Intelligence. *Mind*. 1950;59(236):433–460. https://doi.org/10.1093/mind/LIX.236.433
2. Goldberg Y. Neural Network Methods for Natural Language Processing. Cham: Springer; 2017. 312 p. (Synthesis Lectures on Human Language Technologies). https://doi.org/10.1007/978-3-031-02165-7
3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL]. 2019 May 24. https://doi.org/10.48550/arXiv.1810.04805
4. Zhu X., Zhang M., Hong Y., He R., editos. Natural Language Processing and Chinese Computing. Proceedings of the 9th CCF International Conference, NLPCC 2020, (Zhengzhou, October 14–18, 2020). Cham: Springer; 2020. 857 p. (Lecture Notes in Computer Science. Vol. 12430). https://doi.org/10.1007/978-3-030-60450-9

5. Jia J., Liang W., Liang Y. A review of hybrid and ensemble in deep learning for natural language processing. arXiv preprint arXiv:2312.05589. 2023. https://doi.org/10.48550/arXiv.2312.05589
6. Jurafsky D., Martin J.H., Kehler A., Linden K.V., Ward N. Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition. Upper Saddle River, NJ: Prentice-Hall; 2000. 934 p.
7. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. *Автоматическая обработка текстов на естественном языке и анализ данных*. М.: Изд-во НИУ ВШЭ; 2017. 269 с.
Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinskiy E.S., Lukashevich N.V., Sapin A.S. Automatic natural language processing and data analysis. Moscow: NRU HSE Publ.; 2017. 269 p. (In Russ.).
8. Bhattacharya S., Mazumder A., Banerjee A., Bandyopadhyay C., Nandi S. Automated Reviewer Assignment Process Using Machine Learning Technique. In: Patel A., Kesswani N., Mishra M., Meher P., editos. *Advances in Machine Learning and Big Data Analytics* (*ICMLBDA 2023*). Cham: Springer; 2025, pp. 87–99. (Springer Proceedings in Mathematics & Statistics. Vol. 441). https://doi.org/10.1007/978-3-031-51338-1_7
9. Tan S., Duan Z., Zhao S., Chen J., Zhang Y. Improved Reviewer Assignment Based on Both Word and Semantic Features. *Information Retrieval Journal*. 2021;24(2):175–204. https://doi.org/10.1007/s10791-021-09390-8
10. Adebiyi A., Ogunleye O., Adebiyi M., Okesola O. A Comparative Analysis of TF-IDF, LSI and LDA in Semantic Information Retrieval Approach for Paper-Reviewer Assignment. *ARPN Journal of Engineering and Applied Sciences*. 2019;14(10):3378–3382. https://doi.org/10.36478/jeasci.2019.3378.3382
11. Anjum O., Gong H., Bhat S., Hwu W.M., Xiong J. PaRe: A Paper-Reviewer Matching Approach Using a Common Topic Space. arXiv preprint arXiv:1909.11258. 2019 Sep. https://doi.org/10.48550/arXiv.1909.11258
12. Peng H., Hu H., Wang K., Wang X. Time-Aware and Topic-Based Reviewer Assignment. In: Bao Z., Trajcevski G., Chang L., Hua W., editos. *Database Systems for Advanced Applications* (*DASFAA 2017*). Cham: Springer; 2017:145-157. (Lecture Notes in Computer Science. Vol. 10179). https://doi.org/10.1007/978-3-319-55705-2_11
13. Li C.L., Hu X., Xu M.H., Li K. , Zhang Y., Cheng X.Z. Can Large Language Models Be Trusted Paper Reviewers? A Feasibility Study. arXiv:2506.17311v1 [cs.CY]. 2025 June 18. https://doi.org/10.48550/arXiv.2506.17311
14. Liang W.X., Zhang Y.H., Cao H.C., Wang B., Ding D., Yang X. et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. arXiv:2310.01783v1 [cs.LG]. 2023 Oct 3. https://doi.org/10.48550/arXiv.2310.01783
15. Lee J., Lee J., Yoo J.-J. The Role of Large Language Models in the Peer-Review Process: Opportunities and Challenges for Medical Journal Reviewers and Editors. *Journal of Educational Evaluation for Health Professions*. 2025;22:4. https://doi.org/10.3352/jeehp.2025.22.4
16. Vasiliev Yu. Natural language processing with python and spaCy: A practical introduction. San Francisco, CA: No Starch Press; 2020. 217 p.
17. Lane H., Howard C., Hapke H.M. Natural language processing in action: understanding, analyzing, and generating text with python. 1st ed. Shelter Island, NY: Manning Publications Co.; 2019. 544 p.
18. Bengfort B., Bilbro R., Ojeda T. Applied text analysis with python: enabling language-aware data products with machine learning. 1st ed. Sebastopol, CA: O'Reilly Media; 2018. 330 p.
19. Kiela D., Clark S. A Systematic Study of Semantic Vector Space Model Parameters. In: *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality* (*CVSC*). Kerrville, TX: Association for Computational Linguistics; 2014, pp. 21–30. https://doi.org/10.3115/v1/W14-1503
20. Пугачёв В.С. Теория вероятностей и математическая статистика. М.: Физматлит; 2002. 496 с.
Pugachev V.S. Probability theory and mathematical statistics. Moscow: Fizmatlit; 2002. 496 p. (In Russ.).
21. Большаков Д.Ю. О связях в науке на примере редакционной коллегии научного журнала. *Наука и научная информация*. 2021;4(1-2):23–32. https://doi.org/10.24108/2658-3143-2021-4-1-2-23-32
Bolshakov D.Yu. On Relations in Science: The case of the scientific journal editorial board. *Scholarly Research and Information*. 2021;4(1-2):23–32. https://doi.org/10.24108/2658-3143-2021-4-1-2-23-32
22. Большаков Д.Ю. Дополнение к статье «О связях в науке на примере редакционной коллегии научного журнала». *Наука и научная информация*. 2022;5(1):8–10. https://doi.org/10.24108/2658-3143-2022-5-1-2
Bolshakov D.Yu. Supplement to the article "On relations in science: the case of the scientific journal editorial board". *Scholarly Research and Information*. 2022;5(1):8–10. https://doi.org/10.24108/2658-3143-2022-5-1-2

23. Diestel R. Graph theory. 6th ed. Berlin, Heidelberg: Springer; 2025. 455 p. https://doi.org/10.1007/978-3-662-70107-2
24. van der Maaten L. J. P., Hinton G. E. Visualizing data using t-SNE. *Journal of Machine Learning Research*. 2008;9(86):2579–2605. Available from: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (accessed: 13.02.2025).
25. Brezina V., Gablasova D. A frequency dictionary of British English: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2024. 340 p.
26. Davies M., Gardner D. A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists. 1st ed. London, New York: Routledge; 2010. 368 p.
27. Buckwalter T., Parkinson D. A frequency dictionary of Arabic: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2011. 578 p.
28. Tiberius C., Schoonheim T. A frequency dictionary of Dutch: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2014. 320 p.
29. Davies M. H., Davies K. H. A frequency dictionary of Spanish: Core vocabulary for learners. 2nd ed. London, New York: Routledge; 2018. 350 p.
30. Xiao R., Rayson P., McEnery T. A frequency dictionary of Mandarin Chinese: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2009. 390 p.
31. Lee S. H., Jang S. B., Seo S. K. A frequency dictionary of Korean: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2017. 358 p.
32. Tschirner E., Möhring J. A frequency dictionary of German: Core vocabulary for learners. 2nd ed. London, New York: Routledge; 2020. 304 p.
33. Davies M., Raposo Preto-Bay A. M. A frequency dictionary of Portuguese: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2008. 336 p.
34. Miller C., Aghajanian-Stewart K. A frequency dctionary of Persian: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2018. 366 p.
35. Sharoff S., Umanskaya E., Wilson J. A frequency dictionary of Russian: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2013. 400 p.
36. Aksan Y., Aksan M., Mersinli U. U., Demirhan U. U. A frequency dictionary of Turkish: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2017. 349 p.
37. Lonsdale D., Bras Y. L. A frequency dictionary of French: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2009. 320 p.
38. Cermák F., Kren M. A frequency dictionary of Czech: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2011. 296 p.
39. Tono Y., Yamazaki M., Maekawa K. A frequency dictionary of Japanese: Core vocabulary for learners. 1st ed. London, New York: Routledge; 2013. 384 p.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Денис Юрьевич Большаков, кандидат технических наук, начальник отдела научно-технических изданий и специальных проектов аппарата генерального директора АО «Концерн ВКО «Алмаз – Антей», заместитель главного редактора научно-технического журнала «Вестник Концерна ВКО «Алмаз – Антей» / Journal of "Almaz – Antey" Air and Space Defence Corporation, Москва, Российская Федерация; https://orcid.org/0000-0001-7694-1454; e-mail: press@almaz-antey.ru

INFORMATION ABOUT THE AUTHOR

Denis Yu. Bolshakov, Cand. Sci. (Eng.), Head of the Department of Scientific and Technical Issues and Special Projects of the Office of the Director General, Almaz – Antey Air and Space Defence Corporation, JSC, Deputy Editor-in-Chief of the Journal of "Almaz – Antey" Air and Space Defence Corporation, Moscow, Russian Federation; https://orcid.org/0000-0001-7694-1454; e-mail: press@almaz-antey.ru

Поступила в редакцию / Received 19.06.2025
Поступила после рецензирования / Revised 07.07.2025
Принята к публикации / Accepted 27.07.2025