ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Т. 8, № 1(25)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.8 DOI: 10.54708/SIIT-2026-no1-p127 EDN: DIIKOG

Классификация текстов на основе семантической близости с использованием встраиваемых моделей

Р. А. Ишкинин • Д. А. Ризванов

Уфимский университет науки и технологий

В статье рассматривается задача автоматической классификации русскоязычных новостных текстов с использованием методов семантического анализа на основе встраиваемых моделей. Предложенный подход основан на вычислении косинусного сходства между векторным представлением классифицируемого документа и центроидами классов в пространстве эмбеддингов. Экспериментальные результаты демонстрируют, что модель E5 обеспечивает наилучшие показатели классификации (F1-score = 0,91) без необходимости дополнительного дообучения.

Классификация текстов; обработка естественного языка; семантическая близость; встраиваемые модели; эмбеддинги; косинусное расстояние; трансформерные модели.

Введение

Современный этап развития информационного общества характеризуется экспоненциальным ростом объёмов текстовых данных. Ежедневно генерируются петабайты информации, и для новостных агентств, аналитических центров и служб безопасности способность оперативно структурировать этот поток является ключевым фактором успеха [Jou17]. Современные большие языковые модели (Large Language Models, LLM) демонстрируют впечатляющие возможности в области обработки естественного языка, однако их внедрение в промышленные системы часто сопряжено с высокими требованиями к вычислительным ресурсам [Jou16, Dev19]. Это ставит перед исследователями актуальную задачу: поиск метода, который обеспечивает высокое качество классификации при минимальных вычислительных затратах. В контексте специальности «Управление в организационных системах» данная задача приобретает особую значимость, поскольку автоматическая классификация текстов является важным компонентом систем поддержки принятия решений [Bop22, Kop24, Гус24].

Настоящая работа позиционируется как исследование в рамках создания минимально жизнеспособного продукта (MVP) для системы автоматической сортировки новостей. Ключевая гипотеза исследования заключается в том, что семантический потенциал, заложенный в предварительно обученные встраиваемые модели (embedding models), достаточен для решения задачи классификации без дорогостоящего этапа дообучения (fine-tuning) [Han12].

Исследования в области классификации русскоязычных текстов активно развиваются. Работы [Sha23, Пес22, Кот21] демонстрируют сравнительный анализ различных моделей векторных представлений для решения схожих задач, подтверждая актуальность поиска наиболее эффективных архитектур. В исследовании [Пет24] рассматриваются методы семантического

Рекомендовано к публикации программным комитетом XI Международной научной конференции ITIDS'2025 «Информационные технологии интеллектуальной поддержки принятия решений», Уфа, 13–15 ноября 2025 г.

Ишкинин Р. А., Ризванов Д. А. Классификация текстов на основе семантической близости с использованием встраиваемых моделей // СИИТ. 2026. Т. 8, № 1(25). С. 127-133. DOI: 10.54708/SIIT-2026-no1-p127. EDN: DIIKOG.

Ishkinin R. A., Rizvanov D. A. "Text classification based on semantic similarity using embedding models" // SIIT. 2026. Vol. 8, no. 1(25), pp. 127-133. DOI: 10.54708/SIIT-2026-no1-p127. EDN: DIIKOG (In Russian).

анализа на основе трансформерных моделей для специфической области нормативных документов, что подчёркивает применимость данных технологий в различных предметных областях.

Цель данного исследования — определить наиболее эффективную встраиваемую модель для ресурсосберегающей классификации новостных текстов как основного модуля систем поддержки принятия решений.

Для достижения поставленной цели необходимо решить следующие задачи:

1) сформировать и предобработать масштабный корпус русскоязычных новостных текстов;

2) реализовать алгоритм классификации на основе вычисления косинусного расстояния до центроидов классов;

3) провести сравнительный анализ производительности моделей MPNet [Son20], ruBERT [Kyp20], E5 [Wan23] и fastText [Jou16] на подготовленном корпусе;

4) выполнить количественный и качественный анализ результатов, включая разбор типовых ошибок;

5) сформулировать выводы о жизнеспособности подхода и дать рекомендации по дальнейшему развитию системы.

Материалы и методы

Корпус данных

В качестве эмпирической базы исследования использовался корпус данных, состоящий из приблизительно 10 тысяч новостных документов. Источниками послужили открытые данные популярных русскоязычных Telegram-каналов и новостных сайтов крупнейших средств массовой информации, в частности, тематические разделы «Происшествия»[1]. Весь корпус был вручную размечен и распределён по 93 тематическим категориям, каждая из которых соответствует определённому событию или теме.

Выбор данного корпуса обусловлен его методологической ценностью в качестве тестового стенда. Большое количество классов (n = 93) и семантическая близость некоторых из них создают сложную задачу, позволяющую надёжно оценить разделительную способность исследуемых моделей. Успешная апробация метода на этом материале подтверждает его жизнеспособность для переноса на другие целевые домены, например, экономические новости, которые непосредственно используются в моделях принятия управленческих решений.

Алгоритм классификации

Выбранный метод исключает этап обучения в традиционном понимании и состоит из двух последовательных шагов [Ман11, Agg18].

Первый шаг — вычисление центроидов классов. Для каждой из K категорий вычисляется вектор-центроид, представляющий собой усреднённый вектор всех документов, принадлежащих данной категории. Этот шаг выполняется однократно на имеющихся размеченных данных:

$$C_k = (1/|N_k|)\, \Sigma_{i \in N_k} \mathrm{Embed}(d_i), \tag{1}$$

где C_k — вектор-центроид для класса k; $|N_k|$ — количество документов в классе k; N_k — множество индексов документов, принадлежащих классу k; $\mathrm{Embed}(d_i)$ — функция, преобразующая документ d_i в векторное представление с помощью одной из исследуемых моделей (E5, ruBERT, MPNet, fastText).

Второй шаг — классификация нового документа. Для нового документа d_new сначала вычисляется его эмбеддинг v_new = Embed(d_new). Затем класс документа определяется

[1] Новости по тегам // Интерфакс: [сайт]. URL: https://www.interfax.ru/tags.
Происшествия в России // Коммерсантъ: [сайт]. URL: https://www.kommersant.ru/rubric/6.

путём нахождения центроида, к которому он наиболее близок в векторном пространстве. В качестве метрики близости используется косинусное сходство [Ман11, Sal75], которое эффективно измеряет семантическую близость независимо от длины векторов:

$$\text{similarity}(A, B) = (A \cdot B) \,/\, (\|A\| \cdot \|B\|). \quad (2)$$

Итоговый класс документа определяется по формуле [Ман11]:

$$\text{pred} = \text{argma}\ x_k \in 1, \ldots, K\}\ \text{similarity}(\text{v_new}, C_k). \quad (3)$$

Эксперименты и результаты

Количественный анализ

Для оценки качества классификации использовались стандартные метрики, усреднённые по всем классам: доля правильных ответов (accuracy), точность (precision), полнота (recall) и F-мера (F1-score) [Sok09]. Точность показывает, какой процент предсказанных положительных примеров действительно является положительным. Полнота отражает способность модели находить все положительные примеры в наборе данных. F-мера представляет собой гармоническое среднее между точностью и полнотой.

Результаты сравнительного анализа представлены в таблице.

Таблица

Сравнение метрик качества моделей

Модель	Precision	Recall	F1-score	Accuracy
E5	0,92	0,90	0,91	0,91
ruBERT	0,91	0,91	0,89	0,90
MPNet	0,85	0,83	0,83	0,86
fastText	0,82	0,79	0,79	0,79

Как видно из представленных данных, модель E5 демонстрирует наилучший F1-score — метрику, которая гармонично сочетает точность и полноту. Это свидетельствует о том, что данная модель обеспечивает оптимальный баланс между ошибочным отнесением документа к неправильной категории (ложноположительная ошибка) и пропуском документов, относящихся к нужной категории (ложноотрицательная ошибка). Следует отметить, что модель fastText, не учитывающая контекст слов, показывает наихудший результат среди исследуемых моделей.

Графический анализ

Для более глубокого понимания поведения моделей были построены гистограммы распределения значений косинусного сходства для истинно-положительных и истинно-отрицательных случаев (рис. 1).

Анализ гистограмм позволяет сделать следующие выводы. У моделей MPNet и ruBERT наблюдается значительное пересечение между распределениями истинно-положительных и истинно-отрицательных случаев. Это означает, что существует широкий диапазон значений сходства, в котором модель не может уверенно отличить релевантный класс от нерелевантного, что приводит к ошибкам классификации.

У модели E5 распределения разделены наиболее чётко. Распределение для истинно-отрицательных случаев смещено влево (к меньшим значениям сходства), а для истинно-положительных — вправо (к большим значениям). Область их пересечения минимальна, что объясняет более высокие показатели качества данной модели.

Качественный анализ модели E5

Проведено исследование корреляции между количеством документов в классе и процентом ошибок классификации (рис. 2). Результаты опровергают гипотезу о большей подверженности ошибкам малых классов.

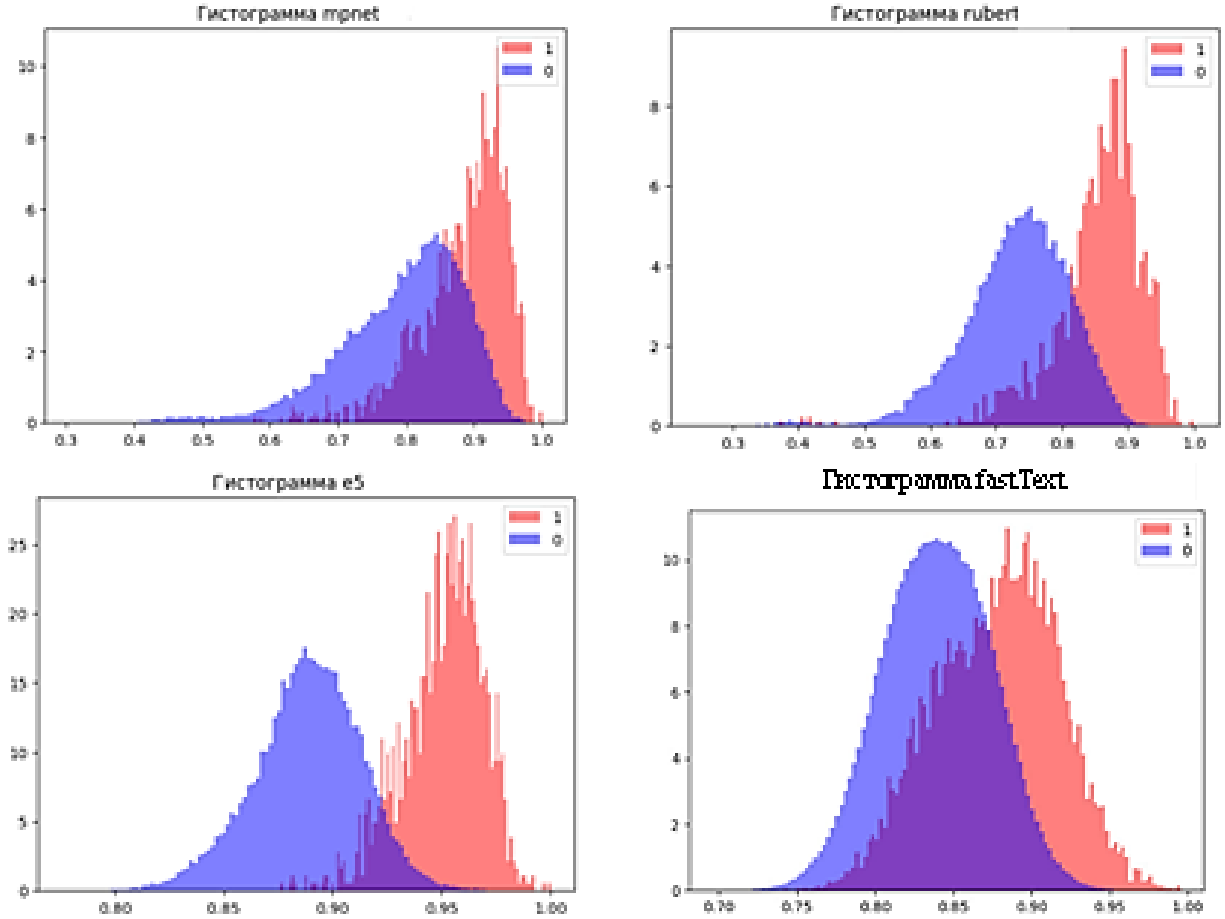

Рис. 1 Гистограммы распределения косинусного сходства для исследуемых моделей

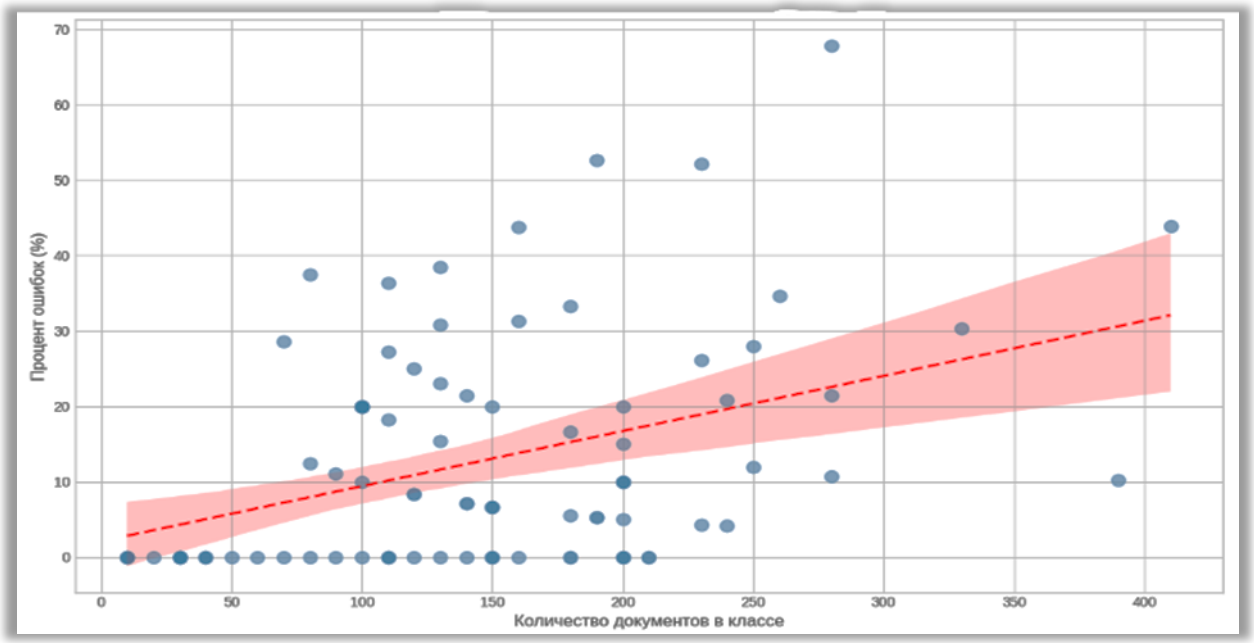

Рис. 2 Зависимость процента ошибок от количества документов в классе

Наблюдается положительная тенденция: увеличение числа документов в классе коррелирует с ростом процента ошибок. Данный феномен может быть обусловлен повышенным семантическим разнообразием внутри крупных классов, что снижает репрезентативность их центроидов.

Для оценки внутренней схожести тем была построена тепловая карта косинусного сходства между центроидами всех классов (рис. 3).

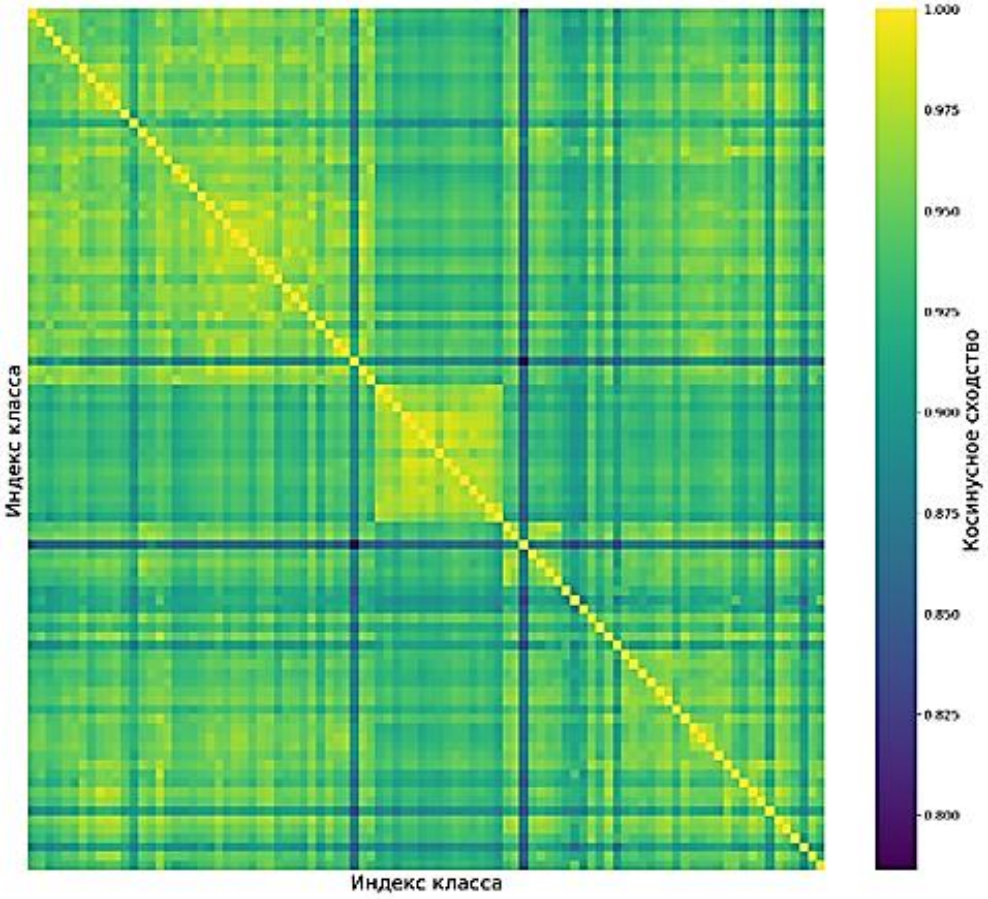

Рис. 3 Тепловая карта косинусного сходства между центроидами классов

Тепловая карта наглядно демонстрирует наличие «тематических кластеров» — групп классов, которые семантически близки друг к другу (обозначены яркими квадратами вне главной диагонали). Это указывает на то, что основной источник ошибок заключается не в недостатках модели, а в объективной сложности и смысловом пересечении некоторых категорий в наборе данных. Ошибки классификации носят системный характер и концентрируются на границах семантически близких кластеров.

Обсуждение результатов

К сильным сторонам реализованного подхода следует отнести: высокую базовую точность — использование предобученной модели E5 позволяет достичь высокого качества классификации (F1-score = 0,91) без необходимости дообучения; глубокое семантическое понимание — модель продемонстрировала способность улавливать контекстуальные нюансы, а не только поверхностные ключевые слова; простоту и скорость внедрения — расчёт центроидов и косинусного сходства являются быстрыми и не ресурсоёмкими операциями.

К слабым сторонам и точкам роста относятся: смешение семантически близких классов — основная проблема заключается в классификации документов из тем, которые объективно схожи по содержанию; недостаток специфического контекста — модель оперирует только

семантикой текста, не учитывая такие важные метаданные, как местоположение или дата события.

В качестве перспективных направлений для развития можно выделить следующие. Во-первых, обогащение векторов контекстом: предлагается усовершенствовать метод путём конкатенации эмбеддинга текста с дополнительными эмбеддингами, соответствующими извлечённым из текста сущностям (Named Entity Recognition, NER), таким как локации и даты [Rei19]. Во-вторых, иерархическая классификация: для групп семантически близких классов может быть внедрена двухуровневая система, где на первом этапе определяется общая тема, а на втором — специализированная модель уточняет подкатегорию. В-третьих, дообучение (fine-tuning) модели: при наличии достаточного объёма размеченных данных возможно дообучение модели E5 на целевом датасете [Wan23].

Заключение

Использование предобученной модели E5 для классификации текстов методом сравнения с центроидами классов показало себя как высокоэффективный базовый подход. Модель продемонстрировала глубокое понимание семантики и контекста, обеспечив точность классификации на уровне 91 %.

Полученные результаты подтверждают жизнеспособность предложенного ресурсосберегающего подхода для создания систем автоматической сортировки новостного потока в рамках систем поддержки принятия решений. Дальнейшее повышение качества классификации связано не столько с заменой модели, сколько с обогащением входных данных дополнительным контекстом и применением иерархических методов классификации.

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

[Agg18] Aggarwal C. C. Machine Learning for Text. Cham: Springer, 2018.

[Dev19] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. Minneapolis, 2019. P. 4171–4186.

[Han12] Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Waltham: Morgan Kaufmann, 2012. 744 p.

[Jou16] Joulin A., Grave E., Bojanowski P., et al. FastText.zip: Compressing text classification models // arXiv preprint arXiv:1612.03651. 2016.

[Jou17] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of Tricks for Efficient Text Classification // Proc. 15th Conf. EACL. 2017. Vol. 2. P. 427–431.

[Rei19] Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proc. EMNLP-IJCNLP. Hong Kong, 2019. P. 3982–3992.

[Sal75] Salton G., Wong A., Yang C. S. A vector space model for automatic indexing // Communications of the ACM. 1975. Vol. 18, No. 11. P. 613–620.

[Sha23] Shalyapina A., Kobozeva I. Comparative Analysis of Russian Text Embedding Models for Classification Tasks // Proc. AINL. 2023. P. 112–121.

[Sok09] Sokolova M., Lapalme G. A systematic analysis of performance measures for classification tasks // Information Processing & Management. 2009. Vol. 45, № 4. P. 427–437. EDN: YZWCTH.

[Son20] Song K., Tan X., Qin T., et al. MPNet: Masked and Permuted Pre-training for Language Understanding // Advances in NeurIPS. 2020. Vol. 33. P. 16857–16867.

[Wan23] Wang L., Yang N., Huang X., et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training // Proc. 17th Conf. EACL. Dubrovnik, 2023. P. 450–465.

[Вор22] Воронцов К. В. Лекции по алгоритмам восстановления регрессии и классификации. М., 2022. 71 с. [[Vorontsov K. V. Lectures on regression and classification algorithms. M., 2022. 71 p. (In Russian).]]

[Гус24] Гусаренко А. С., Миронов В. В. Совместная программная обработка разнородных конструкторских документов в учебном ИТ-проектировании // СИИТ. 2024. Т. 6, № 3(18). С. 102-118. EDN: QATAMS. [[Gusarenko A. S., Mironov V. V. Joint software processing of heterogeneous design documents in educational IT design // SIIT. 2024. Vol. 6, No. 3(18). P. 102-118. (In Russian).]]

[Кор24] Коровин Е. А., Чиглинцева С. А., Сазонова Е. Ю., Сметанина О. Н. Медицинская рекомендательная система на основе автоматического извлечения знаний из текстов // СИИТ. 2024. Т. 6, № 4(19). С. 111-121. EDN: OTVTXR. [[Korovina E. A., Chiglintseva S. A., Sazonova E. Yu., Smetanina O. N. Medical recommender system based on automatic knowledge extraction from texts // SIIT. 2024. Vol. 6, No. 4(19). P. 111-121. (In Russian).]]

[Кот21] Котельников Е. В., Сысоев А. А. Сравнительный анализ методов классификации коротких текстов на русском языке // Онтология проектирования. 2021. Т. 11, № 2. С. 222–234. [[Kotelnikov E. V., Sysoev A. A. "Comparative analysis of short text classification methods in Russian" // Ontology of Designing. 2021. Vol. 11, No. 2. P. 222–234. (In Russian).]]

[Кур20] Куратов Ю. А., Артамонов М. С. Адаптация многоязычных трансформерных моделей для русского языка // Труды ИСП РАН. 2020. Т. 32, № 2. С. 135–146. [[Kuratov Yu. A., Artamonov M. S. "Adaptation of multilingual transformer models for Russian" // Proc. ISP RAS. 2020. Vol. 32, No. 2. P. 135–146. (In Russian).]]

[Ман11] Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск / пер. с англ. М.: Вильямс, 2011. 528 с. EDN: QYIRXL. [[Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Moscow: Williams, 2011. (In Russian).]]

[Пес22] Пескова О. В., Романов Д. А. Применение векторных представлений текстов для задачи классификации обращений граждан // Программные продукты и системы. 2022. Т. 35, № 4. С. 605–614. [[Peskova O. V., Romanov D. A. "Application of text vector representations for citizen appeals classification" // Software & Systems. 2022. Vol. 35, No. 4. P. 605–614. (In Russian).]]

[Пет24] Петров А. В. Методы семантического анализа текстов нормативных документов на основе трансформерных моделей: дис. ... канд. техн. наук. СПб., 2024. 145 с. [[Petrov A. V. Methods of semantic analysis of regulatory documents based on transformer models: PhD thesis. St. Petersburg, 2024. (In Russian).]]

ОБ АВТОРАХ | ABOUT THE AUTHORS

ИШКИНИН Роберт Азаматович
Уфимский университет науки и технологий, Россия.
robbtish@gmail.com
Аспирант по спец. «Управление в организационных системах».

РИЗВАНОВ Дмитрий Анварович
Уфимский университет науки и технологий, Россия.
ridmi@mail.ru ORCID: 0000-0003-2378-5587.
Д-р техн. наук (Уфимск.ун-т науки и и технологий, 2019). Иссл. в обл. поддержки принятия решений в сложных системах, разработка многоагентных систем.

ISHKININ Robert Azamatovich
Ufa University of Science and Technology, Russia.
robbtish@gmail.com
Postgraduate student of the specialty "Management in organizational systems".

RIZVANOV Dmitry Anvarovich
Ufa University of Science and Technology, Russia.
ridmi@mail.ru ORCID: 0000-0003-2378-5587.
Doctor of Engineering Sciences (Ufa Univ. of Science and Technologies, 2019). Research in the field of decision support in complex systems, multi-agent systems development.

МЕТАДАННЫЕ | METADATA

Заглавие: Классификация текстов на основе семантической близости с использованием встраиваемых моделей.

Авторы: Ишкинин Р. А., Ризванов Д. А.

Аннотация: В статье рассматривается задача автоматической классификации русскоязычных новостных текстов с использованием методов семантического анализа на основе встраиваемых моделей. Предложенный подход основан на вычислении косинусного сходства между векторным представлением классифицируемого документа и центроидами классов в пространстве эмбеддингов. Экспериментальные результаты демонстрируют, что модель E5 обеспечивает наилучшие показатели классификации (F1-score = 0,91) без необходимости дополнительного дообучения.

Ключевые слова: Классификация текстов; обработка естественного языка; семантическая близость; встраиваемые модели; эмбеддинги; косинусное расстояние; трансформерные модели.

Язык: Русский.

Статья поступила в редакцию 15 февраля 2026 г.

Title: Text classification based on semantic similarity using embedding models.

Authors: Ishkinin R. A., Rizvanov D. A.

Abstract: Abstract: This paper addresses the task of automatic classification of Russian-language news texts using semantic analysis methods based on embedding models. The proposed approach relies on computing cosine similarity between the vector representation of the document being classified and class centroids in the embedding space. Experimental results demonstrate that the E5 model achieves the best classification performance (F1-score = 0.91) without additional fine-tuning.

Key words: Text classification; natural language processing; semantic similarity; embedding models; embeddings; cosine distance; transformer models.

Language: Russian.

The article was received by the editors on 15 February 2026.