Научная статья
УДК 004.91
doi: 10.34822/1999-7604-2022-1-63-71

СРАВНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ НОВОСТНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Эдуард Артурович Челышев*[1]✉*, Шамиль Алиевич Оцоков*[2]*,
***Марина Викторовна Раскатова*[3]*, Павел Щёголев*[4]**
[1, 2, 3, 4]*Национальный исследовательский университет «МЭИ», Москва, Россия*
[1]*chel.ed@yandex.ru*✉*, http://orcid.org/0000-0001-8417-8823*
[2]*Shamil24@mail.ru, http://orcid.org/0000-0001-7451-5443*
[3]*marina@raskatova.ru, http://orcid.org/0000-0001-7671-3312*
[4]*Shchegolevsp@mpei.ru, http://orcid.org/0000-0001-9954-8858*

Аннотация. В работе рассмотрена задача классификации русскоязычных новостных текстов с использованием таких алгоритмов машинного обучения, как наивный байесовский классификатор, случайный лес деревьев решений, логистическая регрессия и искусственная нейронная сеть. Для решения задачи использовались тексты новостного интернет-портала Lenta.ru, относящиеся к девяти различным классам – рубрикам новостных статей. Программная реализация в рамках данной работы проводилась с использованием языка программирования Python. Проведена предварительная обработка текстовых данных: удаление нерелевантных символов и приведение их к общему регистру, токенизация, нормализация, удаление стоп-слов и векторизация текстов. Для реализации искусственной нейронной сети в рамках данной работы использовались библиотеки Tensorflow и Keras языка программирования Python. Для каждой из использованных моделей машинного обучения были определены значения гиперпараметров, дающих наивысшее качество классификации, с использованием ряда метрик: precision, recall и F-мера. Проведен сравнительный анализ использованных алгоритмов. Указаны возможные пути дальнейшей работы в рамках рассматриваемой задачи.

Ключевые слова: классификация текстов, машинное обучение, корпус, токен, стоп-слово, лемматизация, искусственная нейронная сеть, качество классификации

Для цитирования: *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 (42). С. 63–71. DOI 10.34822/1999-7604-2022-1-63-71.*

Original article

COMPARING CLASSIFICATION METHODS FOR NEWS TEXTS IN RUSSIAN USING MACHINE LEARNING ALGORITHMS

***Eduard A. Chelyshev*[1]✉*, Shamil A. Otsokov*[2]*, Marina V. Raskatova*[3]*, Pavel Shchegolev*[4]**
[1, 2, 3, 4]*National Research University "Moscow Power Engineering Institute", Moscow, Russia*
[1]*chel.ed@yandex.ru*✉*, http://orcid.org/0000-0001-8417-8823*
[2]*Shamil24@mail.ru, http://orcid.org/0000-0001-7451-5443*
[3]*marina@raskatova.ru, http://orcid.org/0000-0001-7671-3312*
[4]*Shchegolevsp@mpei.ru, http://orcid.org/0000-0001-9954-8858*

Abstract. The article discusses the problem of classification of news texts in Russian using such machine learning algorithms as naive Bayes classifier, random decision forests, logistic regression, and artificial neural network. The texts of the Internet news portal Lenta.ru were selected from nine different classes – sections of news articles to solve the problem. The software implementation in the framework of the study was carried out using the Python programming language. The preprocessing of text data included removal of

irrelevant characters and their reduction to a common register, tokenization, normalization, removal of stop words and vectorization of texts. Tensorflow and Keras libraries of the Python programming language were used to implement an artificial neural network. For each of the machine learning models used, hyperparameters values were determined in order to achieve the highest classification quality using a number of metrics: precision, recall and F-measure. A comparative analysis of the algorithms used was carried out. Possible ways for further study within the problem in question are specified.

Keywords: text classification, machine learning, corpus, token, stop word, lemmatization, artificial neural network, classification quality

For citation: *Chelyshev E. A., Otsokov Sh. A., Raskatova M. V., Shchegolev P. Comparing Classification Methods for News Texts in Russian Using Machine Learning Algorithms // Proceedings in Cybernetics. 2022. No. 1 (42). P. 63–71. DOI 10.34822/1999-7604-2022-1-63-71.*

ВВЕДЕНИЕ

Для современного мира характерен быстрый рост объема информационных баз данных, скорости обработки информации и объема хранения данных [1]. Именно поэтому в последние годы все более популярными становятся средства автоматической обработки информации, в том числе основанные на алгоритмах машинного обучения.

Для обработки естественного языка (англ. natural language processing) возможно применение машинного обучения, которое используется во многих сферах: перевод, поиск информации, распознавание речи, определение ее строения и тональности, система электронного документооборота и пр. [2].

Одной из задач обработки естественного языка является задача классификации текстов, которая может использоваться в новостных агрегаторах, рубрикаторах научных текстов и пр. При этом стоит отметить, что автоматическая классификация текстов значительно эффективнее ручной обработки больших объемов информации [3].

МАТЕРИАЛЫ И МЕТОДЫ

В работе был использован корпус новостных статей интернет-портала lenta.ru [4]. Условимся, что в дальнейшем в работе статьи корпуса будут именоваться документами. Для каждого документа приведены его содержание и заголовок, дата публикации и URL-ссылка на веб-страницу с оригинальным текстом, рубрика (т. е. сфера общественной жизни, к которой относится данная публикация). Из исходного корпуса были выделены документы, относящиеся к девяти рубрикам: «Дом», «Интернет и СМИ», «Культура», «Наука и техника», «Политика», «Путешествия», «Силовые структуры», «Спорт», «Экономика и бизнес». Из выделенных документов был сформирован итоговый корпус.

Проведена предварительная обработка текстов документов, включающая следующие этапы: удаление нерелевантных символов и приведение их к общему регистру, токенизация, нормализация, удаление стоп-слов и векторизация [5].

Удаление нерелевантных символов в данной работе выполняли с использованием регулярных выражений. В качестве нерелевантных рассматривали все символы, за исключением буквенных и пробелов. Также были удалены присутствующие в тексте URL-ссылки. Все оставшиеся в текстах символы были преобразованы в строчную форму. На этапе токенизации текст был разбит на отдельные токены, т. е. слова. Для выполнения нормализации использовали лемматизацию [6], для лемматизации токенов – морфологический анализатор русского языка, реализованный в библиотеке Pymorphy2 языка программирования Python [7]. Затем из числа токенов, преобразованных в ходе нормализации, были удалены стоп-слова, т. е. часто встречающиеся, но не несущие существенной лексической нагрузки слова [8].

Векторизация текстов проводилась в два этапа. На первом этапе каждому токену из документа корпуса устанавливали соответствие его векторному представлению. На втором – вычисляли векторное представление документа как среднее арифметическое векторных представлений всех токенов, входящих в документ.

Векторизация токенов производилась с использованием модели векторизации FastText, предобученной на русскоязычных текстах [9]. Модели векторизации в сравнении с некоторыми другими методами векторизации текстов (например, мешка слов) имеют ряд преимуществ. Во-первых, векторные представления токенов, сгенерированные моделями векторизации, имеют размерность значительно меньшую, чем размерность используемого словаря. Так, например, в данной работе токенам, а следовательно, и документам ставилось в соответствие 300-мерное векторное представление. Во-вторых, модели векторизации учитывают семантические, т. е. смысловые отношения токенов [10].

Подготовленные таким образом данные были разделены на обучающую и тестовую выборки, размер тестовой выборки составил 25 % от общего объема корпуса.

Постановка задачи. В работе рассматривается задача классификации новостных текстов. Учитывая проведенную предварительную обработку текстов, данную задачу можно формализовать следующим образом: пусть имеется документ $d_i \in D, i = 1, 2, \ldots, K$, где $D = \{d_1, d_2, \ldots, d_K\}$ – множество документов в корпусе, а K – размерность корпуса, причем документ представлен как вектор n-мерного векторного пространства, т. е. $d = (v_1, v_2, \ldots, v_n)$. В данной работе $n = 300$. Пусть также задан некий фиксированный набор классов $C = \{c_1, c_2, \ldots, c_m\}$. В данной работе $m = 9$ по числу рубрик в итоговом корпусе.

Используя обучающую выборку с применением метода обучения, необходимо получить классифицирующую функцию $g: D \to C$, которая отображает множество документов во множестве классов [11].

Построение классификаторов. Первым рассмотренным классификатором в данной работе выступает наивный байесовский классификатор (НБК, англ. Naive Bayes), который строится на «наивном» предположении о независимости признаков.

Если пространство признаков принимается непрерывным, то при построении НБК зачастую предполагается, что вероятность появления признаков задается нормальным распределением. Тогда вероятность того, что объект, обладающий значением x_i некоторого i-го признака, принадлежит классу C_k, задается с использованием формулы (1). Такой алгоритм называется гауссовским НБК (англ. Gaussian Naive Bayes) [12]. С учетом особенностей выполненной векторизации текстового корпуса именно такая разновидность НБК является подходящей для решения рассматриваемой задачи:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right), \qquad (1)$$

где σ_k – среднеквадратичное отклонение значений i-го признака для объектов класса C_k;

μ_y – математическое ожидание значений i-го признака для объектов класса C_k.

В работе был также рассмотрен классификатор на основе метода логистической регрессии (ЛР, англ. Logistic Regression). Данный метод является методом бинарной классификации и позволяет для каждого объекта получить значение вероятности принадлежности данного объекта к каждому из классов. Пусть одному классу соответствует значение $y = -1$, а другому – $y = +1$, т. е. множество возможных меток $Y = \{-1, +1\}$. Тогда вероятность того, что объект x принадлежит классу y, определяется по формуле (2):

$$P(y|x) = \sigma(\langle w, x\rangle y), \qquad (2)$$

где w – вектор весов;

$\langle x, y\rangle$ – скалярное произведение векторов x и y;

$\sigma(z)$ – логистическая функция, определяемая по формуле (3):

$$\sigma(z) = \frac{1}{1+e^{-z}}. \qquad (3)$$

В случае, если классов более, чем два, задачу мультиклассификации можно рассматривать как ряд задач бинарной классификации, каждую из которых можно решить, используя ЛР. Кроме того, метод ЛР может быть обобщен для задачи мультиклассификации. Предположим, что в задаче мультиклассификации присутствует N непересе-

кающихся классов. Тогда многоклассовая ЛР подразумевает построение набора из N линейных моделей, каждая из которых имеет свой собственный вектор весов w_k, $k = 1, 2, \dots, N$. Затем для каждого объекта, используя построенные линейные модели, можно определить численную оценку $z_k = \langle w_k, x \rangle$ принадлежности объекта к k-му классу. Для преобразования вектора численных оценок в вектор вероятностей используется функция Softmax, которая представляет собой многомерное обобщение логистической функции, т. е. вектор p вероятностей принадлежности объекта к каждому из классов вычисляется по формуле (4) [13]:

$$p = (p_1, p_2, \dots, p_N) = \mathrm{S}(z_1, \mathrm{z}_2, \dots, \mathrm{z}_N) = \left(\frac{\exp(\mathrm{z}_1)}{\sum_{\mathrm{k}=1}^{\mathrm{N}} \exp(\mathrm{z}_\mathrm{k})}, \dots, \frac{\exp(\mathrm{z}_\mathrm{N})}{\sum_{\mathrm{k}=1}^{\mathrm{N}} \exp(\mathrm{z}_\mathrm{k})}\right), \quad (4)$$

где N – количество классов в задаче мультиклассификации;

p_k – вероятность принадлежности к k-му классу.

В работе рассмотрен классификатор на основе алгоритма случайного леса (СЛ, англ. Random Forest) дерева решений (СЛДР). Стоит отметить, что сам по себе алгоритм на основе дерева решений (ДР) имеет существенный недостаток, а именно склонность к переобучению. Переобучением называется нежелательное явление, при котором обученный алгоритм машинного обучения на данных, не входящих в обучающую выборку, показывает качество значительно хуже, чем на обучающей выборке. Причиной переобучения является излишне точное повторение алгоритмом зависимостей обучающей выборки, из-за чего он теряет обобщающую способность в целом. По этой причине использование ансамблевого метода, такого как СЛДР, представляется более оправданным [14].

Алгоритм СЛДР заключается в построении некоторого множества различных ДР на одном и том же наборе данных. На этапе обучения при построении каждого из деревьев в отдельности производится случайный выбор некоторого подмножества объектов внутри обучающей выборки. Именно выбранные таким образом объекты в дальнейшем будут использоваться для обучения этого ДР. Для каждого разветвления внутри ДР модель также случайным образом выбирает учитываемые признаки. Таким образом, возможное переобучение каждого из деревьев в отдельности на своем наборе данных компенсируется за счет усреднения по множеству различных ДР [15].

При построении ДР был использован алгоритм CART (англ. Classification and Regression Tree). Для задания функционала качества при построении решающего дерева в данной работе использовался критерий информативности Джини, который вычисляется по формуле (5), при этом нужно отметить, что, чем меньше критерий Джини для вершины ДР, тем меньше неопределенность в ней:

$$G(p) = 1 - \sum_{i=1}^{n} p_i^2, \quad (5)$$

где T – набор объектов в вершине;

n – число классов, содержащихся в наборе T;

$p = (p_1, p_2, \dots, p_n)$ – вектор вероятностей (относительных частот) для каждого из классов в наборе T.

Последним использованным классификатором является искусственная нейронная сеть (ИНС), реализованная с использованием библиотек Tensorflow и Keras языка программирования Python (рис. 1).

ИНС содержит четыре слоя. Входной слой является полносвязным. Он был реализован с использованием встроенного класса Dense библиотеки Keras. Входной слой содержит 300 нейронов по числу компонент векторного представления документа. Первый скрытый слой был выполнен с использованием класса Dropout библиотеки Keras. Во время обучения ИНС данный слой случайным образом отключает некоторую долю своих нейронов, в результате чего уменьшается вероятность переобучения. Доля отключаемых нейронов определяется коэффициентом отключения, который также может рассматриваться как гиперпараметр модели. Второй скрытый слой был реализован с использованием класса BatchNormalization. Он осуществляет статистическую нормализацию выхода предыдущего слоя. В качестве функ-

ции активации на скрытых слоях используется функция ReLU. Выходной слой также является полносвязным (класс Dense библиотеки Keras) и содержит 9 нейронов, каждый из которых соответствует отдельной рубрике. Функция Softmax используется в качестве функции активации выходного слоя [16].

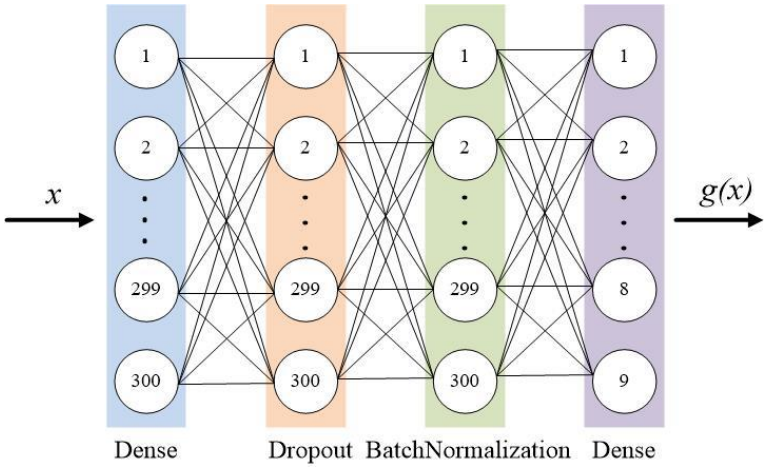

Рис. 1. Структура искусственной нейронной сети
Примечание: составлено авторами.

Оценка качества классификации. Для оценки качества классификации были использованы метрики precision (точность) J_p и recall (полнота) J_r, определяемые формулами (6) и (7) соответственно. Метрика precision является долей верно классифицированных объектов в общем числе объектов, отнесенных классификатором к данному классу. Метрика recall указывает, насколько полно классификатор распознал класс: чем меньше отношение числа объектов в классе, которые были ошибочно отнесены к другому классу, к числу объектов данного класса, тем выше значение метрики recall. В работе использовалась также F_β-мера, которая является комбинированной метрикой, в которой параметр β является весом метрики precision. Частным случаем данной метрики при β = 1 является F_1-мера J_F, определяемая формулой (8), которая и использовалась в данной работе. Также при обучении ИНС для контроля возможного переобучения использовалась метрика accuracy J_a, задаваемая формулой (9). Данная метрика равняется доле верно классифицированных объектов в общем числе объектов всех классов [17]:

$$J_p = \frac{G_p^+}{G_p^+ + G_p^-}, \quad (6)$$

$$J_r = \frac{G_p^+}{G_p^+ + G_n^-}, \quad (7)$$

где G_p^+ – число объектов, для которых классификатор верно определил принадлежность к текущему классу (англ. true positives);

G_n^+ – число объектов, для которых классификатор верно определил, что они не принадлежат к текущему классу (англ. true negatives);

G_p^- – число объектов, для которых классификатор неверно определил, что они принадлежат к текущему классу (англ. false positives);

G_n^- – число объектов, для которых классификатор неверно определил, что они не принадлежат к текущему классу (англ. false negatives).

$$J_F = 2 \times \frac{J_p \times J_r}{J_p + J_r}, \quad (8)$$

$$J_a = \frac{G_p^+ + G_n^+}{G_p^+ + G_p^- + G_n^+ + G_n^-}. \quad (9)$$

Рассмотренные выше метрики классификации предназначены для бинарной классификации. В случае мультиклассификации данные метрики вычисляются для каждого класса в отдельности, после чего можно вычислить средние значения метрик как пока-

затели качества построенного классификатора в целом.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Программные реализации НБК, ЛР и СЛДР были взяты из библиотеки scikit-learn языка программирования Python. С использованием алгоритма решетчатого поиска с применением скользящего контроля для данных моделей классификации были определены значения гиперпараметров, дающих наилучшее качество классификации. Обучение ИНС велось при числе эпох $m = 50$. В ходе обучения на каждой эпохе производили измерения величин $J_a^{\text{обуч}}$ и $J_a^{\text{тест}}$, т. е. значений метрики accuracy на обучающей и тестовой выборках соответственно (рис. 2).

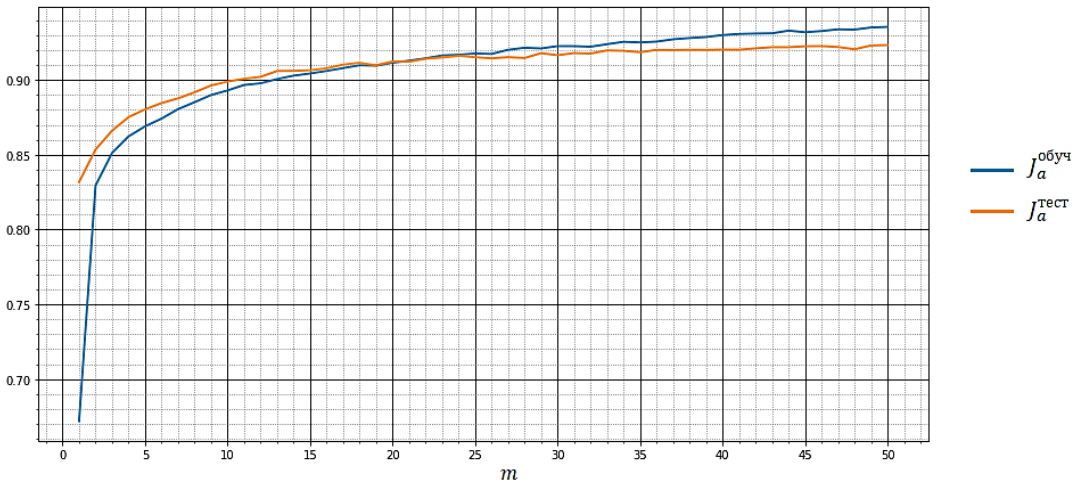

Рис. 2. Зависимость значений метрики accuracy от числа эпох
Примечание: составлено авторами на основе экспериментальных данных.

Для каждого из классификаторов были определены средние значения метрик классификации (табл. 1, рис. 3).

Таблица 1

Сводная таблица значений метрик классификации

Классификатор	J_p	J_r	J_F
Наивный байесовский классификатор	0,81459	0,79775	0,75367
Логистическая регрессия	0,90216	0,90236	0,90222
Случайный лес решающих деревьев	0,88318	0,88310	0,88221
Искусственная нейронная сеть	0,9253	0,9250	0,9251

Примечание: составлено авторами на основе экспериментальных данных.

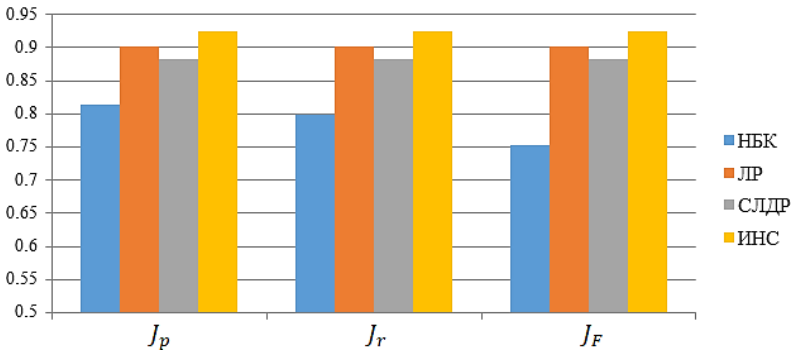

Рис. 3. Гистограмма значений метрик классификации
Примечание: составлено авторами.

Видно, что наибольшие значения метрик классификации показали разработанная авторами ИНС и ЛР. При этом достаточно близкие к ним значения показал и СЛДР. Наивный байесовский классификатор как простейшая модель при этом показал значительно более низкое качество классификации.

Проведем анализ двух классификаторов, показавших наилучшие результаты (табл. 2).

Таблица 2

Значения метрик классификации для ИНС и ЛР

Класс (рубрика)	ИНС			ЛР		
	J_p	J_r	J_F	J_p	J_r	J_F
Дом	0,959	0,9241	0,941	0,874	0,849	0,862
Интернет и СМИ	0,868	0,883	0,876	0,829	0,812	0,820
Культура	0,943	0,931	0,937	0,929	0,936	0,932
Наука и техника	0,918	0,901	0,909	0,909	0,905	0,907
Политика	0,883	0,912	0,897	0,858	0,873	0,866
Путешествия	0,955	0,965	0,960	0,834	0,823	0,829
Силовые структуры	0,919	0,934	0,927	0,851	0,837	0,844
Спорт	0,977	0,986	0,982	0,979	0,977	0,978
Экономика и бизнес	0,903	0,891	0,897	0,907	0,918	0,913

Примечание: составлено авторами на основе экспериментальных данных.

На основании данных табл. 2 можно заключить, что ИНС распознала лучше, чем ЛР восемь классов (рубрик) из девяти. Исключением является только рубрика «Экономика и бизнес», для которой наивысшие значения метрик классификации показала именно ЛР. Такой результат можно объяснить тем, что данная рубрика являлась самой объемной в используемом наборе данных и была представлена наибольшим количеством документов.

ЗАКЛЮЧЕНИЕ

В ходе проведенного исследования рассматривалось решение задачи классификации русскоязычных новостных текстов с применением алгоритмов машинного обучения. Использованный корпус документов (новостных статей) был предварительно обработан и каждому документу был поставлен в соответствие вектор 300-мерного векторного пространства. Были реализованы и обучены четыре классификатора: наивный байесовский классификатор, логистическая регрессия, случайный лес деревьев решений и предложенная коллективом авторов искусственная нейронная сеть. В целом наилучшее качество при этом показала именно искусственная нейронная сеть. При этом рубрика «Экономика и бизнес» была распознана наилучшим образом логистической регрессией.

Однако задачу классификации текстов вообще и новостных в частности нельзя считать окончательно решенной. Стоит отметить, что в работе не использовались некоторые другие алгоритмы классификации, например метод k ближайших соседей и машина опорных векторов.

Авторы планируют продолжить исследования в двух направлениях. Во-первых, прирост качества классификации могут показать искусственные нейронные сети с другими архитектурами. Во-вторых, полезным может оказаться использование ансамблей различных методов, как рассмотренных выше, так и тех, которые еще не были проанализированы в данной работе.

Список источников

1. Reinsel D., Gantz J., Rydning J. The Digitalization of the World – From Edge to Core. IDC White Paper, 2018. 28 p. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (дата обращения: 11.01.2022).

References

1. Reinsel D., Gantz J., Rydning J. The Digitalization of the World – From Edge to Core. IDC White Paper, 2018. 28 p. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (accessed: 11.01.2022).

2. Батура Т. В. Методы автоматической классификации текстов // Программн. продукты и системы. 2017. Т. 30, № 1. С. 85–99.
3. Шаграев А. Г. Модификация, разработка и реализация методов классификации новостных текстов : дис. … канд. техн. наук. М., 2014. 108 с.
4. News Dataset from Lenta.ru. URL: https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta (дата обращения: 08.02.2022).
5. Челышев Э. А., Оцоков Ш. А., Раскатова М. В. Автоматическая рубрикация текстов с использованием алгоритмов машинного обучения // Вестн. Рос. нового ун-та. Сер.: Сложные системы: модели, анализ, управление. 2021. № 4. С. 175–182. DOI 10.25586/RNU.V9187.21.04.P.175.
6. Вершинин Е. В., Тимченко Д. К. Исследование применения стемминга и лемматизации при разработке систем адаптивного перевода текста // Наука. Исследования. Практика : сб. изб. ст. по материалам Междунар. науч. конф. СПб. : Гуманитар. национал. исслед. ин-т «НАЦРАЗВИТИЕ», 2020. С. 77–79.
7. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Proceedings of the 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015. P. 330–342. DOI 10.1007/978-3-319-26123-2_31.
8. Мартынов В. А., Плотникова Н. П. Нормализация и фильтрация текста для задачи кластеризации // XLVIII Огаревские чтения : материалы науч. конф. В 3 ч. Саранск, 06–13 декабря 2019 г. Саранск : Национал. исслед. Мордов. гос. ун-т им. Н. П. Огарева, 2020. С. 448–452.
9. Korogodina O., Klyshinsky E., Karpik O. Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings // Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2. Saint Petersburg, 2020.
10. Жеребцова Ю. А., Чижик А. В. Сравнение моделей векторного представления текстов в задаче создания чатбота // Вестник НГУ. Сер.: Лингвистика и межкультурная коммуникация. 2020. Т. 18, № 3. С. 16–34. DOI 10.25205/1818-7935-2020-18-3-16-34.
11. Рубцова Ю. С. Методы и алгоритмы построения информационных систем для классификации текстов социальных сетей по тональности : дис. … канд. техн. наук. Новосибирск, 2019. 141 с.
12. Fadlil A., Riadi I., Aji S. DDoS Attacks Classification Using Numeric Attribute-Based Gaussian Naive Bayes // International Journal of Advanced Computer Science and Applications. 2017. No. 8. P. 42–50. DOI 10.14569/IJACSA.2017.080806.
13. Aggarwal C. C., Zhai C. Mining Text Data. Boston : Springer, 2012. 524 p.
14. Полин Я. А., Зудилова Т. В., Ананченко И. В., Войтюк Т. Е. Деревья решений в задачах классификации: особенности применения и методы повышения качества классификации // Современ.

2. Batura T. V. Automatic Text Classification Methods // Software & Systems. 2017. Vol. 30, No. 1. P. 85–99. (In Russian).
3. Shagraev A. G. Modifikatsiia, razrabotka i realizatsiia metodov klassifikatsii novostnykh tekstov : Cand. Sci. Dissertation (Engineering). Moscow, 2014. 108 p. (In Russian).
4. News Dataset from Lenta.ru. URL: https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta (accessed: 08.02.2022).
5. Chelyshev E. A., Otsokov Sh. A., Raskatova M. V. Automatic Text Rubrication Using Machine Learning Algorithms // Vestnik of Russian New University. Series: Complex Systems: Models, Analysis, Management. 2021. No. 4. P. 175–182. DOI 10.25586/RNU.V9187.21.04.P.175. (In Russian).
6. Vershinin E. V., Timchenko D. K. Research of the Application of Stemming and Lemmatization Applying to Adaptive Text Translation System // Nauka. Issledovaniia. Praktika : Collection of selected articles in proceedings of the International Research Conference. Saint Petersburg : Gumanitar. national. issled. in-t “NATsRAZVITIE”, 2020. P. 77–79. (In Russian).
7. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Proceedings of the 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015. P. 330–342. DOI 10.1007/978-3-319-26123-2_31.
8. Martynov V. A., Plotnikova N. P. Text Normalization and Filtration for Clustering Task // XLVIII Ogarevskie chteniia : Proceedings of the Research Conference. In 3 parts. Saransk, December 6–13, 2019. Saransk : Ogarev Mordovia State University, 2020. P. 448–452. (In Russian).
9. Korogodina O., Klyshinsky E., Karpik O. Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings // Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2. Saint Petersburg, 2020.
10. Zherebtsova Yu. A., Chizhik A. V. Text Vectorization Methods for Retrieval-Based Chatbot // NSU Vestnik. Series: Linguistics and Intercultural Communication. 2020. Vol. 18, No. 3. P. 16–34. DOI 10.25205/1818-7935-2020-18-3-16-34. (In Russian).
11. Rubtsova Yu. S. Metody i algoritmy postroeniia informatsionnykh system dlia klassifikatsii tekstov sotsialnykh setei po tonalnosti : Cand. Sci. Dissertation (Engineering). Novosibirsk, 2019. 141 p. (In Russian).
12. Fadlil A., Riadi I., Aji S. DDoS Attacks Classification Using Numeric Attribute-Based Gaussian Naive Bayes // International Journal of Advanced Computer Science and Applications. 2017. No. 8. P. 42–50. DOI 10.14569/IJACSA.2017.080806.
13. Aggarwal C. C., Zhai C. Mining Text Data. Boston : Springer, 2012. 524 p.
14. Polin Ya. A., Zudilova T. V., Ananchenko I. V., Voityuk T. E. Decision Trees in Classification Prob-

наукоемкие технологии. 2020. № 9. С. 59–63. DOI 10.17513/snt.38215.
15. Bertsimas D., Dunn J. Optimal classification trees // Machine Learning. 2017. Vol. 106. P. 1039–1082.
16. Челышев Э. А., Оцоков Ш. А., Раскатова М. В. Разработка информационной системы для автоматической рубрикации новостных текстов // Междунар. журн. информацион. технологий и энергоэффективности. 2021. Т. 6, № 3 (21). С. 11–17.
17. Vujovic Z. D. Classification Model Evaluation Metrics // International Journal of Advanced Computer Science and Applications. 2021. Vol. 12, No. 6. P. 599–606.

Информация об авторах

Э. А. Челышев – магистрант.
Ш. А. Оцоков – доктор технических наук.
М. В. Раскатова – кандидат технических наук.
П. Щёголев – ассистент.

lems: Application Features and Methods for Improving the Quality of Classification // Modern High Technologies. 2020. No. 9. P. 59–63. DOI 10.17513/snt.38215. (In Russian).
15. Bertsimas D., Dunn J. Optimal Classification Trees // Machine Learning. 2017. Vol. 106. P. 1039–1082.
16. Chelyshev E. A., Otsokov Sh. A., Raskatova M. V. Development of Information System for Automatic Rubrication of News Texts // Mezhdunar. zhurn. information. tekhnologii i energoeffektivnosti. 2021. Vol. 6, No. 3 (21). P. 11–17. (In Russian).
17. Vujovic Z. D. Classification Model Evaluation Metrics // International Journal of Advanced Computer Science and Applications. 2021. Vol. 12, No. 6. P. 599–606.

Information about the authors

E. A. Chelyshev – Master's Degree Student.
Sh. A. Otsokov – Doctor of Sciences (Engineering).
M. V. Raskatova – Candidate of Sciences (Engineering).
P. Shchegolev – Assistant Professor.