DOI: 10.18137/RNU.V9187.22.03.P.185 УДК 004.89

Э.А. Челышев, Ш.А. Оцоков, М.В. Раскатова, П. Щёголев

РАЗРАБОТКА СИСТЕМЫ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ НОВОСТНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Аннотация. Представлена разработка системы тематической классификации новостных текстов с использованием алгоритмов машинного обучения с использованием выборки новостных статей, каждая из которых относится к одной из девяти рубрик. Описан способ подготовки текстовых данных для их последующей классификации. Для векторизации документов применяется модель векторизации FastText. Для построения классификаторов использованы четыре различных алгоритма классификации. Проведена оценка качества построенных классификаторов по ряду метрик. Также представлено разработанное в рамках системы тематической классификации веб-приложение и его интерфейс.

Ключевые слова: обработка естественного языка, машинное обучение, классификация, рубрика, нормализация, метрика, веб-приложение.

E.A. Chelyshev, Sh.A. Otsokov, M.V. Raskatova, P. Shchegolev

DEVELOPMENT OF NEWS TEXT THEMATIC CLASSIFICATION SYSTEM USING MACHINE LEARNING ALGORITHMS

Abstract. The article describes the development of a system of thematic classification of news texts using machine learning algorithms. The news article dataset is used in the work, each article belonging to one of nine headings. The method of text data preparation for their subsequent classification is described. The FastText vectorization model is used for document vectorization. Four different classification algorithms were used to build classifiers. The quality of the constructed classifiers was evaluated according to a number of metrics. The paper also describes a web application developed within the framework of the thematic classification system and its interface.

Keywords: natural language processing, machine learning, classification, category, normalization, performance measure, web application.

Введение

Информационный объем накопленных человечеством данных неуклонно растает. Безусловно, в таких условиях становится все сложнее ориентироваться в информационном потоке. По этой причине возрастает роль различных средств автоматической обработки информации.

Одним из средств автоматической обработки информации является система определения тематической принадлежности текста (автоматической рубрикации). Автоматическое определение тематической принадлежности текста может использоваться, например, в новостных агрегаторах, рубрикаторах научных текстов, системах электронного документооборота, поиске информации и др. [2; 6;15].

В данной работе рассматривается задача автоматической рубрикации русскоязычных новостных статей. Результатом является система тематической классификации, определяющая принадлежность новостной статьи к одной из девяти преопределенных рубрик, соответствующих различным сферам общественной жизни.

Челышев Эдуард Артурович
магистрант. Московский энергетический институт (Национальный исследовательский университет), Москва. Сфера научных интересов: машинное обучение по прецедентам, искусственные нейронные сети, машинная обработка текстов на естественном языке, C++. Автор более 10 опубликованных научных работ.
Электронный адрес: chel.ed@yandex.ru

Оцоков Шамиль Алиевич
доктор технических наук, доцент кафедры вычислительных машин, систем и сетей. Московский энергетический институт (Национальный исследовательский университет), Москва. Сфера научных интересов: машинное обучение, машинная арифметика. Автор более 30 опубликованных научных работ.
Электронный адрес: Shamil24@mail.ru

Раскатова Марина Викторовна
кандидат технических наук, доцент кафедры вычислительных машин, систем и сетей. Московский энергетический институт (Национальный исследовательский университет), Москва. Сфера научных интересов: разработка программного обеспечения, информационные системы. Автор 40 опубликованных научных работ.
Электронный адрес: marina@raskatova.ru

Щёголев Павел
старший преподаватель кафедры вычислительных машин, систем и сетей. Московский энергетический институт (Национальный исследовательский университет), Москва. Сфера научных интересов: разработка программного обеспечения, языки и методы программирования.
Электронный адрес: ShchegolevsP@mpei.ru

Стоит отметить, что вопросу тематической классификации посвящен ряд публикаций в отечественных и иностранных изданиях. Так, например, статья [17] посвящена решению проблемы классификации русскоязычных новостных текстов. В работе выполнена предварительная обработка используемых текстов. Для векторизации документов применялся метод TF-IDF. Для решения задачи использовались различные алгоритмы машинного обучения, два из которых показали наилучшие результаты – RuBERT (англ. Russian Bidirectional Encoder Representations from Transformers), предобученный на русскоязычных текстах, и машина опорных векторов. Средние значения F1-меры для каждой из моделей равны 0,882 и 0,877 соответственно.

В данной работе для решения задачи тематической классификации новостных текстов используются такие алгоритмы машинного обучения, как наивный байесовский классификатор, случайный лес решающих деревьев, логистическая регрессия и искусственная нейронная сеть. Подробно рассмотрен процесс подготовки текстовых данных для их машинной обработки. В статье также представлено веб-приложение разработанной системы тематической классификации.

Исходные данные

В рамках данной работы использовался текстовый корпус, содержащий новостные статьи интернет-портала Lenta.ru за период с 1999 по 2019 год, представленный в виде файла формата CSV. Для каждой статьи приведены ее содержание и заголовок, дата пу-

бликации и URL-ссылка на нее. Кроме того, указана рубрика, к которой относится данная публикация.

Из текстового корпуса были выделены новостные статьи, каждая из которых относится к одной из девяти рубрик: «Дом», «Интернет и СМИ», «Культура», «Наука и техника», «Политика», «Путешествия», «Силовые структуры», «Спорт», «Экономика и бизнес». Распределение количества статей по рубрикам представлено в Таблице 1. В дальнейшем условимся называть новостные статьи документами.

Таблица 1

Распределение документов по рубрикам

Рубрика	Количество статей
Дом	21734
Интернет и СМИ	44663
Культура	53796
Наука и техника	53136
Политика	40716
Путешествия	6408
Силовые структуры	19596
Спорт	64413
Экономика и бизнес	86926

Рассматриваемый текстовый корпус не является сбалансированным: самая объемная рубрика «Экономика и бизнес» содержит 86 926 документов, в то время как в наименьшую по объему рубрику «Путешествия» входит только 6408 документов, что меньше количества документов в предыдущей рубрике более чем в 13 раз. Для обучения моделей классификации была выполнена балансировка текстового корпуса. При балансировке удалялись преимущественно статьи, которые имели более раннюю дату публикации, в то время как недавние по дате публикации документы были сохранены.

Полученные данные были подготовлены с использованием языка программирования Python. Подготовка включала в себя следующие этапы.

1. Удаление нерелевантных символов и приведение символов в нижний регистр.
2. Токенизация.
3. Нормализация.
4. Удаление стоп-слов [13].

На первом этапе с использованием регулярных выражений были удалены присутствующие в текстах документов URL-ссылки, а также все нерелевантные символы, к которым в данной работе относятся все небуквенные символы, исключая пробелы. Все сохранившиеся символы были приведены в нижний регистр. В ходе токенизации получившиеся тексты были разбиты на отдельные смысловые единицы – токены.

Нормализация в обработке естественного языка – это процесс приведения различных форм одного токена к нормальной форме. Данная задача может быть решена методами морфологического анализа – стеммингом и лемматизацией [1].

Стемминг (англ. stemming) – метод морфологического анализа, при котором происходит отсечение от основы некоторых частей слова (префиксов, суффиксов и окончаний)

[7]. Существенным недостатком данного метода является высокая вероятность ошибки, которая может проявляться в недостаточном или избыточном стеммировании [10].

Лемматизация (англ. lemmatization) – метод морфологического анализа, при котором происходит приведение слова к его начальной, то есть словарной, форме, называемой леммой. Лемматизация гарантирует достоверное приведение к начальной форме, однако при этом является более требовательной к вычислительным ресурсам [9].

В данной работе для решения задачи нормализации текстов применялась лемматизация, которая осуществлялась с использованием морфологического анализатора русского языка, реализованного в библиотеке pymorphy2 языка программирования Python [14].

Стоп-слова (шумовые слова) – токены, которые часто встречаются в тексте, но для модели машинного обучения являются шумом. Как правило, это частицы, союзы, предлоги и подобные части речи. Удаление стоп-слов помогает удалить шумовую составляющую, а также сократить количество обрабатываемых токенов [5]. Все шумовые слова были удалены из используемого набора данных.

Для векторизации документов была использована модель векторизации ***FastText***, предобученная на русскоязычном корпусе GeoWAC. Преимуществами моделей векторизации перед другими методами векторизации является то, что они, во-первых, учитывают семантические, то есть смысловые отношения токенов; во-вторых, генерируют векторные представления размерности гораздо меньшей, чем размерность словаря [3]. В ходе векторизации в данной работе каждому токену было поставлено в соответствие векторное представление размерности 300. Векторное представление для документа вычислялось как среднее арифметическое векторных представлений всех токенов, входящих в данный документ.

Подготовленные данные были разделены на обучающую и тестовую выборки, причем на тестовую выборку приходится около 25 % всех документов корпуса.

Построение моделей машинного обучения

Тематическая классификация с точки зрения машинного обучения по прецедентам является задачей многоклассовой классификации. Каждому объекту x (в данном случае объектом является текстовый документ) ставится в соответствие некоторая метка $f(x)$, определяемая моделью классификации и обозначающая тематическую принадлежность данного текстового документа к конкретному классу – рубрике.

Для решения задачи тематической классификации с использованием языка программирования Python были обучены четыре модели классификации: ***наивный байесовский классификатор*** (далее – НБК), ***случайный лес решающих деревьев*** (далее – СЛРД), ***логистическая регрессия*** (далее – ЛР) и ***искусственная нейронная сеть*** (далее – ИНН), которая была реализована с использованием библиотеки Keras. Определение значений гиперпараметров для моделей классификации велось методом решетчатого поиска с использованием скользящего контроля (кросс-валидации) [11].

Разработанная искусственная нейронная сеть содержит четыре слоя (см. Рисунок 1). Входной слой является полносвязным и реализован с использованием встроенного класса Dense библиотеки Keras. Он содержит 300 нейронов, каждый из них получает на вход соответствующую координату 300-мерного вектора, представляющего отдельный текстовый документ. Первый скрытый слой реализован с использованием встроенного класса Dropout. Принцип работы данного слоя заключается в следующем. При работе с каждым из объектов обучающей выборки некоторые нейроны этого слоя случайным образом от-

ключаются [12]. Доля отключенных нейронов определяется коэффициентом исключения. Такой процесс снижает вероятность переобучения. Второй скрытый слой разработанной нейронной сети – это слой Batch Normalization. Он предназначен для статистической нормализации значений, полученных на выходах предыдущих слоев [8]. Выходной слой также полносвязный (Dense) и содержит 9 нейронов, каждый из которых соответствует определенному классу – рубрике.

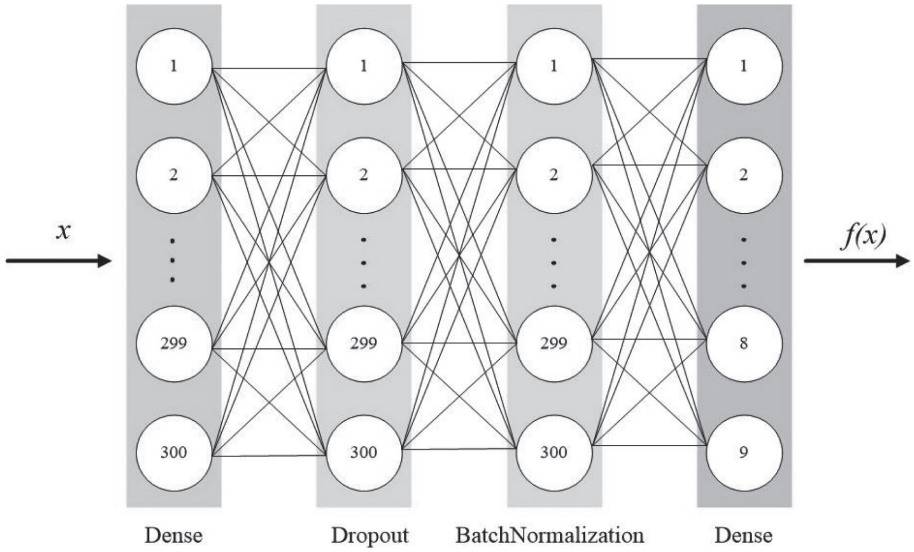

Рисунок 1. Структура искусственной нейронной сети

Для количественной оценки качества классификации были использованы следующие метрики: precision, recall и F1-мера. Метрики precision, recall определяются формулами

$$P = \frac{TP}{TP + FP}; \tag{1}$$

$$R = \frac{TP}{TP + FN}, \tag{2}$$

где *TP* – количество объектов, которые были правильно классифицированы как относящиеся к данному классу; *TN* – количество объектов, которые были правильно классифицированы как не относящиеся к данному классу; *FP* – количество объектов, которые были ошибочно классифицированы как относящиеся к данному классу; *FN* – количество элементов, которые были ошибочно классифицированы как не относящиеся к данному классу.

В качестве комбинированной метрики в данной работе была использована F1-мера, которая является частным случаем F-меры, определяемой в соответствии с формулой при $\beta = 1$ [16]:

$$F_\beta = (1 + \beta^2)\frac{P \cdot R}{\beta^2 \cdot P + R}, \tag{3}$$

где β имеет смысл веса метрики precision.

Результаты оценки качества построенных классификаторов представлены в Таблице 2 и на Рисунке 2.

Таблица 2

Значения метрик классификации

Классификатор	Среднее значение метрики precision	Среднее значение метрики recall	Среднее значение F1-меры
НБК	0,81459	0,79775	0,75367
ЛР	0,90216	0,90236	0,90222
СЛРД	0,88318	0,88310	0,88221
ИНС	0,9253	0,9250	0,9251

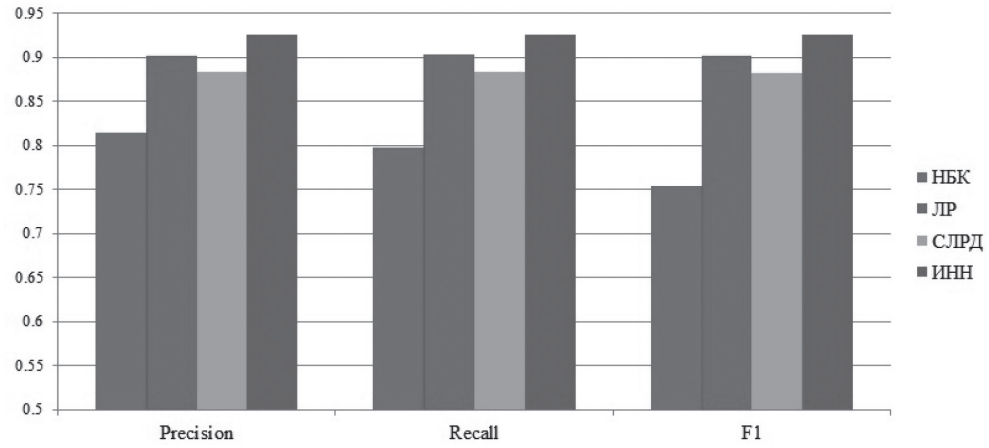

Рисунок 2. Гистограмма значений метрик классификации

Как видно, наилучшее качество классификации показала искусственная нейронная сеть, второй – логистическая регрессия.

Разработка веб-приложения

Для предоставления удобного пользовательского интерфейса и обеспечения доступа широкого круга лиц к разработанной системе было создано веб-приложение. Разработка веб-приложения велась с использованием фреймворка Django языка программирования Python [4]. В качестве системы управления базами данных использовалась MySQL.

Веб-приложение содержит набор веб-страниц, посвященных отдельным рубрикам, а также заглавную страницу и страницу с информацией о проекте. Веб-страница отдельной рубрики включает в себя набор блоков, каждый из которых содержит информацию отдельной новостной статьи: дату публикации, заголовок и начальную часть текста. Пример типовой веб-страницы разработанного веб-приложения представлен на Рисунке 3. При нажатии на содержимое такого блока осуществляется перенаправление на веб-страницу оригинальной новостной статьи на веб-сайте новостного ресурса.

В качестве модели классификации, используемой в веб-приложении, применяется разработанная модель классификации на основе искусственной нейронной сети. Она импортируется в веб-приложение в виде файла формата json, а ее веса – в файле формата h5.

Рисунок 3. Интерфейс веб-приложения

Заключение

Системы автоматического определения тематической принадлежности текстов, в том числе новостных, могут существенно упростить поиск информации и помочь пользователю успешно ориентироваться в информационном пространстве.

В ходе работы были построены и обучены следующие модели классификации: наивный байесовский классификатор, случайный лес решающих деревьев, логистическая регрессия и искусственная нейронная сеть. Разработанные модели были оценены по ряду метрик классификации. Наивысшие показатели качества были продемонстрированы искусственной нейронной сетью.

Для обеспечения конечному пользователю возможности удобным образом взаимодействовать с системой тематической классификации новостных текстов было разработано веб-приложение с использованием фреймворка Django и системы управления базами данных MySQL.

Литература

1. *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., 2020. С. 77–79.

2. *Гусев П.Ю.* Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 1. С. 39–47. DOI: 10.25205/1818-7900-2021-19-1-39-47.

3. *Жеребцова Ю.А., Чижик А.В.* Сравнение моделей векторного представления текстов в задаче создания чатбота // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18, № 3. С. 16–34. DOI: 10.25205/1818-7935-2020-18-3-16-34.

4. *Жилин В.А.* Сравнение веб-фреймворков Django и Rubyonrails // Россия молодая: сборник материалов VII Всероссийской научно-практической конференции молодых ученых с международным участием, Кемерово, 21–24 апреля 2015 года. Кемерово: Кузбасский государственный технический университет им. Т.Ф. Горбачева, 2015. С. 157.

5. *Мартынов В.А., Плотникова Н.П.* Нормализация и фильтрация текста для задачи кластеризации // XLVIII Огарёвские чтения: материалы научной конференции, Саранск, 06–13 декабря 2019 года. В 3 частях / Саранск: Национальный исследовательский мордовский государственный университет им. Н.П. Огарёва, 2020. С. 448–452.
6. *Ткаченко А.Л.* Решение задачи классификации документов вуза на основе методов интеллектуального анализа // Вестник кибернетики. 2021. № 1 (41). С. 12–19. DOI: 10.34822/1999-7604-2021-1-12-19.
7. *Челышев Э.А., Оцоков Ш.А., Раскатова М.В.* Автоматическая рубрикация текстов с использованием алгоритмов машинного обучения // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ, управление. 2021. № 4. С. 175–182. DOI: 10.25586/RNU.V9187.21.04.P.175. ISSN: 2414-9187.
8. *Челышев Э.А., Оцоков Ш.А., Раскатова М.В., Щёголев П.* Сравнение методов классификации русскоязычных новостных текстов с использованием алгоритмов машинного обучения // Вестник кибернетики. 2022. № 1 (45). С. 63–71. DOI: 10.34822/1999-7604-2022-1-63-71. EDNVHTYBB.
9. *Якиль К.А., Рязанова Н.Ю.* Фильтрация SMS-спама // Автоматизация. Современные технологии. 2016. № 9. С. 19–24.
10. *Яцко В.А.* Алгоритмы и программы автоматической обработки текста // Вестник Иркутского государственного лингвистического университета. 2012. № 1 (17). С. 150–161.
11. *Aggarwal C.C., Zhai C.* (2012) A Survey of Text Classification Algorithms. Mining Text Data.
12. *Chelyshev E.A., Raskatova M.V.* (2022) Information System for Automatic News Text Classification: Proc. 6th International Conference on Information Technologies in Engineering Education, Inforino 2022. Moscow, 12–15 April 2022. DOI: 10.1109/Inforino53888.2022.9782937. EDN UUZQPI.
13. *Hartmann J., Huppertz J., Schamp C., Heitmann M.* (2019) Comparing automated text classification methods. International Journal of Research in Marketing, vol. 36, pp. 20–38.
14. *Korobov M.* (2015) Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, pp. 320–332.
15. *Manning C.D., Raghavan P., Schütze H.* (2008) Introduction to Information Retrieval. Cambridge University Press.
16. *Vujovic Z.* (2021) Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications, vol. 12, pp. 599–606.
17. *Vychegzhanin S., Kotelnikov E., Milov V.* (2021) Comparative analysis of machine learning methods for news categorization in Russian. CEUR Workshop Proceedings, vol. 2922, pp. 100–108.

References

1. Vershinin E.V., Timchenko D.K. (2020) *Issledovanie primeneniya stemming i lemmatizacii pri razrabotke system adaptivnogo perevoda teksta. Nauka. Issledovaniya. Praktika: sb. izb. st. po materialam mezhdunar. nauch. konf.* [Study of the use of stemming and lemmatization in the development of systems for adaptive text translation // Nauka. Research. Practice: Sat. izb. Art. according to the materials of the international scientific conf.] SPb, pp. 77–79 (in Russian).
2. Gusev P.Yu. (2021) *Razrabotka sistemy klassifikacii tekstov po nauchnym special'nostyam s primeneniem metodov mashinnogo obucheniya* [Development of a text classification system for scientific specialties using machine learning methods]. *Vestnik NGU. Seriya: Informacionnye tekhnologii*, vol. 19, No. 1, pp. 39–47. DOI: 10.25205/1818-7900-2021-19-1-39-47 (in Russian).
3. Zherebcova Yu.A., Chzhik A.V. (2020) *Sravnenie modelej vektornogo predstavleniya tekstov v zadache sozdaniya chatbota* [Comparison of models of vector representation of texts in the task of creating a chat-

bot]. *Vestnik NGU. Seriya: Lingvistika i mezhkul'turnaya kommunikaciya*, vol. 18, No. 3, pp. 16–34. DOI: 10.25205/1818-7935-2020-18-3-16-34 (in Russian).

4. Zhilin V.A. (2015) *Sravnenieveb-frejmvorkov Django i Rubyonrails* [Comparison of Django and Rubionrails web frameworks]. *Rossiya molodaya: Sbornik materialov VII Vserossijskoj nauchno-prakticheskoj konferencii molodyh uchenyh s mezhdunarodnym uchastiem, Kemerovo, 21–24 aprelya 2015 goda* [Proc. of the VII All-Russian scientific and practical conference of young scientists with international participation, Kemerovo, April 21–24, 2015]. Kemerovo, Kuzbasskij gosudarstvennyj tekhnicheskij universitet im. T.F. Gorbacheva, p. 157 (in Russian).

5. Martynov V.A., Plotnikova N.P. (2020) *Normalizaciya i fil'traciya teksta dlya zadach iklasterizacii* [Text normalization and filtering for clustering tasks]. *XLVIII Ogaryovskie chteniya: materialy nauchnoj konferencii, Saransk, 06–13 dekabrya 2019 goda* [HLVIIII Ogaryov readings: materials of the scientific conference, Saransk, December 06–13, 2019]. Saransk, Nacional'nyj issledovatel'skij mordovskij gosudarstvennyj universitet im. N.P. Ogaryova, pp. 448–452 (in Russian).

6. Tkachenko A.L. (2021) Reshenie zadachi klassifikacii dokumentov vuza na osnove metodov intellektual'nogo analiza [Solution of the problem of classification of university documents based on the methods of intellectual analysis]. *Vestnik kibernetiki*, No. 1 (41), pp. 12–19. DOI: 10.34822/1999-7604-2021-1-12-19 (in Russian).

7. Chelyshev E.A., Ocokov Sh.A., Raskatova M.V. (2021) *Avtomaticheskaya rubrikaciya tekstov s ispol'zovaniem algoritmov mashinnogo obucheniya* [Automatic categorization of texts using machine learning algorithms]. *Vestnik Rossijskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz, upravlenie*, No. 4, pp. 175–182. DOI: 10.25586/RNU.V9187.21.04.P.175 (in Russian).

8. Chelyshev E.A., Ocokov Sh.A., Raskatova M.V., Shchyogolev P. (2022) Sravnenie metodov klassifikacii russkoyazychnyh novostnyh tekstov s ispol'zovaniem algoritmov mashinnogo obucheniya [Comparison of classification methods for Russian-language news texts using machine learning algorithms]. *Vestnik kibernetiki*, No. 1 (45), pp. 63–71. DOI: 10.34822/1999-7604-2022-1-63-71. EDN VHTYBB (in Russian).

9. Yakil' K.A., Ryazanova N.Yu (2016) *Fil'traciya SMS-spama* [Fil'traciya SMS spama]. *Avtomatizaciya. Sovremennye tekhnologii*, No. 9, pp. 19–24 (in Russian).

10. Yako V.A. (2012) *Algoritmy i programmy avtomaticheskoj obrabotki teksta* [Algorithms and programs for automatic text processing]. *Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta*, No. 1 (17), pp. 150–161 (in Russian).

11. Aggarwal C.C., Zhai C. (2012) A Survey of Text Classification Algorithms. Mining Text Data.

12. Chelyshev E.A., Raskatova M.V. (2022) Information System for Automatic News Text Classification: Proc. 6th International Conference on Information Technologies in Engineering Education, Inforino 2022. Moscow, 12–15 April 2022. DOI: 10.1109/Inforino53888.2022.9782937. EDN UUZQPI.

13. Hartmann J., Huppertz J., Schamp C., Heitmann M. (2019) Comparing automated text classification methods. International Journal of Research in Marketing, vol. 36, pp. 20–38.

14. Korobov M. (2015) Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, pp. 320–332.

15. Manning C.D., Raghavan P., Schütze H. (2008) Introduction to Information Retrieval. Cambridge University Press.

16. Vujovic Z. (2021) Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications, vol. 12, pp. 599–606.

17. Vychegzhanin S., Kotelnikov E., Milov V. (2021) Comparative analysis of machine learning methods for news categorization in Russian. CEUR Workshop Proceedings, vol. 2922, pp. 100–108.