

# ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT



УДК 004.89, 004.032.26

<https://doi.org/10.23947/2687-1653-2020-20-4-430-436>

## Автоматизация распределения информации в адаптивных системах электронного документооборота с применением машинного обучения



А. Д. Обухов

ФГБОУ ВО «Тамбовский государственный технический университет» (г. Тамбов, Российская Федерация)

**Введение.** Системы электронного документооборота (СЭД) используются для хранения, обработки и передачи больших объемов информации. Автоматизация данных процессов является актуальной проблемой, требующей комплексного решения. Ее решение позволит снизить временные и материальные затраты на проектирование и осуществить переход к более совершенным, адаптивным СЭД. Статья посвящена разработке новых методов автоматизации процесса распределения информации в СЭД.

Целью работы является повышение точности распределения информации в СЭД за счет перехода от аналитических или алгоритмических решений к использованию новых методов, функционирующих на основе технологий машинного обучения. Использование нейронных сетей для достижения поставленной цели также позволит повысить эффективность разработки программного обеспечения за счет автоматизации процессов анализа и обработки информации.

**Материалы и методы.** Предложен новый метод автоматизированного распределения информации на основе технологий машинного обучения, включающий математическое описание правил распределения информации. Сформулированный перечень условий распределения информации позволяет реализовать программное обеспечение на основе нейронных сетей для решения задачи автоматического распределения данных в СЭД.

**Результаты исследования.** Метод автоматизированного распределения информации апробирован на примере предметной области СЭД при решении задачи анализа правильности информации, введенной пользователем. В ходе экспериментальных исследований получено, что предложенный метод, функционирующий на основе технологий машинного обучения, обеспечивает лучшую точность (на 8 % выше) и более эффективен (в соответствии с метриками Джилба и цикломатической сложности)

**Обсуждение и заключения.** Полученные результаты подтверждают эффективность и точность предложенного метода. Представленные результаты могут использоваться для автоматизации процессов распределения и проверки информации в адаптивных СЭД, а также других информационных системах. На основе разработанного метода также возможно решение связанных задач: поиск дубликатов и похожих документов, классификация и размещение по категориям файлов.

**Ключевые слова:** системы электронного документооборота, распределение информации, автоматизация, адаптивность, машинное обучение.

**Для цитирования:** Обухов, А. Д. Автоматизация распределения информации в адаптивных системах электронного документооборота с применением машинного обучения / А. Д. Обухов // Advanced Engineering Research. — 2020. — Т. 20, № 4. — С. 430–436. <https://doi.org/10.23947/2687-1653-2020-20-4-430-436>

**Финансирование:** работа выполнена по гранту Президента РФ МК-74.2020.9.

© Обухов А. Д., 2020



## Automation of information distribution in adaptive electronic document management systems using machine learning

A. D. Obukhov

Tambov State Technical University (Tambov, Russian Federation)

**Introduction.** Electronic document management systems (EDMS) are used to store, process and transmit large amounts of information. Automation of these processes is a challenge that requires a comprehensive solution. Its solution will

reduce the time and material costs for design and make the transition to a more advanced, adaptive EDMS. The paper is devoted to the development of new methods for automating the process of distributing information in the EDMS. The work objective is to improve the accuracy of the information distribution in the EDMS through moving from analytical or algorithmic solutions to the use of new methods based on machine learning technologies. The application of neural networks in the furtherance of this purpose will also improve the efficiency of software development through automating the analysis and processing of information.

*Materials and Methods.* A new method of the automated information distribution based on machine learning technologies including a mathematical description of the information distribution rules is proposed. The formulated list of conditions for the information distribution provides the implementation of software based on neural networks for solving the problem of automatic data distribution in the EDMS.

*Results.* The method of automated information distribution has been tested on the example of the EDMS subject area when solving the problem of analyzing the correctness of information entered by the user. In the course of experimental studies, it was found that the proposed method, based on machine learning technologies, provides better accuracy (8 % higher) and is more efficient (in accordance with the Jilb metrics and cyclomatic complexity).

*Discussion and Conclusions.* The results obtained confirm the efficiency and accuracy of the method proposed. The presented results can be used to automate the processes of distribution and verification of information in adaptive EDMS, as well as in other information systems. Based on the method developed, it is also possible to solve connected problems: search for duplicates and similar documents, classification and placement by file categories.

**Keywords:** electronic document management systems, information distribution, automation, adaptability, machine learning.

**For citation:** A. D. Obukhov. Automation of information distribution in adaptive electronic document management systems using machine learning. *Advanced Engineering Research*, 2020, vol. 20, no. 4, p. 430–436. <https://doi.org/10.23947/2687-1653-2020-20-4-430-436>

**Funding information:** the research is done under grant no. MK-74.2020.9 from President of the Russian Federation.

**Введение.** Системы электронного документооборота (СЭД) широко используются для хранения, обработки и передачи информации различного типа: документов, электронных таблиц, графической и аудиоинформации, технической и бухгалтерской документации и так далее. Основная функциональность по работе с документами за годы развития СЭД сформирована. Дальнейшее развитие направлено на автоматизацию человеческой деятельности, повышение гибкости и надежности, реализацию интеллектуальных модулей по поддержке принятия решений [1–3].

Однако подобная гибкость и адаптивность СЭД приводит к дополнительным временным и материальным затратам, увеличивает сложность проектирования и модернизации системы. Поэтому актуальным вопросом при разработке адаптивных информационных систем, в том числе СЭД, является автоматизация проектирования системы, что включает множество отдельных задач по автоматизации процессов анализа, обработки и передачи документов, решение которых в комплексе позволяет снизить нагрузку на разработчиков в ходе реализации функций адаптивности в СЭД.

В рамках исследования будет рассмотрен вопрос классификации информации и ее последующего автоматического распределения по заданным категориям в СЭД. Данная подзадача является одной из наиболее распространенных при организации электронных хранилищ и архивов, картотек, заполнении форм документов. Под распределением понимается размещение данных и файлов на заданных позициях: по категориям, носителям, каталогам и так далее [4]. Классические алгоритмы не всегда способны определить ошибку в размещении информации без привлечения эксперта или модератора.

Рассматривая подходы к решению задачи распределения информации, можно выделить подходы, основанные на применении машинного обучения<sup>1</sup>. Использование технологий искусственного интеллекта для решения задач классификации доказало свою эффективность в ходе многочисленных исследований и экспериментов [5–8]. При решении поставленной задачи и реализации метода распределения информации для автоматизации процесса классификации данных будут использованы технологии машинного обучения.

На основе анализа процесса движения информации можно выделить следующие актуальные задачи по классификации и распределению данных в адаптивных СЭД:

— классификация и определение соответствия введенных данных той категории, в которой их расположил пользователь [9];

<sup>1</sup> Umadevi S. A survey on data mining classification algorithms / S. Umadevi, K. S. J. Marseline // 2017 International Conference on Signal Processing and Communication (ICSPPC): IEEE, 2017. P. 264–268.

- классификация данных по категориям<sup>2</sup> и, в случае ошибки, перемещение их в правильные категории либо выдача пользователю предупреждения [10];
- определение дублирования данных по признакам [11].

Для решения поставленных задач предлагается разработать метод автоматизированного распределения информации в адаптивных СЭД, обобщающий существующие подходы к классификации информации и основанный на применении методов машинного обучения для автоматизации процессов обработки данных.

**Материалы и методы.** Формализуем основные этапы метода автоматизированного распределения данных в адаптивных СЭД.

В основе метода — формирование и обучение нейронной сети для классификации информации. Поэтому на первом этапе реализации метода необходимо подготовить множество информационных объектов  $X = \{x_1, \dots, x_N\}$  для обучения и тестирования нейронной сети. Объект  $x_i \in X$  может быть представлен текстовой или численной информацией, введенной пользователем в поля форм, либо файлами, загруженными через интерфейс СЭД [12]. Процесс подготовки данных может включать нормализацию, токенизацию, лемматизацию, извлечение свойств и атрибутов файлов. Для собранных данных задается множество категорий  $Y = \{y_1, \dots, y_M\}$ , то есть существует непрерывное отображение  $X \rightarrow Y$ . Аппроксимируем это отображение нейронной сетью  $NN$ :

$$NN(X) = Y.$$

При достаточном объеме тренировочных данных можно обеспечить необходимую точность классификации. В соответствии с исследованиями, представленными в работе [13], на каждый выходной признак необходимо предоставить не менее 50 экземпляров тренировочных данных. Тогда для  $\forall x_i \in X$  получим  $NN(x_i) = y_j$ , где  $j = 1..M$ .

На второй этапе метода осуществляется распределение информации. Пусть в СЭД задан набор информационных объектов  $X = \{x_1, \dots, x_N\}$  и множество соответствующих им категорий  $Y = \{y_1, \dots, y_M\}$ . Для распределения информации необходимо осуществить проверку следующих условий:

- если  $\forall x_i \in X$ ,  $NN(x_i) = y_j$  и объект  $x_i$  помещен пользователем в категорию  $y_j$ , тогда положение  $x_i$  остается без изменений;
- если  $\exists x_i \in X$ ,  $NN(x_i) = y_m$  (где  $y_m$  — категория вредоносных объектов), то объект  $x_i$  удаляется из СЭД, администратору системы отправляется предупреждение, пользователь, добавивший объект  $x_i$ , блокируется;
- если  $\exists x_i \in X$ ,  $NN(x_i) = y_j$ , но объект  $x_i$  введен пользователем в категорию  $y_k$  ( $y_k \neq y_j$ ), то необходимо осуществить перераспределение информации в СЭД. Рассмотрим все возможные варианты перераспределения:

1. Категория  $y_j$  свободная (пустая), тогда объект  $x_i$  переносится из категории  $y_k$  в  $y_j$ .
2. Категория  $y_j$  занята некоторым объектом  $x_q$  (таким, что  $NN(x_q) = y_j$ ), тогда объект  $x_i$  переносится в буфер, пользователю выдается предупреждение о неправильном вводе  $x_i$  и дублировании информации.
3. Категория  $y_j$  занята некоторым объектом  $x_q$  (таким, что  $NN(x_q) \neq y_j$ ), тогда объект  $x_q$  переносится в буфер, а объект  $x_i$  переносится из категории  $y_k$  в  $y_j$ , пользователю выдается предупреждение о неправильном вводе  $x_q$ .

Рассмотренные в рамках метода этапы и условия распределения информации включают основные сценарии добавления информации в СЭД. За счет применения технологий машинного обучения возможно автоматизировать распределение информации по заданным категориям.

**Результаты исследования.** Для реализации и интеграции метода в адаптивную СЭД будем использовать микросервисный подход. Нейронная сеть реализуется посредством библиотеки Keras (Python), после этого импортируется в микросервис, который может быть реализован на основе любого фреймворка, например, Flask. Данные между СЭД и микросервисом передаются по протоколу HTTP в формате JSON, что обеспечивает совместимость с любой реализацией СЭД [14].

В качестве примера задачи распределения данных будем использовать форму карточки документа в СЭД с 5 полями: наименование документа; автор документа; контактная информация (адрес) автора; дата создания; описание документа. В ходе экспериментального исследования будет моделироваться ввод данных пользователей в форму и заполнение информации в указанные 5 полей. Нейронная сеть будет классифицировать введенную информацию и проверять ее на соответствие заданной категории.

Для формирования набора исходных данных используем генератор на основе открытых баз русских имен, городов и стран, а также последовательностей из русских слов и цифр. Это позволит получить конструкции, соответствующие реальным данным, но не хранящие персональные данные реальных людей. Текстовые данные будут обработаны с помощью токенизатора для перевода их в численный формат. Максимальная длина предложения ограничена 20 словами. Сгенерируем тренировочную выборку. Для каждой категории сформируем по 10 000 элементов для каждой категории. Также добавим 50 000 элементов с неправильными и ошибочными данными. Таким образом получим массив из 100 тысяч элементов.

Зависимости между входными признаками и категориями представлены на тепловой карте на рис. 1. Карта позволяет визуализировать матрицу корреляций и провести наглядный анализ данных для определения того, как параметры влияют друг на друга и на выходные переменные [15].

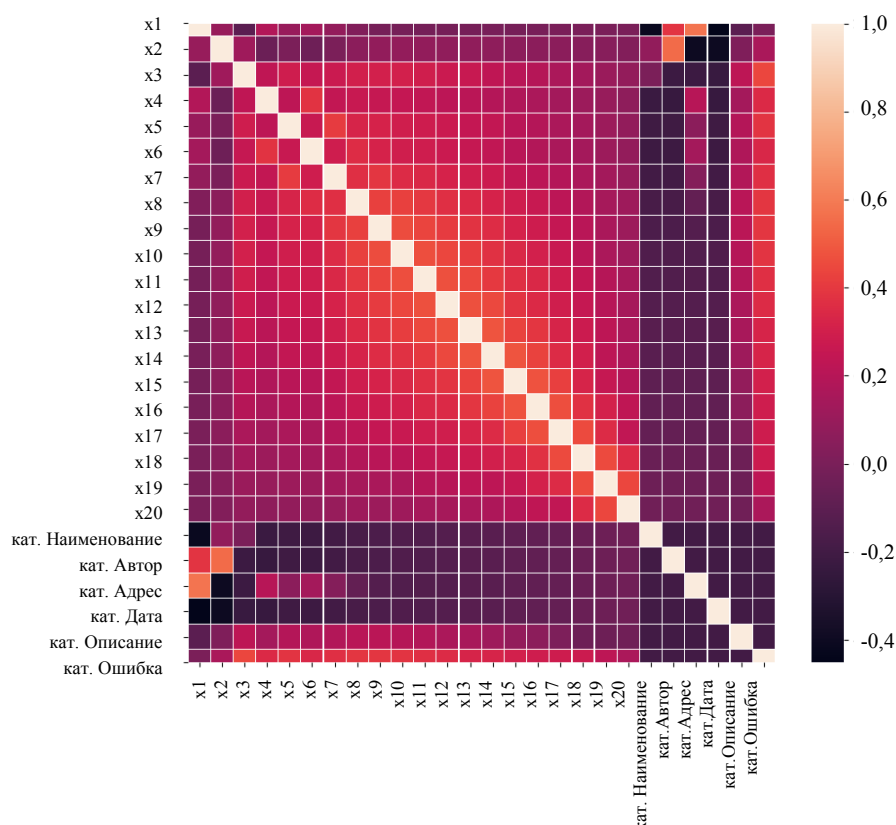


Рис. 1. Тепловая карта исходных данных

Процесс обучения сети отражен на рис. 2. Итоговая точность нейронной сети на контрольном наборе после 5 эпох составила 97,8 %.

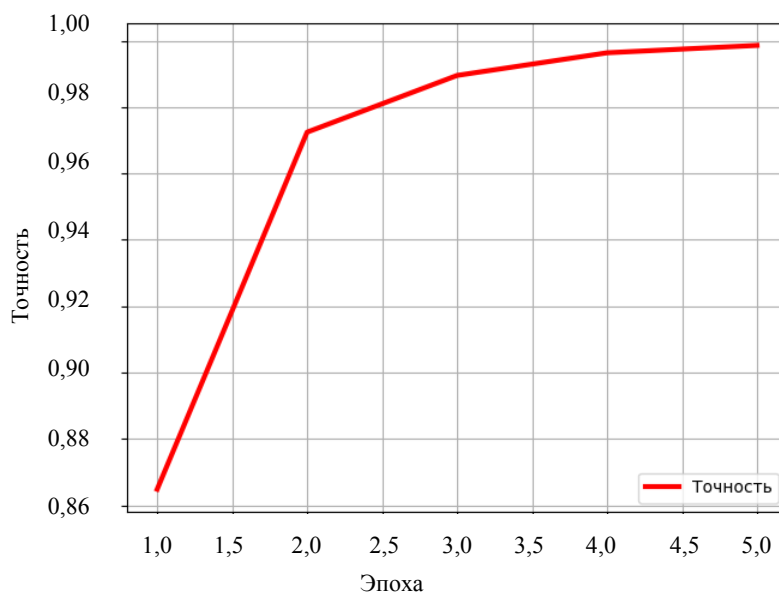


Рис. 2. Процесс обучения нейронной сети классификации

Количество ошибок не превысило 7 % для категории «Автор» и 1 % для категории «Описание». В остальных категориях некорректные данные были распознаны в 100 % случаев.

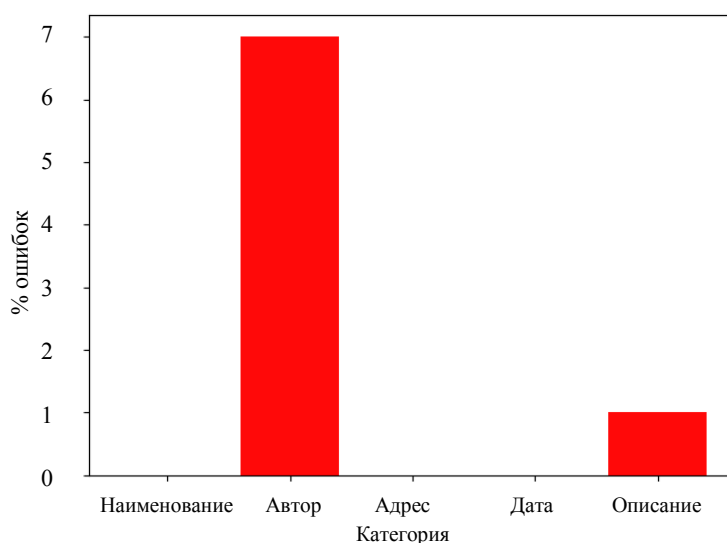


Рис. 3. Тестирование нейронной сети на ввод некорректных данных

Второй эксперимент (рис. 4) заключается в вводе данных из  $j$  категорий в  $i$  ( $j \neq i$ ). Формируется набор из 100 элементов для  $i$  каждой категории, состоящий из элементов других категорий (по 25 каждой категории). График отражает количество тестов, в которых нейронная сеть неправильно распознала данные других категорий, как соответствующие текущей категории. Во втором эксперименте был допущен 1 % в категории «Наименование», остальные категории обработаны без ошибок.

Далее сравним точность предложенного метода (далее обозначен как «нейросетевой метод») с классическим решением задачи классификации и распределения информации (обозначен как «классический метод»). Для классического метода получены следующие результаты: 16 % ошибок в категории «Наименование» и 25 % — в «Описание», что существенно хуже показателей нейросетевого метода (рис. 5). В среднем использование нейросетевого метода обеспечивает повышение точности на 8 %.

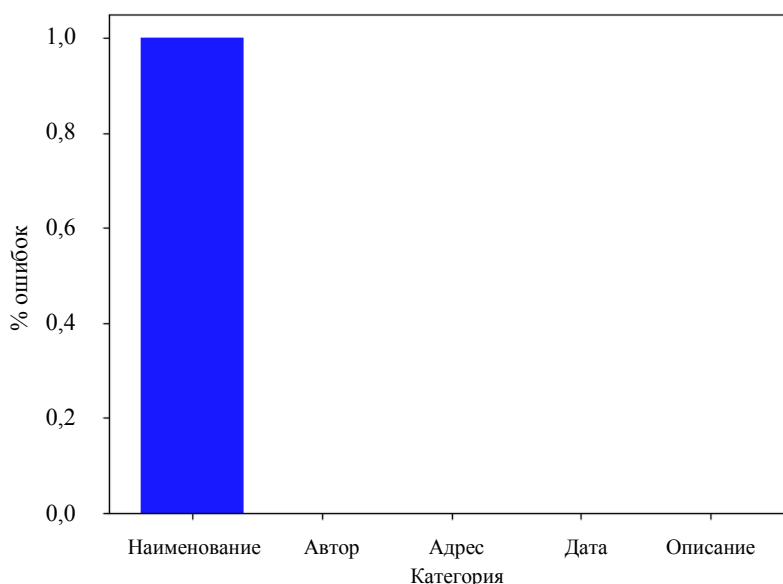


Рис. 4. Тестирование нейронной сети на соответствие категории

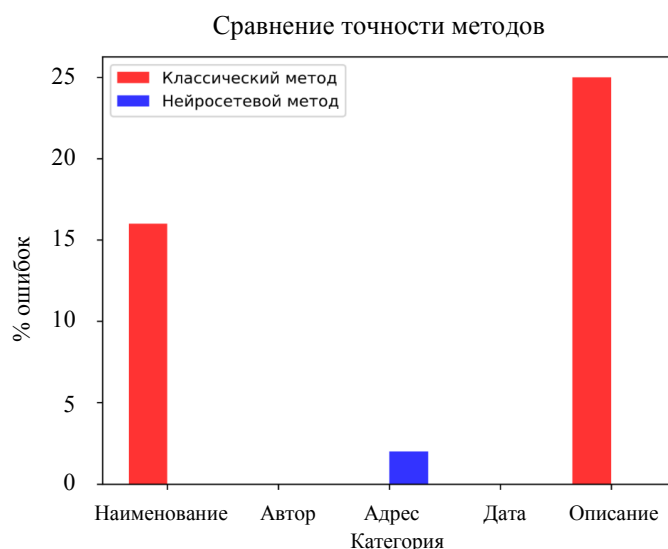


Рис. 5. Сравнение точности нейросетевого и классического методов

Далее осуществим сравнение методов по метрикам сложности реализации программного обеспечения. В качестве метрик используются цикломатическая сложность, выраженная числом McCabe (общая и усредненная) [16], а также сложность по метрике Джилба [17]. Итоговые результаты представлены в таблице 1.

Таблица 1

Сравнение классического и нейросетевого методов

| Метрика                             | Классический метод | Нейросетевой метод |
|-------------------------------------|--------------------|--------------------|
| Цикломатическая сложность (общая)   | 26                 | 22                 |
| Цикломатическая сложность (средняя) | A (3,71)           | A (1,37)           |
| Метрика Джилба                      | 0,8                | 0,15               |

Таким образом, сложность нейросетевого метода по метрикам цикломатической сложности и Джилба ниже. Точность разработанного метода выше почти на 8 %. Стоит заметить, что сложность реализации классического и нейросетевого метода классификации и распределения данных относительно невелики, так что их можно считать сравнимыми. Однако, полученные результаты по точности классификации подтверждают большую эффективность нейросетевого подхода.

**Обсуждение и заключения.** В работе поставлена задача автоматизации процессов классификации и распределения данных в адаптивных СЭД. Предложенный метод автоматизированного распределения данных включает формализацию процесса классификации и распределения данных, использование технологий машинного обучения для автоматизации решения задачи. Сформулированный перечень условий распределения информации позволяет реализовать программное обеспечение на основе нейронных сетей, решающее задачу автоматического распределения данных.

Для апробации разработанного метода проведены экспериментальные исследования на базе сгенерированных данных о документах в СЭД. Точность обученной нейронной сети составила около 98 %. Дополнительные испытания показали, что нейронная сеть может обнаруживать некорректно распределенные данные почти в 100 % случаев, причем, при наихудших условиях ошибка не превысила 7 %. Таким образом подтверждена эффективность и точность предлагаемого метода. Разработанный метод, по сравнению с классической реализацией на основе алгоритмического обеспечения, показывает следующий положительный эффект: повышение средней точности на 8 %, снижение сложности реализации.

Представленные результаты могут использоваться для автоматизации процессов распределения и проверки информации в адаптивных СЭД, а также других информационных системах. Кроме того, на основе предложенного метода возможно решение связанных задач: поиск дубликатов и похожих документов, классификация и размещение по категориям файлов.

#### Библиографический список

1. Кузнецова, Е. В. Актуальные проблемы электронного документооборота в органах власти / Е. В. Кузнецова // Вопросы управления. — 2013. — №. 4. — С. 73–77.

2. Zhong, R. Y. Intelligent manufacturing in the context of industry 4.0: a review / R. Y. Zhong [et al.] // Engineering. — 2017. — Vol. 3 (5). — P. 616–630. DOI: 10.1016/J.ENG.2017.05.015
3. Xu, D. Enhancing e-learning effectiveness using an intelligent agent-supported personalized virtual learning environment: An empirical investigation / D. Xu [et al.] // Information & Management. — 2014. — Vol. 51 (4). — P. 430–440. DOI:10.1016/j.im.2014.02.009
4. Кузнецов, С. Д. Распределенные горизонтально масштабируемые решения для управления данными / С. Д. Кузнецов, А. В. Посконин // Труды Института системного программирования РАН. — 2013. — Т. 24. — С. 327–358.
5. Сравнительный анализ методов машинного обучения для решения задачи классификации документов научно-образовательного учреждения / М. Н. Краснянский, А. Д. Обухов, Е. М. Соломатина, А. А. Воякина // Вестник ВГУ, Серия : Системный анализ и информационные технологии. — 2018. — № 3. — С. 173–182.
6. Karampidis, K. File type identification-computational intelligence for digital forensics / K. Karampidis, G. Papadourakis // Journal of Digital Forensics, Security and Law. — 2017. — Vol. 12 (2). — P. 6. DOI: 10.15394/jdfsl.2017.1472
7. Kim, D. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec / D. Kim [et al.] // Information Sciences. — 2019. — Vol. 477. — P. 15–29.
8. Zheng, J. Hierarchical neural representation for document classification / J. Zheng [et al.] // Cognitive Computation. — 2019. — Vol. 11 (2). — P. 317–327. DOI:10.1007/s12559-018-9621-6
9. Bodström, T. State of the art literature review on network anomaly detection with deep learning / T. Bodström, T. Hämmäläinen // In book: Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, Cham, 2018. — P. 64–76. DOI: 10.1007/978-3-030-01168-0\_7
10. Datta, S. Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs / S. Datta, S. Das // Neural Networks. — 2015. — Vol. 70. — P. 39–52. DOI: 10.1016/j.neunet.2015.06.005
11. Irolla, P. The duplication issue within the Drebin dataset / P. Irolla, A. Dey // Journal of Computer Virology and Hacking Techniques. — 2018. — Vol. 14 (3). — P. 245–249. DOI: 10.1007/s11416-018-0316-z
12. Goldberg, Y. Neural network methods for natural language processing / Y. Goldberg // Synthesis Lectures on Human Language Technologies. — 2017. — Vol. 10 (1). — P. 1–309. DOI: 10.2200/S00762ED1V01Y201703HLT037
13. Beleites, C. Sample size planning for classification models / C. Beleites [et al.] // Analytica chimica acta. — 2013. — Vol. 760. — P. 25–33. DOI:10.1016/j.aca.2012.11.007
14. Obukhov, A. Algorithm of adaptation of electronic document management system based on machine learning technology / A. Obukhov, M. Krasnyanskiy, M. Nikolyyukin // Progress in Artificial Intelligence. — 2020. — Vol. 9. — P. 287–303. DOI: 10.1007/s13748-020-00214-2
15. Bazgir, O. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks / O. Bazgir [et al.] // Nature Communications. — 2020. — Vol. 11 (1). — P. 1–13. DOI: 10.1038/s41467-020-18197-y
16. Luo, A. A Structural Complexity Metric Method for Complex Information Systems / A. Luo [et al.] // JSW. — 2019. — Vol. 14 (7). — P. 332–339. DOI: 10.17706/jsw.14.7.332-339
17. Смирнов, А. В. Методы оценки и управления качеством программного обеспечения / А. В. Смирнов // Известия СПбГЭТУ «ЛЭТИ». — 2019. — № 2. — С. 20–25.

Сдана в редакцию 29.10.2020

Запланирована в номер 27.11.2020

*Об авторах:*

**Обухов Артем Дмитриевич**, доцент кафедры «Системы автоматизированной поддержки принятия решений» ФГБОУ ВО «Тамбовский государственный технический университет, (392000, РФ, г. Тамбов, ул. Советская, 106), кандидат технических наук, доцент, ResearcherID: [M-9836-2019](https://orcid.org/0000-0002-3450-5213), ScopusID: [56104232400](https://orcid.org/0000-0002-3450-5213), ORCID: <http://orcid.org/0000-0002-3450-5213>, [Obukhov.art@gmail.com](mailto:Obukhov.art@gmail.com)

*Автор прочитал и одобрил окончательный вариант рукописи.*