УДК 004

ДВОИЧНАЯ КЛАССИФИКАЦИЯ АВИАЦИОННЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННОЙ СЕТИ

С.А. ПЕТРУХИН, Г.Е. ГЛУХОВ, Н.Н. ЛАДЫГИНА

Государственный научно-исследовательский институт гражданской авиации, г. Москва, Российская Федерация

Аннотация. Для автоматизации процессов обработки возрастающих объемов информации в Информационно-аналитическом центре, связанных с решением ряда важнейших задач в авиации, рассматривается возможность создания и практического использования нейронной сети. Описывается процесс построения работающей нейронной сети – двоичного классификатора официальных/неофициальных текстов, имеющих отношение к гражданской авиации. В качестве архитектуры нейронной сети применяется персептрон с одним скрытым слоем нейронов с логистической функцией активации. Обучение сети осуществляется методом градиентного спуска. Обосновывается определение количества нейронов для входного и выходного слоя. Подробно рассматривается процесс экспериментального подбора достаточного для качественной работы сети количества нейронов в скрытом слое. А также подбор коэффициента обучения нейронной сети. Приводятся данные о времени обучения нейронной сети в зависимости от количества эпох обучения и графики качества работы нейросети-классификатора на различных этапах построения. Проводятся сравнительные тесты качества работы нейросети и демонстрируются результаты таких тестов для нейронной сети на начальном и конечном этапах проектирования. Итогом работы будет являться построенная нейронная сеть, пригодная для практического применения в деятельности Информационно-аналитического центра ФГУП ГосНИИ ГА и для авиации.

Ключевые слова: авиационный текст, классификация текста, машинный анализ текстов, машинная обработка естественных языков, машинное обучение, нейронные сети, многослойный персептрон, градиентный спуск, логистическая функция

BINARY CLASSIFICATION OF AVIATION TEXTS USING A NEURAL NETWORK

S.A. PETRUKHIN, G.E. GLUKHOV, N.N. LADYGINA

The State Scientific Research Institute of Civil Aviation, Moscow, Russian Federation

Abstract. To automate the processing of increasing amounts of information in the Information and Analytical Center associated with the solution of a number of important tasks in aviation, the possibility of creating and practical use of a neural network is considered. The process of building a working neural network – a binary classifier of official / unofficial texts related to civil aviation is described. As the architecture of the neural network, a perceptron with one hidden layer of neurons with a logistic activation function is used. The network is trained using the gradient descent method. Determination of the number of neurons for the input and output layers is substantiated. The process of experimental selection of a sufficient number of neurons in the hidden layer for high-quality network operation is considered in detail. And also the selection of the neural network learning coefficient. The data on the training time of the neural network depending on the number of training epochs and graphs

of the quality of the neural network-classifier at various stages of construction are given. Comparative tests of the quality of the neural network are carried out and the results of such tests for the neural network at the initial and final design stages are demonstrated. The result of the work will be a built neural network suitable for practical use in the activities of the Information and Analytical Center of the FSUE GosNII GA and for aviation.

Keywords: aviation text, text classification, machine text analysis, machine processing of natural languages, machine learning, neural networks, multilayer perceptron, gradient descent, logistic function

Введение

В своей деятельности Информационно-аналитический центр ФГУП ГосНИИ ГА оперирует различными документами, текстами и сведениями, имеющими как прямое, так и косвенное отношение к гражданской авиации. Поиск, обработка и систематизация таких данных производится сотрудниками Центра в целях обеспечения деятельности следующих направлений:

– ведение (обслуживание) центральной нормативно-методической библиотеки гражданской авиации;

– мониторинг и оценка аутентичности компонентов воздушных судов;

– разработка и эксплуатация информационных систем поддержания летной годности;

– безопасность полетов и авиационная безопасность.

Учитывая возрастающий поток документов и текстов, которые необходимо обрабатывать сотрудникам, актуальной становится задача автоматизации поиска и обработки документов.

При обработке информации может производиться автоматический отбор текстов, имеющих отношение к авиации в соответствии с ранее разработанной методикой [1]. При этом отбираются как официальные, так и неофициальные (любительские) тексты. К первым относятся приказы, распоряжения, инструкции, пресс-релизы, эксплуатационная и техническая документация и т.п. Ко вторым – тексты, написанные блогерами, любителями авиационной истории, современной и/или перспективной авиатехники, пилотами-любителями, и обычно содержат личные впечатления или личное отношение к тем или иным событиям.

Для качественного автоматического анализа информации необходимо предварительно разделить официальные и неофициальные тексты, имеющие отношение к авиации, и именно о них будет идти речь в статье. Фактически это классическая задача машинной классификации текста [2]. В данном случае необходимо провести классификацию по признаку официальный/неофициальный текст. Решение задач двоичной классификации хорошо изучено применительно к нейронным сетям [3–5].

Проектирование основных параметров нейросети-классификатора

Для обучения нейронной сети было собрано несколько десятков официальных и неофициальных текстов. Кодирование текста в цифровую форму, пригодную для подачи на вход нейросети, возможно несколькими способами. В данном случае был выбран стемминг (нестрогое выделение основы слова) [6], с последующим подсчетом относительной доли каждой основы к общему количеству извлеченных из текста основ. За основу слов в данном случае брались начальные символы слов. Был проведен подсчет уникальных основ слов для некоторых текстов, имеющих отношение к авиации, при выделении разного количества начальных символов. Ниже приводятся результаты подсчета (табл. 1).

Фактически количество основ будет являться количеством нейронов входного слоя [7], поэтому, с целью получения приемлемого времени обучения нейронной сети, желательно

ограничить количество входных нейронов, исходя из этого были выбраны трехбуквенные основы.

Для классификации на два класса хорошо подходит персептрон [8] с одним скрытым слоем, где выход каждого нейрона входного и скрытого слоя подается на вход всех нейронов последующего слоя. В качестве функции активации используется логистическая функция, а в качестве метода обучения – хорошо себя зарекомендовавший метод градиентного спуска [9, 10].

Таблица 1

Количество уникальных основ для различного числа начальных символов

Основы	3-символьные	4-символьные	5-символьные
Количество	~ 2000	~ 7000	~ 16000
Примеры основ для слов: летчик, аэропорт, аэровокзал, бортпроводник	 лет аэр аэр бор	 летч аэро аэро борт	 летчи аэроп аэров бортп

Количество нейронов входного слоя равно количеству полученных основ, а именно – 1198. Нейронов в выходном слое должно быть два. Активность первого будет сигнализировать о том, что текст официальный, активность второго – неофициальный. Соответственно при обучении нейросети мы будем стремиться к тому, чтобы при подаче официальных текстов получить состояние выходных нейронов (1,0), при подаче любительских – (0,1). На рис. 1 представлена архитектура применяемой нейронной сети. Количество нейронов во входном и выходном слоях определено, количество нейронов в скрытом слое на начальном этапе построения сети неизвестно.

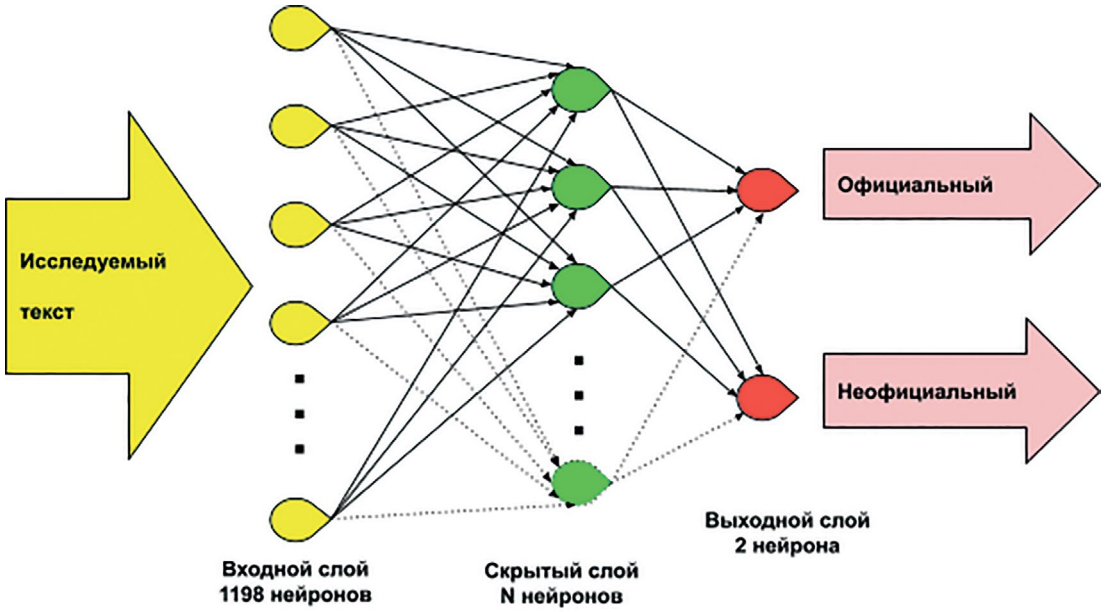

Рис. 1. Архитектура применяемой нейронной сети (персептрон)

Собранные для обучения тексты были переработаны в обучающие выборки. Каждая содержит 1198 относительных долей основ к общему количеству найденных в тексте основ и целевое значение выходных нейронов (1,0) или (0,1).

Для оценки количества нейронов в скрытом слое на начальном этапе, для более быстрого (хотя и грубого) обучения, был установлен большой коэффициент обучения, равный 10, и небольшое количество эпох обучения, равное 500 [11]. Для оценки работы обученной нейросети были сформированы 20 тестовых выборок (на основе десяти официальных и десяти неофициальных текстов).

Результаты подбора количества нейронов для скрытого слоя с текущими параметрами обучения представлены на рис. 2, где k – коэффициент обучения, c – количество эпох обучения и n – количество нейронов в скрытом слое, которое менялось от 3 до 12.

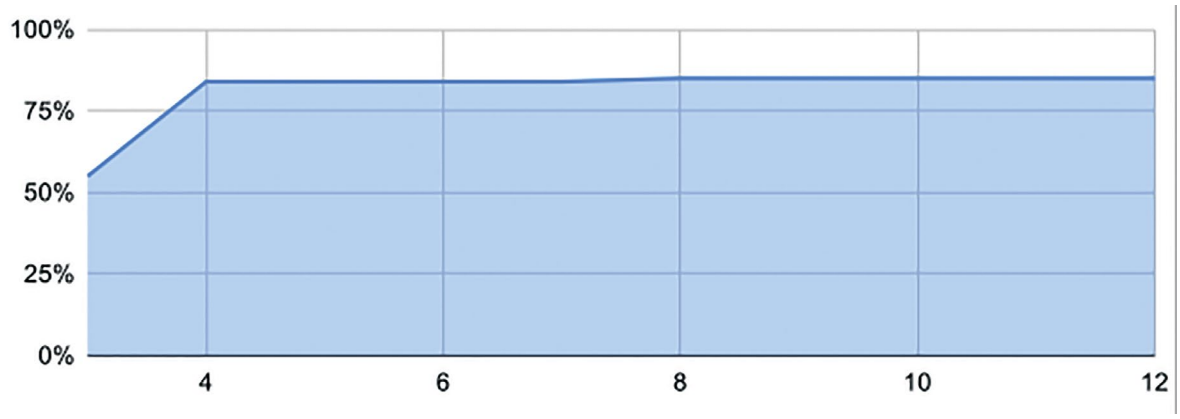

Рис. 2. График среднего приближения к правильному ответу (k=10, c=500, n=3..12)

При трех нейронах сеть еще не может различать виды текстов (правильных ответов немногим более 50 %), при четырех – происходит резкий скачок качества работы сети. При дальнейшем увеличении количества нейронов наблюдается плавное, очень медленное увеличение качества работы нейронной сети, и после шести и восьми нейронов происходит выход на «плато» и дальнейшее увеличение нейронов не дает видимого улучшения.

Возможно ли, что с большим количеством нейронов в скрытом слое произойдет кардинальное улучшение качества классификации? Для проверки этого предположения была предпринята попытка обучить нейросеть с большим количеством нейронов. Начиная с пяти нейронов в скрытом слое, на каждом шаге их количество увеличивалось на 5. Полученные результаты показаны на рис. 3.

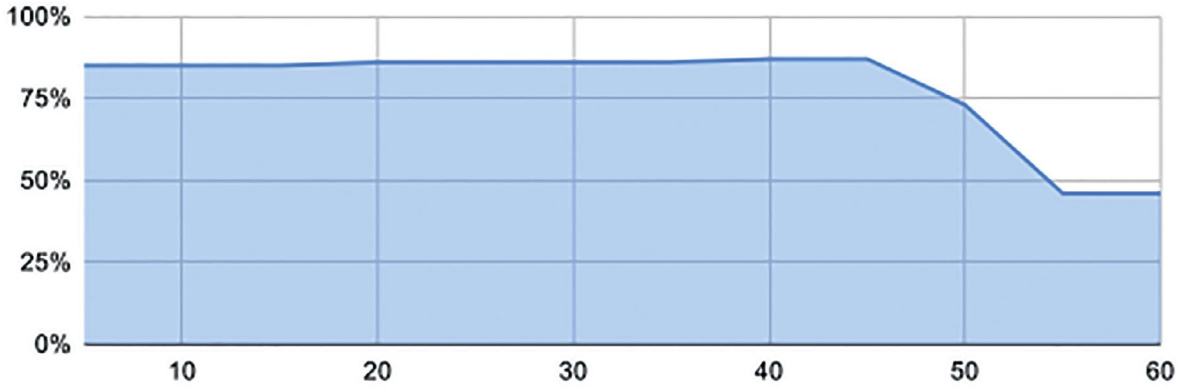

Рис. 3. График среднего приближения к правильному ответу (k=10, c=500, n=5..60)

Обнаружено, что и большее количество нейронов не дает существенного улучшения качества работы. Более того, видим, что при 50 нейронах наступает ухудшение классификации, а при 55 и 60 нейросеть полностью теряет способность к разделению текстов на классы. Очевидно, что для увеличившегося количества нейронных связей не хватает 500 эпох для обучения. При попытке увеличения эпох обучения до 1000 результат выглядит лучше (рис. 4), теперь спад качества распознавания более пологий, однако по-прежнему очевидно недообучение нейросети.

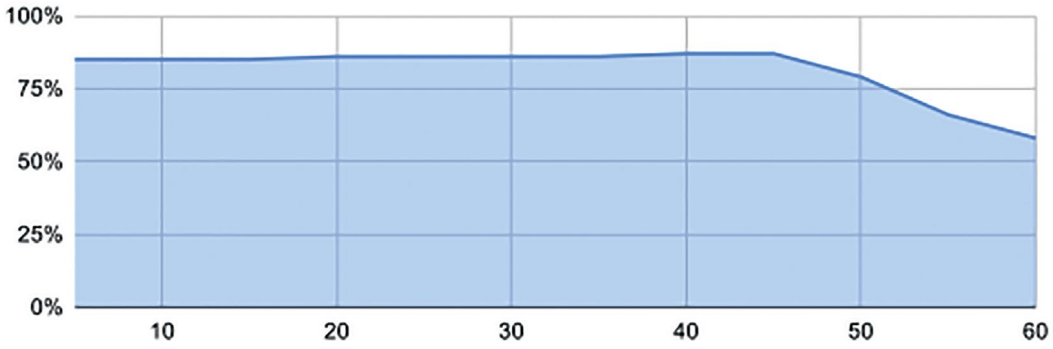

Рис. 4. График среднего приближения к правильному ответу (k=10, c=1000, n=5..60)

При значительном увеличении количества нейронов в скрытом слое, кривая обучения все-таки поднимается вверх (качество распознавания улучшается), хотя и очень медленно. Судя по графику, возможно улучшить работу сети, кардинально увеличив количество нейронов в скрытом слое и одновременно эпох обучения, чтобы нейросеть «успевала» обучиться. Это часто встречающаяся ситуация, когда для существенного улучшения качества распознавания количество нейронов должно быть увеличено зачастую на порядки. Однако такое увеличение приведет к сильному росту времени обучения, которое на имеющемся оборудовании (обычный пользовательский персональный компьютер), будет исчисляться сутками.

Приняв во внимание временные ограничения, а также то, что хотя качество обучения при увеличенном количестве нейронов улучшается, но 2 текста из 20 классифицируются неверно и при 4 нейронах в скрытом слое и при 40, принято решение ограничиться небольшим количеством нейронов в скрытом слое. На данном этапе было выяснено, что достаточное количество нейронов скрытого слоя равно 6–8, то есть нейросеть выделяет 6–8 параметров, существенно влияющих на классификацию. Возможно с более тонкими настройками обучения (коэффициентом обучения и количеству эпох), нейросеть обнаружит дополнительные параметры, влияющие на классификацию текстов, поэтому количество нейронов для скрытого слоя целесообразно установить с небольшим запасом, и итоговое количество было принято равным 10.

Создание рабочей нейросети-классификатора

Для построения итоговой нейронной сети был кардинально уменьшен коэффициент обучения до 0,1 и увеличено количество эпох обучения до 10 тысяч. При таких параметрах среднее приближение к правильному ответу составило около 77 %, тогда как ранее, при более грубых параметрах, оно составляло около 85 %, то есть, при сильно уменьшенном коэффициенте обучения, опять наблюдается недообучение сети. Поэтому, количество эпох обучения увеличено до 100 тыс. и одновременно начат подбор коэффициента обучения (рис. 5).

Теперь ситуация улучшилась, нейросеть опять достигла среднего приближения к правильному ответу около 85 % при коэффициенте обучения 0,2 и 0,3. При более низком и более высоких коэффициентах обучения результат оказывается хуже. По-видимому, это объясняется

тем, что при более низком коэффициенте нейросеть не достигает локального минимума функции ошибки, а при более высоком проскакивает его.

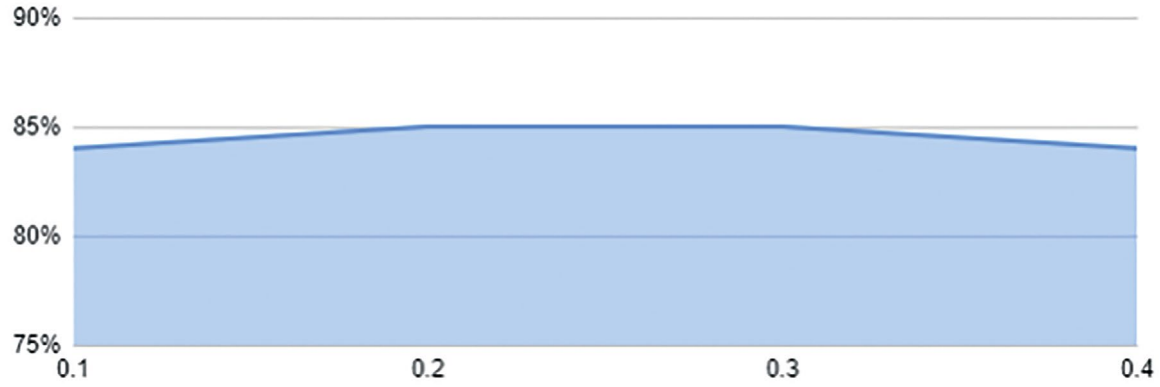

Рис. 5. График среднего приближения к правильному ответу (k=0.1..0.4, c=1000000, n=10)

Для достижения наилучшего результата установлен коэффициент обучения 0,2 и количество эпох обучения равное миллиону, при этом время обучения значительно увеличилось (табл. 2), и составило почти 1,5 часа при расчетах на обычном пользовательском компьютере (на одном ядре процессора Intel Core i3 – 6006U на частоте 2 Ггц, без применения аппаратных нейроускорителей, графических процессоров или специализированных облачных сервисов).

Таблица 2

Время обучения нейронной сети в зависимости от количества эпох

Количество эпох обучения	100	1 тысяча	10 тысяч	100 тысяч	1 миллион
Время обучения, сек	0,61	5,33	53,28	519,43	5149,82

Результаты тестовых классификаций начальной нейронной сети и итоговой приведены ниже (табл. 3), активным считается нейрон, значение которого достигло более 0,5 (выделенные ячейки в таблице). Тесты номеров 1–10 содержали официальные тексты, а тесты номеров 11–20 – неофициальные. Значения выходных нейронов округлены с точностью до сотых.

Таблица 3

Результаты тестов начальной и итоговой нейронных сетей

№ теста	Начальная нейросеть (k=10, n=4, c=500)		Итоговая нейросеть (k=0,2; n=10, c=1млн)	
	Нейрон 1 (официальный текст)	Нейрон 2 (неофициальный текст)	Нейрон 1 (официальный текст)	Нейрон 2 (неофициальный текст)
1	0,98	0,02	1,00	0,00
2	0,95	0,05	0,99	0,01
3	0,84	0,16	0,94	0,06
4	0,84	0,16	0,94	0,06

Продолжение таблицы 3

№ теста	Начальная нейросеть (k=10, n=4, c=500)		Итоговая нейросеть (k=0,2; n=10, c=1млн)	
	Нейрон 1 (официальный текст)	Нейрон 2 (неофициальный текст)	Нейрон 1 (официальный текст)	Нейрон 2 (неофициальный текст)
5	0,94	0.06	1,00	0,00
6	0,98	0,02	1,00	0,00
7	1,00	0,00	1,00	0,00
8	0,35	0,65	0,27	0,73
9	0,61	0,39	0,52	0,48
10	0,75	0,25	0,82	0,18
11	0,01	0.99	0,00	1,00
12	0.00	1,00	0,00	1,00
13	0,24	0,76	0,17	0,83
14	0,00	1,00	0,00	1,00
15	0,88	0,12	0,95	0,05
16	0,26	0,74	0,48	0,52
17	0,00	1,00	0,00	1,00
18	0,01	0,99	0,01	0,99
19	0,00	1,00	0,00	1,00
20	0,00	1,00	0,00	1,00

В подавляющем большинстве случаев тексты классифицированы верно, при этом, значения выходных нейронов итоговой нейросети почти всегда приближены к единице или нулю, что говорит об «уверенности» нейросети в полученных результатах. В то же время два текста, один официальный и один любительский, были классифицированы неверно.

Выводы

Правильная классификация 90 % текстов является хорошим результатом для нейросети в первом приближении и сопоставима с ручной классификацией человеком. Такая нейросеть, используемая совместно с методикой, уже пригодна для практического применения в качестве автоматического фильтра официальных текстов, имеющих отношение к авиации из произвольной массы текстов. Возможно дальнейшее улучшение качества классификации, а также расширение на базе созданной нейронной сети возможности классификации по другим признакам (например: приказ, распоряжение, директива), кроме уже разработанного (официальный/неофициальный текст). Также возможно на базе созданной сети введение других способов обработки текста и извлечения из него полезной информации.

В исследованных текстах были замечены два вырожденных случая, которые повлияли на качество классификации, а именно:

1) текст написан блогером в строгом официальном стиле (а не типичном любительском);

2) авиакомпания выпустила пресс-релиз для пассажиров с просторечными выражениями, намеренно в блогерском (т.е. с точки зрения нейросети – неофициальном) стиле. В этой связи, применительно к классификации текста нейросетью, правильно говорить именно о стиле написания текста, а не о его источнике (официальный/неофициальный), хотя в подавляющем большинстве случаев тип источника и стиль текста совпадают.

ЛИТЕРАТУРА

1. Петрухин С.А., Брусникин В.Ю., Шарыпов А.Н., Губанов О.В., Коваль С.В., Карапетян А.Г. Возможные подходы к идентификации авиационной информации, опубликованной в сети Интернет // Научный вестник ГосНИИ ГА. 2019. № 25. С. 65–74.

2. Прикладная и компьютерная лингвистика [Текст] : монография / Под ред.: И. С. Николаева, О.В. Митрениной, Т. М. Ландо. М.: Ленанд; М.: URSS, 2016. 320 с.

3. Бенджамин Бенгфорт, Ребекка Билбро, Тони Охеда. Прикладной анализ текстовых данных. СПб. ПИТЕР. 2020. 368 с.

4. Эдвард А. Патрик. Основы теории распознавания образов. Москва: Советское Радио, 1980. 408 с.

5. Брайан Макмахан, Делип Рао. Глубокое обучение при обработке естественного языка. СПб: ПИТЕР, 2020. 256 с.

6. Каллан Р. Основные концепции нейронных сетей: Пер. с англ. М.: Издательский дом «Вильямс», 2001. 287 с.

7. Барский А.Б. Нейронные сети: распознавание, управление, принятие решений. М.: Финансы и статистика, 2004. 176 с.

8. Джоэл Грас. Наука о данных с нуля. Санкт-Петербург: БХВ-Петербург. 2019. 336 с.

9. Крисилов В.А., Олежко В.Н. Методы ускорения обучения нейронных сетей: Доклад. Одесса. ОГПУ. 2000. С. 2

10. Емельянов В.В., Курейчик В.В., Курейчик В.М. Теория и практика эволюционного моделирования. Москва. Физматлит, 2003. 432 с.

11. Тарик Рашид. Создаем нейронную сеть. СПб: Диалектика, 2018. 272 с.

REFERENCES

1. Petrukhin S.A., Brusnikin V.Yu., Sharypov A.N., Gubanov O.V., Koval S.V., Karapetyan A.G. Possible approaches to the identification of aviation information published on the Internet. *Nauchnyj vestnik GosNII GA= Scientific Bulletin of The State Scientific Research Institute of Civil Aviation.* 2019, no 25, pp. 65–74. (In Russian).

2. Applied and Computational Linguistics. Ed. I.S. Nikolaev, O.V. Mitrennina, T.M. Lando. 2nd Edition. Moscow. Lenand Publ., 2016, 230 p. (In Russian).

3. Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda. Applied analysis of text data. St. Petersburg. Piter Publ., 2020, 368 p. (In Russian).

4. Edward A. Patrick. Fundamentals of the theory of pattern recognition. Moscow. Sovetskoe Radio Publ., 1980, 408 p. (In Russian).

5. Brian McMahan, Delip Rao. Deep learning in natural language processing. St. Petersburg. Piter Publ., 2020, 256 p. (In Russian).

6. Robert Callan. Basic Concepts of Neural Networks.. Moscow, St. Petersburg, Kiev. Williams Publishing House, 2001, 287 p. (In Russian).

7. Barsky A.B. Neural networks: recognition, control, decision making. Moscow. Finansy i Statistika Publ., 2004, 176 p. (In Russian).

8. Joel Grasse. The science of data from scratch. St. Petersburg. BHV-Petersburg Publ., 2019, 336 p. (In Russian).

9. Krisilov V.A., Olezhko V.N. Methods of accelerating the learning of neural networks. Report. Odessa. OGPU Publ., 2000, p. 2.

10. Emelyanov V.V., Kureichik V.V., Kureichik V.M. Theory and practice of evolutionary modeling. Moscow. Physmatlit Publ., 2003, 432 p. (In Russian).

11. Tarik Rashid. Create a neural network. St. Petersburg. Dialectika Publ., 2018, 272 p. (In Russian).

СВЕДЕНИЯ ОБ АВТОРАХ

Петрухин Сергей Александрович, инженер Информационно-аналитического центра ФГУП Государственный научно-исследовательский институт гражданской авиации, ул. Михалковская, 67, корпус 1, Москва, Российская Федерация, 125438, e-mail: petruhin@mlgvs.ru.

Глухов Геннадий Евгеньевич, эксперт Системы добровольной сертификации объектов гражданской авиации, заместитель директора центра по информационным технологиям, ФГУП Государственный научно-исследовательский институт гражданской авиации, ул. Михалковская, 67, корпус 1, Москва, Российская Федерация, 125438; e-mail: glukhov@mlgvs.ru.

Ладыгина Наталья Наильевна, старший специалист, ФГУП Государственный научно-исследовательский институт гражданской авиации, ул. Михалковская, 67, корпус 1, Москва, Российская Федерация, 125438; e-mail: ladygina@mlgvs.ru.

ABOUT THE AUTHORS

Petrukhin Sergey A., Engineer of Center, The State Scientific Research Institute of Civil Aviation, Mikhalkovskaya Street,67, Building 1, 125438 Moscow, Russian Federation, e-mail: petruhin@mlgvs.ru.

Glukhov Gennady E., Expert of System of Voluntary Certification of Civil Aviation Facilities, Deputy Director of the Center for Information Technology, The State Scientific Research Institute of Civil Aviation, Mikhalkovskaya Street, 67, Building 1, 125438 Moscow, Russian Federation; e-mail: glukhov@mlgvs.ru.

Ladygina Natalia N., Chief Specialist, The State Scientific Research Institute of Civil Aviation, Mikhalkovskaya Street,67, Building 1, 125438 Moscow, Russian Federation; e-mail: ladygina@mlgvs.ru.