УДК 004.62, 159.9.07
DOI: 10.25559/SITITO.16.202001.127-131

Нейросетевой метод классификации текстов групп социальной сети Вконтакте

К. С. Николаев*, Н. М. Давлетшин, А. А. Бердников
Казанский (Приволжский) федеральный университет, г. Казань, Россия
420008, Россия, Республика Татарстан, г. Казань, ул. Кремлёвская, д. 18
* konnikolaeff@yandex.ru

Аннотация

Классификация текстовых данных с помощью нейросетей находит свое применение в многочисленных проектах. Применяются различные структуры искусственных нейросетей с варьирующимися параметрами. В данной статье описан процесс создания и тренировки нейросети, направленной на классификацию записей из социальной сети Вконтакте. Цель работы состоит в том, чтобы проверить применимость нейросетей для решения поставленной задачи. В качестве программного инструмента для создания нейросети использовался пакет TensorFlow. Данные для обучения нейросети получены из записей групп с определенной тематикой (биология, информационные технологии, кино). Алгоритм автоматической классификации текстов групп социальной сети Вконтакте, апробированный в исследовании, будет использован при разработке содержательного компонента функциональной психометрической модели когнитивно-поведенческих предикторов активности личности в рамках ее образовательной деятельности.

Ключевые слова: нейронная сеть, классификация, TensorFlow, социальные сети, Вконтакте, прогностика, психометрия, предикторы, виртуальное поведение, метод.

Финансирование: исследование выполнено за счет гранта Российского научного фонда в рамках проекта № 19-18-00253 «Нейросетевая психометрическая модель когнитивно-поведенческих предикторов жизненной активности личности на базе социальных сетей».

Для цитирования: Николаев, К. С. Нейросетевой метод классификации текстов групп социальной сети Вконтакте / К. С. Николаев, Н. М. Давлетшин, А. А. Бердников. – DOI 10.25559/SITITO.16.202001.127-131 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 1. – С. 127-131.

Neural Network Method of Text Classification of Social Network Vkontakte

K. S. Nikolaev*, N. M. Davletshin, A. A. Berdnikov
Kazan (Volga region) Federal University, Kazan, Russia
18 Kremlevskaya St., Kazan 420008, Republic of Tatarstan, Russia
* konnikolaeff@yandex.ru

Abstract

Classification of text data using neural networks is used in numerous projects. Different structures of artificial neural networks with varying parameters can be used. This article describes the process of creating and training a neural network aimed at the classification of records from the social network Vkontakte. The aim of the work is to check the applicability of neural networks to solve the problem. Tensorflow package was used as a software tool to create a neural network. Data for neural network training are obtained from the records of groups with specific topics (biology, information technology, cinema). The algorithm of automatic of Vkontakte groups texts, tested in the study, will be used in the development of the content component of the functional psychometric model of cognitive – behavioral predictors of personal activity in the framework of its educational activities.

Keywords: Neural network, classification, TensorFlow, social networks, Vkontakte, prognostics, psychometrics, predictors, virtual behavior, method.

Funding: the study was supported by a grant from the Russian Science Foundation within the framework of project No. 19-18-00253 "Neural Network Psychometric Model of Cognitive-Behavioral Predictors of Personality's Vital Activity Based On Social Networks."

For citation: Nikolaev K.S., Davletshin N.M., Berdnikov A.A. Neural Network Method of Text Classification of Social Network Vkontakte. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(1):127-131. DOI: https://doi.org/10.25559/SITITO.16.202001.127-131

Введение

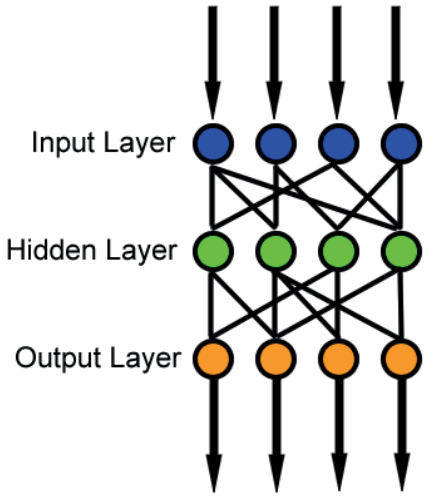
Задача классификации данных (текстовых объектов, изображений, математических данных) часто возникает как подзадача при анализе статистических данных. Особенно популярен анализ данных из социальных сетей, так как из них можно получить большие массивы разнообразных данных. Обработка таких данных вручную не представляется возможным, поэтому на помощь приходят автоматические методы обработки.

Искусственные нейронные сети являются вычислительными моделями, использующими приближающие функции. Нейронные сети показывают себя весьма эффективными в задачах, связанных с классификацией и сегментацией данных. Некоторые сложные нейронные сети используются для моделирования изменяющихся во времени процессов с множеством сложно предсказуемых параметров.

В задачах анализа текстов на естественном языке обычно применяются методы, основанные на машинном обучении либо статистические методы. В работах [1,7] приведен сравнительный анализ эффективности нейронных сетей и метода опорных векторов при классификации текстов. Их результаты позволяют утверждать, что нейронные сети сравнимы по точности с методом опорных веторов, но более экономичны в объемах необходимых для классификации данных. Yan-liang Jin и другие [2] используют двунаправленные рекуррентные нейронные сети для определения значимых для классификации слов в каждом предложении текста. В качестве входных данных они используют тексты на китайском и английском языке. В работах [3, 10-13] и на ресурсе[1] рассматривается применение свёрточных нейросетей для классификации данных. В работах [4] экспериментируют с двухэтапной системой для сентимент-анализа, состоящей из рекуррентной нейросети на первом этапе и свёрточной нейросети на финальном этапе анализа. Pengfei Liu и другие [6] анализируют результаты многозадачной рекуррентной нейросети, улучщающей результаты своего обучения с помощью обмена результатами, полученными различными частями нейросети.

В работах [5, 9] рассматриваются подходы, моделирующие тему (LDA, Латентно-семантический анализ) и методы, основанные на машинном обучении (наивный байесовский подход, метод k ближайших соседей и K-means) для выделения тематик текстов. В свою очередь, в работе [8] авторы рассматривают эффективность совмещения наивного байесовского подхода с предварительным применением двухэтапного метода максимальной предельной релевантности для выбора характеристик, в целях его ускорения.

Как показывает опыт указанных авторов, использование нейросетей для классификации текстов является оправданным и весьма эффективным методом. Необходимо выбрать тип и сложность нейронной сети, подходящей для нашей задачи. К самым простым типам нейросетей относятся сети прямой связи (многослойный персептрон), в которых перемещение данных происходит в одном направлении, от слоя входа до слоя выхода, как изображено на Рисунке 1.



Р и с. 1. Структура нейросети прямой связи
F i g. 1. Feedforward neural network structure

Рассмотрим краткое описание сетей прямой связи:

1) Автокодировщик, который нацелен на повышение устойчивости нейросети к шуму в данных;
2) Вероятностная нейронная сеть, в алгоритме которой родительская функция распределения вероятностей (PDF) каждого класса аппроксимируется с помощью окна Парцена и непараметрической функции;
3) Нейронная сеть с временной задержкой, в которой ради распознавания свойств вне зависимости от положения в последовательности ко входным данным добавляются задержки;
4) Свёрточные нейронные сети, которые чередуют свёрточные и субдискретизирующие слои.

В данной работе в качестве структуры нейронной сети используется персептрон с двумя скрытыми слоями. Данные для обучения нейросети извлекаются из публикаций тематических групп социальной сети Вконтакте. Обучение нейросети для решения задачи классификации текстов на русском языке из социальных сетей позволит использовать ее для определения интересов пользователей Вконтакте и определения их личностных характеристик. Полученный алгоритм будет использован при разработке содержательного компонента функциональной психометрической модели когнитивно-поведенческих предикторов активности личности в рамках ее образовательной деятельности.

Подготовка данных

В ходе подготовки данных был реализован скрипт на языке Python с использованием API социальной сети Вконтакте. При создании скрипта были использованы следующие библиотеки: requests, time, json, csv, datetime. С помощью метода get_all_posts данные выгружаются из выбранной группы или персональной страницы.

[1] Batista, D. S. Convolutional Neural Networks for Text Classification [Электронный ресурс]. URL: http://www.davidsbatista.net/blog/2018/03/31/SentenceClassificationConvNets (дата обращения: 06.12.2019).

Таблица 1. **Входные данные**
Table 1. **Input data**

Название группы	Количество подписчиков	Количество извлеченных постов	Тематика	Тренировочная или тестовая коллекция
Типичный программист	478 тысяч	21283	IT	Тренировочная
Hell's Kitchen	156 тысяч	40732	Кино	Тренировочная
Биология | VK	354 тысячи	25723	Биология	Тренировочная
Биология	40 тысяч	4169	Биология	Тренировочная
Флора и Фауна	143 тысячи	8718	Биология	Тестовая

Метод write_in_file_csv выгружает данные в csv файл, который в свою очередь используется в качестве входных данных для обучения нейронной сети. Данные в csv файле были дополнительно перемешаны для того, чтобы можно было без особых потерь в результатах обучения отрезать от массива входных данных части разного размера.

Создание и настройка нейросети

Рабочим инструментом при создании нейросети стала библиотека для языка Python - TensorFlow от компании Google. Данный продукт позволяет быстро создать и настроить нейросеть с необходимыми параметрами.
Для начала экспериментов по обучению нейросети были выбраны параметры, указанные в Для начала экспериментов по обучению нейросети были выбраны параметры, указанные в Для начала экспериментов по обучению нейросети были выбраны параметры, указанные в Для начала экспериментов по обучению нейросети были выбраны параметры, указанные в Для начала экспериментов по обучению нейросети были выбраны параметры, указанные в Таблица 2. Такие параметры показали неплохие результаты на тестовой коллекции, а именно 88% точности классификации текста.2. Такие параметры показали неплохие результаты на тестовой коллекции, а именно 88% точности классификации текста.2. Такие параметры показали неплохие результаты на тестовой коллекции, а именно 88% точности классификации текста.2. Такие параметры показали неплохие результаты на тестовой коллекции, а именно 88% точности классификации текста.2. Такие параметры показали неплохие результаты на тестовой коллекции, а именно 88% точности классификации текста.

Таблица 2. **Стартовые параметры нейросети**
Table 2. **Starting parameters of the neural network**

Параметр	Стартовое значение
Размер шага	0.01
Эпохи обучения	9
Количество элементов первого скрытого слоя	100
Количество элементов второго скрытого слоя	100
Размерность входных данных	421972
Количество текстов из тренировочных данных	30000

В таблице 2 приведены изменения точности классификации в зависимости от изменения тех или иных параметров обучения. Изменению подлежат параметры «Размер шага» и «Количество текстов из тренировочных данных».
Пары параметров с наиболее высокими показателями точности (а именно «20000 входных текстов, шаг обучения 0,014» и «30000 входных текстов, шаг обучения 0,014») были перепроверены на других объемах тестовых данных и результаты оказались следующими: при 8000 записей в тестовом наборе и указанными параметрами нейросети, точность составила 0.878 для первой пары параметров и 0.885 для второй пары.

Таблица 3. **Зависимость точности обучения от параметров нейросети**
Table 3. **Dependence of the training accuracy on the neural network parameters**

Количество входных текстов	Шаг обучения	0,008	0,01	0,012	0,014
20000		0,862564	0,851795	0,862051	0,911795
25000		0,846154	0,865641	0,876410	0,887179
30000		0,843077	0,849231	0,884615	0,901538
35000		0,840513	0,878974	0,883077	0,884355

Выводы

В ходе данной работы мы создали нейронную сеть, которая с достаточной точностью способна распознать тематику текстов из тематических групп и пользовательских страниц. Полученные результаты позволяют с уверенностью сказать, что итоговая нейронная сеть успешно справится с последующими задачами классификации текстов на персональных страницах пользователей и послужит составной частью системы по определению их круга интересов. Результаты выступают составной частью исследования психометрических предикторов образовательной успешности студентов в рамках разработки теоретико-прикладной модели прогнозирования жизненной активности личности в ее образовательной и профессиональной деятельности через социальные сети.

References

[1] Zaghloul W., Lee S.M., Trimi S. Text classification: neural networks vs support vector machines. *Industrial Management & Data Systems*. 2009; 109(5):708-717. (In Eng.) DOI: https://doi.org/10.1108/02635570910957669
[2] Jin Y., Luo C., Guo W., Xie J., Wu D., Wang R. Text Classification Based on Conditional Reflection. *IEEE Access*. 2019;

7:76712-76719. (In Eng.) DOI: https://doi.org/10.1109/ACCESS.2019.2921976

[3] Prusa J. D., Khoshgoftaar T. M. Designing a Better Data Representation for Deep Neural Networks and Text Classification. In: 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA; 2016. p. 411-416. (In Eng.) DOI: https://doi.org/10.1109/IRI.2016.61

[4] Wen Y., Zhang W., Luo R., Wang J. Learning text representation using recurrent convolutional neural network with highway layers. arXiv:1606.06905 [cs.CL], 2016. Available at: https://arxiv.org/abs/1606.06905 (accessed 06.12.2019). (In Eng.)

[5] Rajasundari T., Subathra P., Kumar P.N. Performance Analysis of Topic Modeling Algorithms for News Articles. *Journal of Advanced Research in Dynamical and Control Systems*. 2017; 11:175-183. (In Eng.)

[6] Liu P., Qiu X., Huang X. Recurrent neural network for text classification with multi-task learning. arXiv:1605.05101 [cs.CL], 2016. Available at: https://arxiv.org/abs/1605.05101 (accessed 06.12.2019). (In Eng.)

[7] Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). AAAI Press; 2015. p. 2267-2273. (In Eng.)

[8] Fauzi M.A., Arifin A.Z., Gosaria S.C. Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model. *Indonesian Journal of Electrical Engineering and Computer Science*. 2017; 8(3):610-615. (In Eng.) DOI: http://doi.org/10.11591/ijeecs.v8.i3.pp610-615

[9] Bindu K. R., Parameswaran L., Soumya K.V. Performance Evaluation of Topic Modelling Algorithms with an application of Q A Dataset. *International Journal of Applied Engineering Research*. 2015; 10(73):23-27. (In Eng.)

[10] Jacovi A., Shalom O.S., Goldberg Y. Understanding Convolutional Neural Networks for Text Classification. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Brussels, Belgium; 2018. p. 56-65. (In Eng.) DOI: http://doi.org/10.18653/v1/W18-5408

[11] Shankar S., Robertson D., Ioannou Y., Criminisi A., Cipolla R. Refining Architectures of Deep Convolutional Neural Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV; 2016. p. 2212-2220. (In Eng.) DOI: http://doi.org/10.1109/CVPR.2016.243

[12] Iandola F.N., Han S., Moskewicz M.W., Ashraf K., Dally W.J., Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360v4 [cs.CV], 2016. 13 p. Available at: https://arxiv.org/pdf/1602.07360.pdf (accessed 06.12.2019). (In Eng.)

[13] Teerapittayanon S., McDanel B., Kung H.T. Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices. In: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA; 2017. p. 328-339. (In Eng.) DOI: http://doi.org/10.1109/ICDCS.2017.226

[14] Skansi S. Convolutional Neural Networks. In: Introduction to Deep Learning. Undergraduate Topics in Computer Science. Springer, Cham; 2018. (In Eng.) DOI: https://doi.org/10.1007/978-3-319-73004-2_6

[15] Zhang X., Zhao J., LeCun Ya. Character-level convolutional networks for text classification. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA; 2015. p. 649-657. (In Eng.)

Поступила 06.12.2019; принята к публикации 28.03.2020;
опубликована онлайн 25.05.2020.
Submitted 06.12.2019; revised 28.03.2020;
published online 25.05.2020.

Об авторах:

Николаев Константин Сергеевич, старший лаборант, Казанский (Приволжский) федеральный университет (420008, Россия, Республика Татарстан, г. Казань, ул. Кремлёвская, д. 18), ORCID: http://orcid.org/0000-0003-3204-238X, konnikolaeff@yandex.ru

Давлетшин Нагим Мухаматнурович, лаборант, Казанский (Приволжский) федеральный университет (420008, Россия, Республика Татарстан, г. Казань, ул. Кремлёвская, д. 18), ORCID: http://orcid.org/0000-0002-5807-8527, davlet-9@ya.ru

Бердников Андрей Анатольевич, лаборант, Казанский (Приволжский) федеральный университет (420008, Россия, Республика Татарстан, г. Казань, ул. Кремлёвская, д. 18), ORCID: http://orcid.org/0000-0003-3837-9821, andreyberdnikov1998@gmail.com

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Konstantin S. Nikolaev, Senior Laboratory Assistant, Kazan (Volga region) Federal University (18 Kremlevskaya St., Kazan 420008, Republic of Tatarstan, Russia), ORCID: http://orcid.org/0000-0003-3204-238X, konnikolaeff@yandex.ru

Nagim M. Davletshin, Laboratory Assistant, Kazan (Volga region) Federal University (18 Kremlevskaya St., Kazan 420008, Republic of Tatarstan, Russia), ORCID: http://orcid.org/0000-0002-5807-8527, davlet-9@ya.ru

Andrey A. Berdnikov, Laboratory Assistant, Kazan (Volga region) Federal University (18 Kremlevskaya St., Kazan 420008, Republic of Tatarstan, Russia), ORCID: http://orcid.org/0000-0003-3837-9821, andreyberdnikov1998@gmail.com

All authors have read and approved the final manuscript.