УДК 004.93

З. В. ГАЛИМЗЯНОВА

магистр

Казанский федеральный университет

Ф. Б. СИТДИКОВА

кандидат филологических наук, старший преподаватель

Казанский федеральный университет

КЛАССИФИКАЦИИ ТЕКСТА С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

TEXT CLASSIFICATION USING NEURAL NETWORKS

Аннотация. Искусственные нейронные сети (ИНС) в настоящее время широко используются во всем мире для решения проблем в различных сферах. В статье рассматривается история нейронных сетей, их обучение и применение в классификации текста.

Ключевые слова: искусственные нейронные сети, ИНС, компьютерная лингвистика, обучение ИНС, классификация текста, word2vec.

Abstract. Artificial neural networks (ANN) are now widely used throughout the world to solve problems in various fields. The article deals with the history of neural networks, their training and application in text classification.

Kew words: artificial neural networks, ANN, computer linguistics, training ANN, text classification, word2vec.

Before the advent of computer technologies people had to analyse texts on themselves. Very often this task was impossible because of the huge amount of data that could not be processed in proper time, the work could last for such a long period of time that by the moment of its potential completion, the task might no longer be relevant. Nowadays, in the IT age, the problem of text analyzing can be solved in a short time.

Applications of artificial intelligence methods are becoming more and more widely used in everyday human life.

One of the essential and typical task in artificial intelligence (AI) is document or text classification. Categorization of different documents, (e.g. a web page, novel, media article etc.) has many applications like spam filtering, email routing, sentiment analysis etc.

In other terms classification is the problem of recognition of which of categories set a new object belongs, based on a training the set of data inclusive examples whose category is already known. For example, we can classify a given email into "spam" or "non-spam" classes, we can also assign a genre to a given fiction book, etc.

The above mentioned problem can be solved by means of the following classification methods: the Naive Bayes Classifier [3], Decision Trees [2], Support Vector Machine (SVM) [4] and Neural Networks(NN) [1].

An artificial neural network (ANN) is a mathematical model, as well as its software or hardware implementation based on the same principles as human neural networks. In 1943, Warren McCulloch and Walter Pitts were the first to create a computational model for neural networks [6].

A neural network is a method of solving various problems in machine learning such as prediction and recognition. The network is a set of neurons aligned in a specific structure with different connections and the selected activation function. In other words, it is a graph where the edges have certain weights. Typically, a neural network consists of an input layer, a hidden layer / layers, and an output layer.

Figure 1 represents an example of a neural network consisting of one hidden layer, and an output layer having 1 value.

There are many different ways to train ANN. ANN training is an algorithm which calculates the optimal parameters of the model (the weights and thresholds), in order for a given input to the network to produce a favored output. When the neural network is well trained, it can be used to solve practical problems.

Consider the work of the ANN which was trained on the basis of pre-trained word vectors for sentence classification problems. The word vector is a dense representation of words in a low-dimensional vector space. That is, words or phrases from the dictionary correspond to vectors of real numbers. A simple neural network with a small tuning of hyper-

parameters and static vectors provides excellent results by several criteria [5].

We use the public word2vec vectors [7] that have been trained on the base of 100 billion words from Google News. These vectors have a dimension of 300. The words that are not present in the set of pre-trained words are initialized randomly. We also use reviews of films with one sentence for each review as training and testing data [8]. Classification process includes the detection of both positive and negative reviews.

Then we extract functions for training our ANN. Thus, we find the average value of word vectors sum for each sentence of the training data [9]. We place each such vector in correspondence with the label – a positive or negative review. We train our neural network by the extracting functions and labels. Once our model is trained, we test it. The well trained ANN has an accuracy about 0.79-0.8, and it takes only 2 seconds to determine a single sentence class.

Our results complement the well-proven evidence that uncontrolled pre-training of word vectors like word2vec is an important component of machine learning for natural language processing (NLP). For example, word2vec can be used to translate text in case we have two sets of word vectors trained on two different languages. We do not need to change the source code of word2vec. Teaching the model does not depend on the language, training our vectors depends on the neighboring terms that appear in the co-occurrence window.

Machine learning models, such as ANNs, have recently achieved amasing results in many practical fields. Analyzing different kinds of texts is one of the important and common tasks.

In our paper we have described an experiment with ANN which was built on top of word2vec. and achieved positive results.

References

1. Bishop C. M. Neural Networks for Pattern Recognition. – Oxford, New York: Oxford University Press, 1996.

2. Breiman L. Random Forests. // Machine Learning, 45, 5 –32. – 2001. – P. 5– 32.

3. Cichosz P. Naïve Bayes classifier // Data Mining Algorithms. – John Wiley & Sons, Ltd. – 2015. – P. 118–133.

4. Cortes C., Vapnik V. Support-Vector Networks. // Machine Learning, 20 (3). – 1995. – P. 273–297.

5. Google Code Archive. Long-term storage for Google Code Project Hosting. – URL: https://code.google.com/archive/p/word2vec/. (accessed November 10, 2017).

6. Kim Y. Convolutional Neural Networks for Sentence Classification. – New York University. – 2014.

7. McCulloch W., Walter P. A Logical Calculus of Ideas Immanent in Nervous Activity // Bulletin of Mathematical Biophysics. – 1943, 5 (4). – P. 115–133.

8. Movie Review Data. – URL: https://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/. (accessed November 10, 2017).

9. Neural network models (supervised). – URL: http://scikit-learn.org/stable/modules/neural_networks_supervised.html