e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2022. № 3
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2022. № 3

UDC 004.85

MULTILINGUAL TEXT CLASSIFIER USING PRE-TRAINED UNIVERSAL SENTENCE ENCODER MODEL

Orlovskiy O. V. – Post-graduate student, Computer Systems Software Department, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

Khalili Sohrab – CEO, CreateITTogether LLC Company, Fullerton, CA, USA.

Ostapov S. E. – Professor, Head of Computer Systems Software Department, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

Hazdyuk K. P. – Assistant, Computer Systems Software Department, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

Shumylyak L. M. – Assistant, Computer Systems Software Department, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

ABSTRACT

Context. Online platforms and environments continue to generate ever-increasing content. The task of automating the moderation of user-generated content continues to be relevant. Of particular note are cases in which, for one reason or another, there is a very small amount of data to teach the classifier. To achieve results under such conditions, it is important to involve the classifier pre-trained models, which were trained on a large amount of data from a wide range. This paper deals with the use of the pre-trained multilingual Universal Sentence Encoder (USE) model as a component of the developed classifier and the affect of hyperparameters on the classification accuracy when learning on a small data amount (~ 0.05% of the dataset).

Objective. The goal of this paper is the investigation of the pre-trained multilingual model and optimal hyperparameters influence for learning the text data classifier on the classification result.

Method. To solve this problem, a relatively new approach to few-shot learning has recently been used – learning with a relatively small number of examples. Since text data is still the dominant way of transmitting information, the study of the possibilities of constructing a classifier of text data when learning from a small number of examples (~ 0.002–0.05% of the data set) is an actual problem.

Results. It is shown that even with a small number of examples for learning (36 per class) due to the use of USE and optimal configuration in learning can achieve high accuracy of classification on English and Russian data, which is extremely important when it is impossible to collect your own large data set. The influence of the approach using USE and a set of different configurations of hyperparameters on the result of the text data classifier on the example of English and Russian data sets is evaluated.

Conclusions. During the experiments, a significant degree of relevance of the correct selection of hyperparameters is shown. In particular, this paper considered the batch size, optimizer, number of learning epochs and the percentage of data from the set taken to train the classifier. In the process of experimentation, the optimal configuration of hyperparameters was selected, according to which 86.46% accuracy of classification on the Russian-language data set and 91.13% on the English-language data, respectively, can be achieved in ten seconds of training (training time can be significantly affected by technical means used).

KEYWORDS: few shot learning, low-data learning, pre-trained models, USE, neural networks, data mining, data set, text data classifier.

ABBREVIATIONS

USE is a Universal Sentence Encoder;
SGD is a Stochastic gradient descent;
RMSProp is a Root Mean Squared Propagation;
Adam is a Adaptive Moment Optimization.

NOMENCLATURE

O_T is a set of optimizer's type;
o_T is an element of an set of optimizer's type;
P is a parameters set;
p_j is an element of a parameters set;
N_{par} is a parameters number;
P' is a specific parameters set for each training subset;
M is a toxic messages dataset;
m_i is a toxic message;
M^k is a training subset of the toxic messages;
L is a language of dataset;
S is a size of dataset (in MB);
N_S is a number of records in dataset;
N_{cat} is a classification categories number in the dataset;
N_{sam} is a proportion of the original samples in training subsample (in %);
N_{ep} is a number of executed epochs of neural network training;
Ac is a classification accuracy;
$F()$ is a function depends on M, M^k, P' which describes Ac;
max is a function $F()$ maximum.

INTRODUCTION

Deep learning systems using large amounts of data have repeatedly shown their effectiveness in a wide range of classification problems [1]. However there are often situations in which it seems impossible to prepare a sufficient number of marked examples for classifier training or requires the involvement of resources that do not justify the expected end result. To solve this problem, a relatively new approach to few-shot learning has recently been used – learning with a relatively small number of examples. Since text data is still the dominant way of transmitting information [2], the study of the possibilities of constructing a classifier of text data when learning

DOI 10.15588/1607-3274-2022-3-10

from a small number of examples (~ 0.002 – 0.05% of the data set) is an urgent task.

Another important bonus for improving the efficiency of development time will be the ability to classify text simultaneously in several of the most popular languages using a single model. In particular, this paper investigates the results of the model's work on texts created in Russian and English.

The object of study is the process of toxic message classification.

The subject of study is the investigation of the pre-trained USE-model on the classification accuracy.

The purpose of the work is the development and investigation of the multilanguage classificator on the base of pre-trained USE-model.

1 PROBLEM STATEMENT

The challenge facing the authors of the paper is as follows. For each specific set of toxic messages $M=\{m_i\}$, where $i=1,\ldots, N_S$, it is necessary to select the training subset $M^k \in M$ and choose the best optimizer type $o_T \in O_T$ (with the parameter set $P=\{p_j\}$, where $j=1, \ldots, N_{par}$) so that specific parameters values $P' \in P | \forall M^k \in M$ made it possible to achieve the maximum classification accuracy, i.e. $Ac=F(M, M^k \in M, P') \rightarrow \max$ for each classifier type $o_T \in O_T$. An additional condition imposed on the data subset is that its amount does not exceed 0.05% of the complete dataset, such as $N_{sam} \leq 0.00005\ N_S$.

2 REVIEW OF THE LITERATURE

In our previous paper, an overview of typical approaches used in the development of text data classifiers, in particular on the example of the classification of destructive messages [3] was made. Special attention was paid to the problem of the data preprocessing methods affect for learning process.

This paper deals with the study of the influence of the pre-trained USE model on the accuracy of the classification of text messages with learning process, which uses only several examples per class.

The paper [4] discusses the problem of data augmentation in a small data subset. Initially, the classifier uses several original examples per class, and then several artificially created examples, which aim, if possible, to comprehensively reflect the features of a particular class. Thus, it is expected that several universal artificial examples will help to replace the lack of a large number of instances, each of which reflects a certain aspect of the class in its own way.

Research [5] helps us to better understand how few shot models work in general, how different approaches to their construction differ, what are the advantages and disadvantages of this class of models developed over the last few years. Also in the work special attention is paid to the use of transformers, which is relevant for our model.

The article [6] deals with the affect of pre-trained models when they are used as components of the model. The results obtained by the authors for the problem of text generation by involving a previously trained model in the developed generator encourage us to investigate the effect that such a solution may have for the classification problem.

The problem of classes optimization in the classification process requires special attention, especially with regard to their quantity, potential merger or replacement. This can also greatly affect both the speed of classifier development and the data preparation. Details of the classes composition and their potential modification are demonstrated in [7].

The article [8] demonstrates the examples of the pre-trained model from Google – Universal Sentence Encoder (USE) using [9]. In particular, a wide range of tasks for which the model can be used is shown, where the task of classifying text data is only one of the possibilities.

Investigation in the paper [10] demonstrate the inclusion of the optimal hyperparameters choice in classifier training, including studies of the effectiveness of various optimizers of the data, such as Adam and its modifications.

As mentioned earlier the main purpose of this paper is to study the influence of the pre-trained multilingual model and the optimal parameters for learning the text data classifier on the classification result. To solve the problem, a classifier based on an artificial neural network was used. One of the network layers will be the pre-trained USE model [9]. Different configurations of hyperparameters were tested during the training. The classification results were verified on the two data sets described below.

3 MATERIALS AND METHODS

The experiments were performed using two datasets. The first – "Fake or real news dataset" [11] has the following characteristics, presented in Table 1.

Table 1 – Characteristics of the data set "Fake or real news dataset"

Characteristic	Value
Language, *L*	English
Dataset size, *S*	~ 29 MB
Number of samples, *Ns*	~ 6335
Number of classification cathegories, *Ncat*	1 (fake)

The second dataset, "Russian Language Toxic Comments. Small dataset with labeled comments from 2ch.hk and pikabu.ru" [12], has the following characteristics presented in Table 2.

Note that although both datasets are intended for the classification problem, these tasks are somewhat different. In the first case, we find "fake" or real news, and in the second – toxic or not a certain message.

Data for training process on both datasets were distributed as shown in Table 3.

DOI 10.15588/1607-3274-2022-3-10

Table 2 – Characteristics of the dataset "Russian Language Toxic Comments. Small dataset with labeled comments from 2ch.hk and pikabu.ru"

Characteristic	Value
Language, *L*	Russian
Dataset size, *S*	4.45 MB
Number of samples, *Ns*	~ 11.500
Number of classification cathegories, *Ncat*	1 (toxic)

Table 3 – Data distribution for training process

Training stage	Data distribution
Training process	0.002 – 0.05%
Validation/testing process	99.998 – 99.95%

The investigations were performed using a neural network, the architecture of which is schematically shown in Fig. 1.

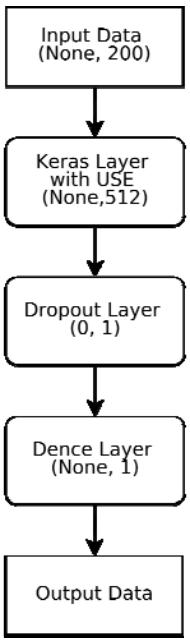

Figure 1 – Scheme of neural network architecture based on USE model

As can be seen from the figure, our classifier contains three layers. The main one is KerasLayer, which includes a massive pre-trained USE model [9]. The Dropout-Dense layer combination block helps to avoid re-learning the classifier and helps to reduce the dimension of the network and, as a result, speed up learning.

Consider the architecture in more detail. A list of English and Russian sentences (depending on the data set) of different lengths is transmitted to the input in the presented classifier based on the neural network. It is known that to learn the network, we can not use as input elements of the word in their usual form, we perform the following manipulations on the data set:

1. Tokenization. For example: we transform every object that looks like "Hello, gentlemen!" to an array of unique words ["hello", "gentlemen"] without punctuation.

2. Indexing. We create a dictionary from the resulting array of all unique words in the dataset, which looks like this: {1: "hello", 2: "gentlemen", ...}.

3. Indexes representation. For each of the objects in the dataset, we form an array in which the words involved in the object are represented as indexes from our dictionary. For example: [1, 2, 1].

Also, one of the typical problems we have to face when preparing data for training is the problem of different sizes of objects among our data. We need to bring the learning elements to the same dimension. This is solved using the padding technique: choose the maximum value of words, such as 200, and fill the remaining spaces in each object with zeros. If the object contains more words than the selected maximum value – each subsequent word after the maximum is cut off. Although in the data sets we have chosen, the size of most sentences does not exceed the value of 100 words, nevertheless, the dimension with a value of 200 is chosen to capture atypical cases, if any.

The basis of the presented classifier is KerasLayer, which is connected to the pre-trained USE model [13] based on the "transformer" architecture presented by researchers from Google in 2019. This model exists in several variations, but in these experiments a multilingual version was chosen (16 languages, including English and Russian). Also noteworthy is the fact that the purpose of the model is not only to classify but also to cluster texts, find their semantic similarities, as well as some multilingual operations. In the experiments conducted, the developed nature of the model related to multilingual classification was useful.

Having received from USE the resulting tensor, we transfer it to the Dropout layer. Its purpose, in this case, is to weed out a certain percentage of nodes, which we will establish, replacing them with a value of zero. This is necessary so that nodes at the next level are forced to process missing data representations. In this way we achieve an effect in which the result of the whole network has the best level of generalization – we avoid the effect of retraining, in which the network can show good results on a familiar data set and far from the desired results on an unfamiliar set. In the presented experiments the value of random exclusion of nodes in 10% was used. Of course, this percentage can be selected empirically.

After passing the Dropout layer we transfer the data to the Dense layer and the RELU activation function is applied to them. Then the result passes through the sigmoidal activation function, where the classification for each of the labels takes place, and we get a value between 0 and 1.

Of course, the presented architecture can be optimized, but this is more of a challenge for the future. Now our task is to determine the influence of the pre-trained USE-model on the results of the classifier.

DOI 10.15588/1607-3274-2022-3-10

4 EXPERIMENTS

Note first of all that in the experiments, a combination of different sets of optimizers, loss functions and other hyperparameters was tested. Unless otherwise noted, the default optimizer was Adam, a loss function: binary crossentropy.

Initially, the classifier was trained in the "basic mode" on 0.002% of examples from the dataset (few-shot learning). The batch parameter is equal 4. It is optimal taking into account the hardware used for training. Number of epochs – 2. With such settings, we obtained the accuracy of the classification in the range of 73.85–74.17%. In this case, and further we mean the range obtained by repeated experiments with the same parameters in the samples described in Table 1 and Table 2.

Next, we conducted an iterative experiment consisting of the following steps:

1. We change one of the key parameters that may affect the resulting classification accuracy (batch; number of epochs in training; dataset percentage included in the sample for training (train); used optimizer). Note that we consider this list not exhaustive of possible options. Nevertheless, the influence of these parameters is investigated in the experiments presented in this paper.

2. We teach the classifier on the selected dataset without changing other parameters.

3. Measure the classification accuracy.

The experiment's results are presented in the next chapter (see Table 4).

5 RESULTS

Note that the configuration with the Adamax optimizer proved to be the best in the considered experiments (№7 in Table 4). We obtained the maximum classification accuracy of 0.9113 on the English-language dataset [11] by repeating the experiment with the same parameters.

Table 4 – The results of the classifier when using different hyperparameters on datasets [11, 12]

No	O_T	N_{ep}	N_{sam}	Ac
1	Adam	2	0.002%	0.7385–0.7417
2	Adam	7	0.002%	0.7826–0.6771
3	Adam	2	0.005%	0.7074–0.7460
4	Adam	10	0.005%	0.6651–0.7657
5	SGD	2	0.005%	0.5552 – 0.6649
6	NAdam	2	0.005%	0.8027 – 0.8104
7	Adamax	2	0.005%	0.8475 – 0.8646
8	RMSProp	2	0.005%	0.7773 – 0.7839

6 DISCUSSIONS

Analyzing the results described in Table 4, we immediately note the key advantage of the few-shot learning approach. Using only 0.002% of samples N_{sam} and two learning epochs N_{ep}, we obtained a quite acceptable result of classification accuracy Ac in the range 0.7385 – 0.7417%. This amount of data used and the number of epochs can significantly reduce network learning time. Depending on the used hardware, the speed of the learning process can vary, however, we can safely say about ten seconds to complete the experiment. This can be extremely relevant when prototyping a certain idea on selected data, when you need to get a quick result and already starting from it to build a further, more detailed experiment. Also, such a scenario may be quite applicable in an area where it is impossible or impractical to collect a relatively large data amount for classifier training, and the value of a quick result on a relatively small amount of "live" data is significant.

Continuing to experiment, we noted that if the number of epochs increases to 7 while maintaining the previous values in the above configuration, we can observe an expansion of the range of classification accuracy (#2 in Table 4). In particular, the lower accuracy limit fell by 6.14%, and the upper accuracy limit increased by only 4.09%. At the same time, after graduating from the 7th epoch, the classifier steadily came to a state of reduced accuracy. Examining this question, we came to the conclusion that it is necessary to continue the selection of the optimal configuration of hyperparameters. In experiments #3 and #4, we tried to increase the number of examples for training to 0.005% of the data samples. As we can see in Table 4, the result is slightly different, but the general trend repeats the result of experiments #1 and #2.

The next hyperparameter for selection was the optimizer type N_{cat}. We first used the SGD optimizer (experiment #5 in Table 4) for two epochs and 0.005% of the sample data. However, the best result in the range (0.6649) was 4.25% behind the worst result obtained with the Adam optimizer for the same other experimental parameters. In our opinion, this is most likely due to the fact that this optimizer performs better when working with other types of data, in contrast to text data in our experiments.

In Experiment #8 we usied the RMSProp optimizer and improved the result obtained with Adam while maintaining other parameters at the same level. Based on the lower bar of the accuracy range, we can reached an improvement of 6.99%. However, the best results in our experiments were achieved using modifications of the Adam optimizer. In particular, using NAdam, we recorded improvements in the lower bar of the accuracy range by 9.53%, and with Adamax by as much as 14.01%! The results are shown in Table 4 in experiments #6 and #7, respectively. Based on the obtained results, we consider this configuration with the described hypeparameters to be optimal when training the classifier on text data.

DOI 10.15588/1607-3274-2022-3-10

CONCLUSIONS

The few-shot learning approach is extremely relevant in a large number of domains, where collecting and preparing a large set of data for learning seems impractical.

The universal knowledge base taken out of the cognition of our datasets is the pre-trained multilingual USE model, which allows simultaneous work with data in 16 languages, of which 2 are used in this work.

In our experiments, the optimal configuration of hyperparameters was selected, according to which 86.46% accuracy of classification on the Russian-language data set and 91.13% on the English-language data, respectively, can be achieved in ten seconds of training (training time can be significantly affected by technical means used).

The scientific novelty. It is shown that even with a small number of examples for learning (36 per class) due to the use of USE and optimal configuration in learning can achieve high accuracy of classification on English and Russian data, which is extremely important when it is impossible to collect your own large dataset.

The practical significance. The obtained results allow to build classifiers of text data with a sufficiently high rate of accuracy in the presence of a small amount of data for learning.

Prospects for further research. In the following studies, you can take into account more hyperparameters to analyze their impact on the final result of the classifier. It is also quite relevant to compare the influence of different pre-trained analog models according to USE, which we relied on in conducting all the experiments described in this paper.

The urgent problem of mathematical support development is solved to automate the sampling at diagnostic and recognizing model building by precedents.

The scientific novelty of obtained results is that the method of training sample selection is firstly proposed. It determines the weights characterizing the term and feature usefulness for a given initial sample of precedents and given feature space partition. It characterizes the individual absolute and relative informativity of instances relative to the centers and the boundaries of feature intervals based on the weight values. This allows to automate the sample analysis and its division into subsamples, and, as a consequence, to reduce the training data dimensionality. This in turn reduces the time and provides an acceptable accuracy of neural model training.

The practical significance of obtained results is that the software realizing the proposed indicators is developed, as well as experiments to study their properties are conducted. The experimental results allow to recommend the proposed indicators for use in practice, as well as to determine effective conditions for the application of the proposed indicators.

Prospects for further research are to study the proposed set of indicators for a broad class of practical problems.

ACKNOWLEDGEMENTS

This investigation is supported by the state budget scientific research project of Yuriy Fedkovych Chernivtsi National Univesity “Investigation, simulation and software development for complex dynamic systems” (state registration number state 0121U109232).

REFERENCES

1. Yann L., Yoshua B., Geoffrey H. Deep learning, *Nature,* 2015, Vol. 521(7553), pp. 436–444.
2. Ma L., Goharian N., Chowdhury A. et al. Extracting unstructured data from template generated web documents, *Information and knowledge management, Twelfth international conference, 2003, proceedings*, 2003, pp. 512–515.
3. Orlovskyi O., Ostapov S. Analysis of the text preprocessing methods influence on the destructive messages classifier, O.Orlovskyi, *Advanced Information Systems,* 2020, Vol. 4(3), pp.104–108.
4. Few-Shot Text Classification with Triplet Networks, Data Augmentation, and Curriculum Learning [Electronic resource], Access mode: https://arxiv.org/abs/2103.07552
5. A Neural Few-Shot Text Classification Reality Check [Electronic resource]. Access mode: https://arxiv.org/abs/2101.12073
6. Few-Shot Text Generation with Pattern-Exploiting Training [Electronic resource]. Access mode: https://arxiv.org/abs/2012.11926
7. Halder K., Akbik A., Krapac J. et al. Task-Aware Representation of Sentences for Generic Text Classification, *Computational Linguistics, 28th International Conference, December 2020, proceedings*, 2020, P. 3202–3213.
8. Reddy T., Williams R., Breazeal C. Text classification for AI education [Electronic resource]. Access mode: https://robots.media.mit.edu/wp-content/uploads/sites/7/2021/01/Text_classifier.pdf
9. Universal-sentence-encoder-multilingual-large. 16 languages (Arabic, Chinese-simplified, Chinese-traditional, English, French, German, Italian, Japanese, Korean, Dutch, Polish, Portuguese, Spanish, Thai, Turkish, Russian) text encoder [Electronic resource]. Access mode: https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3
10. Sriporn K., Tsai C. F., Tsai C. E. et al. Analyzing Malaria Disease Using Effective Deep Learning Approach, *Diagnostics*, 2020, No. 10, pp. 744–749.
11. Fake or real news dataset [Electronic resource]. Access mode:https://github.com/lutzhamel/fake-news/blob/master/data/fake_or_real_news.csv.
12. Russian Language Toxic Comments. Small dataset with labeled comments from 2ch.hk and pikabu.ru [Electronic resource]. Access mode: https://www.kaggle.com/blackmoon/russian-language-toxic-comments.
13. Yang Y.. Cer D., Ahmad A. et al. Multilingual Universal Sentence Encoder for Semantic Retrieval, [Electronic resource]. Access mode: https://aclanthology.org/2020.acl-demos.12.pdf

Received 14.01.2022.
Accepted 22.06.2022.

DOI 10.15588/1607-3274-2022-3-10

УДК 004.85

МУЛЬТИМОВНИЙ КЛАСИФІКАТОР ТЕКСТУ З ВИКОРИСТАННЯМ ПРЕДТРЕНОВАНОЇ МОДЕЛІ UNIVERSAL SENTENCE ENCODER

Орловський О. В. – аспірант, кафедра програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича, Чернівці, Україна.

Khalili, Sohrab – CEO, CreateITTogether LLC Company, Fullerton, Каліфорнія, США.

Остапов С .Е. – професор, завідувач кафедри програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича, Чернівці, Україна.

Газдюк К. П. – ассистент, кафедра програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича, Чернівці, Україна.

Шумиляк Л. М. – ассистент, кафедра програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича, Чернівці, Україна.

АНОТАЦІЯ

Актуальність. Онлайн-платформи продовжують сьогодні генерувати усе більші обсяги інформації. Автоматизація модерування контенту у таких платформах, у зв'язку з цим, залишається актуальною задачею. Особливої уваги потребують випадки, коли з різних причин, доступно лише невеликі обсяги даних для навчання класифікаторів. У таких випадках необхідно залучати попередньо навчені моделі, які використовували для навчання великі об'єми даних широкого діапазону. У цій роботі досліджено питання застосування попередньо навченої мультимовної моделі Universal Sentence Encoder (USE) як компоненту розробленого нами класифікатора, а також впливу різних параметрів на точність класифікації при навчанні на малому об'ємі даних (~ 0,05% обсягу повного набору).

Метод. Для вирішення поставленого завдання використовується відносно новий підхід до навчання, – за допомогою невеликого набору повідомлень. Оскільки текстові повідомлення усе ще домінують як спосіб передавання інформації, застосовується розроблений класифікатор, навчений на невеликому (~ 0,002 – 0,05% повного набору) обсязі даних.

Результати. Показано, що навіть при невеликій кількості прикладів для навчання (36 на клас) за рахунок використання ECE та оптимальної конфігурації в навчанні можна досягти високої точності класифікації за англійськими та російськими даними, що надзвичайно важливо, коли неможливо зібрати свій власний великий набір даних. Оцінено вплив підходу з використанням USE та набору різних конфігурацій гіперпараметрів на результат класифікатора текстових даних на прикладі наборів даних англійською та російською мовами.

Висновки. У ході експериментів показана значна ступінь актуальності правильного підбору гіперпараметрів. Зокрема, у цій роботі розглядалися розмір пакету, оптимізатор, кількість епох навчання та відсоток даних із набору, взятих для навчання класифікатора. У процесі експерименту була обрана оптимальна конфігурація гіперпараметрів, згідно з якою 86,46% точності класифікації за російськомовним набором даних і 91,13% за англомовним відповідно можна досягти за десять секунд навчання (на час навчання можуть істотно вплинути використовувані технічні засоби).

КЛЮЧОВІ СЛОВА: few shot learning, навчання при малій кількості даних, предтреновані моделі, USE, нейронні мережі, інтелектуальний аналіз даних, набір даних, класифікатор текстових даних.

УДК 004.85

МУЛЬТИЯЗЫЧНЫЙ КЛАССИФИКАТОР ТЕКСТА С ИСПОЛЬЗОВАНИЕМ ПРЕДВАРИТЕЛЬНО ТРЕНИРОВАННОЙ МОДЕЛИ UNIVERSAL SENTENCE ENCODER

Орловский А. В. – аспирант, кафедра программного обеспечения компьютерных систем Черновицкого национального унивеситета имени Юрия Федьковича, Украина.

Khalili Sohrab – CEO, CreateITTogether LLC Company, Fullerton, Калифорния, США.

Остапов С. Э. – профессор, заведующий кафедрой программного обеспечения компьютерных систем Черновицкого национального унивеситета имени Юрия Федьковича, Украина.

Газдюк Е. П. – ассистент, кафедра программного обеспечения компьютерных систем Черновицкого национального университета имени Юрия Федьковича, Украина.

Шумиляк Л. М. – ассистент, кафедра программного обеспечения компьютерных систем Черновицкого национального унивеситета имени Юрия Федьковича, Украина.

АННОТАЦИЯ

Актуальность. Онлайн-платформы продолжают сегодня генерировать все более возрастающие объемы информации. Задачи автоматизации модерирования контента ползователей в связи с этим остается актуальной задачей. Особого внимания заслуживают случаи, когда, по разным причинам, доступны очень небольшие объемы данных для обучения классификатора. Для достижения приемлемых результатов необходимо применять предварительно обученные модели, которые использовали большие объемы данных широкого диапазона для предварительного обучения. В данной работе исследуется вопрос применения предварительно обученной мультиязыковой модели Universal Sentence Encoder (USE) в качестве компонента разработанного нами классификатора, а также влияния различных параметров на точность классификации при обучении на малом объема данных (~ 0,05% набора данных).

Метод. Для решения поставленной задачи используется относительно новый подход к обучению – по небольшой выборке сообщений. Поскольку текстовые сообщения все еще доминируют как способ передачи информации, использование классификатора текстовых данных при обучении на небольшой выборке (~ 0,002–0,05% набора данных) сообщений.

DOI 10.15588/1607-3274-2022-3-10

Результаты. Показано, что обучение даже на небольшой выборке (36 на класс) с использованием USE и оптимальной конфигурации при обучении можно достичь высокой верности классификации англо- и русскоязычных текстовых сообщений. Выполнена оценка влияния разных наборов гиперпараметров на результаты классификации.

Выводы. В ходе экспериментов показана актуальность правильного подбора гиперпараметров: размер пакета, тип оптимизатора, количество эпох, размер обучающей выборки. При оптимальных значениях гиперпараметров достигнута вероятность распознавания англоязычных деструктивных сообщений в 91,13%, при этом обучение проводилось всего на протяжении 10 секунд (что, безусловно, зависит от параметров использованных технических средств).

КЛЮЧЕВЫЕ СЛОВА: few shot learning, обучение на малой выборке данных, предварительно обученные модели, USE, нейронные сети, интеллектуальный анализ данных, набор данных, классификатор текстовых данных.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Yann L. Deep learning / L. Yann, B. Yoshua, H. Geoffrey // Nature. – 2015. – Vol. 521(7553). – P. 436–444.
2. Extracting unstructured data from template generated web documents / [L. Ma, N. Goharian, A. Chowdhury et al.] // Information and knowledge management: Twelfth international conference, 2003: proceedings, 2003. – P. 512–515.
3. Orlovskyi O. Analysis of the text preprocessing methods influence on the destructive messages classifier / O. Orlovskyi, S. Ostapov // Advanced Information Systems. – 2020. – Vol. 4(3). – P. 104–108.
4. Few-Shot Text Classification with Triplet Networks, Data Augmentation, and Curriculum Learning [Electronic resource] – Access mode: https://arxiv.org/abs/2103.07552
5. A Neural Few-Shot Text Classification Reality Check [Electronic resource]. – Access mode: https://arxiv.org/abs/2101.12073
6. Few-Shot Text Generation with Pattern-Exploiting Training [Electronic resource]. – Access mode: https://arxiv.org/abs/2012.11926
7. Task-Aware Representation of Sentences for Generic Text Classification / [K. Halder, A. Akbik, J. Krapac et al.] // Computational Linguistics: 28th International Conference, December 2020: proceedings. – 2020. – P. 3202–3213.
8. Reddy T. Text classification for AI education [Electronic resource] / T. Reddy, R. Williams, C. Breazeal. – Access mode: https://robots.media.mit.edu/wp-content/uploads/sites/7/2021/01/Text_classifier.pdf
9. Universal-sentence-encoder-multilingual-large. 16 languages (Arabic, Chinese-simplified, Chinese-traditional, English, French, German, Italian, Japanese, Korean, Dutch, Polish, Portuguese, Spanish, Thai, Turkish, Russian) text encoder [Electronic resource]. – Access mode: https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3
10. Sriporn K. Analyzing Malaria Disease Using Effective Deep Learning Approach / K. Sriporn, C. F. Tsai, C. E. Tsai et al. // Diagnostics. – 2020. – No. 10. – P. 744–749.
11. Fake or real news dataset [Electronic resource]. – Access mode:https://github.com/lutzhamel/fake-news/blob/master/data/fake_or_real_news.csv.
12. Russian Language Toxic Comments. Small dataset with labeled comments from 2ch.hk and pikabu.ru [Electronic resource]. – Access mode: https://www.kaggle.com/blackmoon/russian-language-toxic-comments.
13. Yang Y. Multilingual Universal Sentence Encoder for Semantic Retrieval / Y. Yang, D. Cer, A. Ahmad et al. [Electronic resource] – Access mode: https://aclanthology.org/2020.acl-demos.12.pdf

DOI 10.15588/1607-3274-2022-3-10