UDC 004.855.5

V. Oliinyk, K. Osadcha

DATA AUGMENTATION WITH FOREIGN LANGUAGE CONTENT IN TEXT CLASSIFICATION USING MACHINE LEARNING.

Abstract: This paper addresses the problem of insufficient data for text classification tasks and methods of datasets reinforcement. We explore the influence of data augmentation on classification accuracy. Text dataset is complemented by machine translations of similar data from languages of other families and the advantages of this approach are demonstrated.

An experimental research is conducted for the sentiment analysis problem for hotel reviews data that shows the difference in accuracy of the model trained on original and translated data. Influence of dataset volume on resulting accuracy is studied and compared for both types of the datasets. It is shown that learning of the model graphs are similar and differs only with some minor static error.

Key words: Sentiment analysis, text classification, representative data, data processing, data augmentation, machine learning, MLP.

Introduction

Text classification is one of the important and relevant tasks of machine learning for which state-of-the-art models continue to be created and improved. However, improving classifier models may not always bring better results because often the decisive contribution to the accuracy of their work is made by the data. Weak formalization, errors and multiplicity of contexts are essential features of text data. Machine learning methods, in particular neural networks, do a good job of revealing hidden patterns and, therefore, occupy a dominant position for word processing tasks, but at the same time have their own data requirements.

Collection of texts for machine learning should contain a sufficient amount of vocabulary necessary for training the classifier in the current subject area. When the dataset is too small to cover many variations, it becomes necessary to increase the amount of training data that can be problematic in many tasks. There are a number of approaches for this: usage of generative models, data augmentation by adding data in different languages, data modification, etc.

In this article, we will consider a way of data augmentation by adding text reviews in different languages belonging to different families. They will be translated into the target language - Russian, using a Google translator.

Original and added text data undergoes the same type pre-processing. Processing includes clearing texts from special characters, reducing all words to lower case, stemming words, representing words in vector form.

The research was conducted on a typical task of text classification - sentiment analysis.

Its purpose is to determine whether text data belongs to a positive or negative set. The simplicity of the task allows one to avoid the influence of task complexity on the classification results and adequately evaluate the impact of data.

Analysis of recent research and publications

Cross-Lingual approaches are rather popular in text and sentiment classification [1,2,3]. For instance, in [1] authors use LSTM model in the task of sentiment classification based on reviews. Similar to our experiment, online Google translator was used to put text data to the target language - Chinese. Results of this research demonstrated how different distances in sentences effect accuracy of the model. In our work text translations into the target language are used for increasing initial dataset and main goal is to analyze rationality of this method of data augmentation.

There are also works addressing problem of increasing dataset quality, and its length in particular [4, 5, 6]. In paper [4] method of data augmentation is studied. The author offers to increase the amount of data by creating new sentences. This is achieved by modifying the context in existing text data. As a result, this article proved that their approach works better than the synonym-based augmentation. Our study uses a completely different approach to text data augmentation.

Neural network model

Solving text classification problem requires making mapping between textual data with low formalization with corresponding set of discrete classes. Machine learning techniques have already become a classical tool for this task and neural networks demonstrate state-of-the-art results. Any neural model essentially requires high quality dataset to show adequate results. Training collection of texts should be at least sufficiently representative to achieve this goal.

In practice, the quality of the final models depends much more on the quality of the prepared data than on the model architecture and its optimization. Therefore, simple model of feed forward neural networks – a classic model of multilayer perceptron (MLP) was chosen [7]. The typical architecture of MLP used in this article is shown in Fig. 1, and its specific parameters (the number of neurons and layers) were

 ISSN 1560-8956

selected experimentally for leveling the influence of architecture on the influence of data augmentation in the sample.

This network was trained using ADAM algorithm - classic modification of stochastic gradient descent. At the stage of network training, pairs (batches) of input data were provided to the network sequentially. Such a pair are the values corresponding to the review from the "word bag", and the corresponding previously known parameters of the output vector, which in this task are logical statements of the input vector belonging to a positive or negative class.

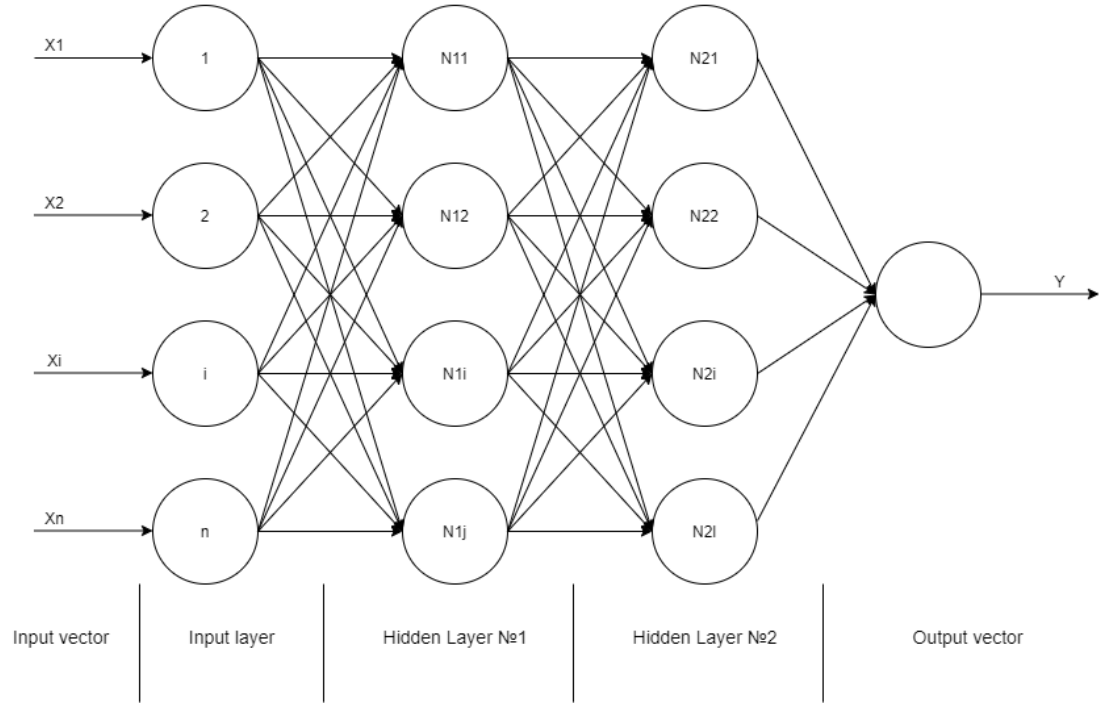

Figure. 1. Multilayer perceptron architecture for the experiment

As a result of applying the trained classifier to an arbitrary text (review), we obtain sentiment class (emotion) of processed textual content.

Data pre-processing

At the first stage, publications relevant to the current task were collected and processed. The data are reviews of vacationers about hotels and their rating on a scale from 1 to 10. They were taken from sources such as hotels.turizm.ru, ru.hotels.com, tripadvisor.com. In total, 10,000 reviews were gathered: 4,000 reviews originally written in Russian, others in different languages. Most of the reviews are in English, there are also French, German, Ukrainian, Japanese and others. Summary there were 5,000 positive and as many negative reviews.

The ratings of the reviews were not evenly distributed - therefore, a decision was made to replace multiclass classification with binary so that this does not affect the result. The expected result consists of a set of 0 and 1, where 0 means that the response is negative; 1 - positive. Actual rating of the review was simple rounded to desired values as follows: values from 1 to 5 was replaced with 0, from 6 to 10 - with 1.

The minimum length of the text is one word, the maximum is 120. Figure 2 shows a histogram of the distribution of the number of words in the texts.

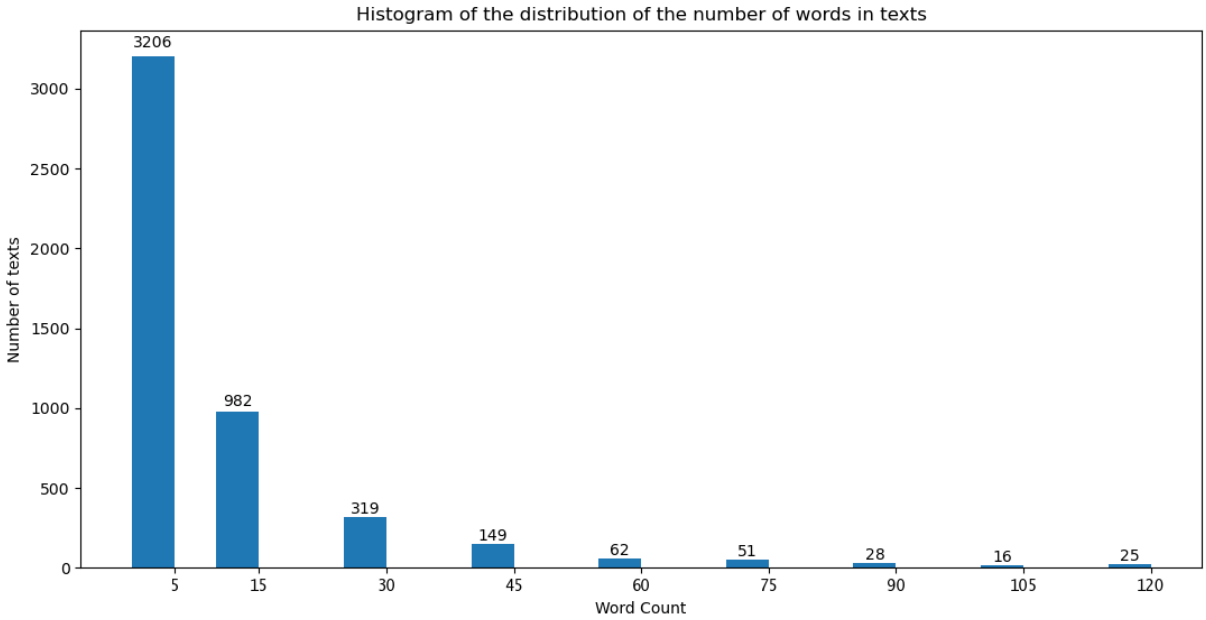

Figure. 2. Histogram of the distribution of the number of words in texts

Analysis of words distribution shows that sample texts in general have small number of words; therefore, we should rely on them in conclusions. Majority of sentences (64%) have up to five words and another 20 percent consist of 5-15 words, the remaining groups are less significant. The length of the input data vector 45 was selected according to the results of the distribution of the number of words in sentences. It allows to use over 93% of gathered samples without data loss. Those text samples whose length exceeds 45 are truncated to this value.

The purpose of pre-processing is to present data in a normalized form that will acceptable to selected models. It also improves achieved results due to initial data generalization. Text pre-processing consists of two stages: cleaning and conversion to vector form.

At the first stage, all the data went through such processing:

1. All words have been converted to lowercase;
2. All punctuation marks were removed (for all data);

 ISSN 1560-8956

3. Control characters removed (for all data);
4. Spelling was checked (only for reviews in Russian);
5. Stemming of words was performed (for all data).

All data was translated to the target Russian language. Python library that implemented the Google Translate API was used for this purpose.

Next, the texts were converted to vector form using the “Bag of words” approach. The statistical measure “tf – idf” was used to determine the weight of each word improving the quality of classification.

Environment and software implementation

The software solution was implemented in Python 3.6.8. Python's built-in capabilities were used to download pages with reviews from travel sites. LXML library helped with HTML page processing. Reviews have been translated into target language thanks to googletrans library version 2.4.0.

The sklearn library was used to prepare training and test dataset that was stored in MySQL database. And Keras and Tensorflow were our frameworks for training neural networks.

Stemming, spelling check and text normalization were made for data pre-processing. The nltk.stem package was used to remove morphological affixes from words and PyEnchant library - to check the spelling of the texts. Preprocessed texts were put to lowercase, cleared from special characters and converted to ‘bag of words’ using Keras.

For data visualization, in particular, for plotting graphs, the matplotlib package was used.

Experimental research

Experiments planning

The main idea of the experiment is to compare the results of the neural network before and after adding translated reviews to the main dataset. In addition, we will study dependency of model’s accuracy on dataset volume with original and translated data. Thus, it will be possible to verify the expediency of this approach.

The size of the dictionary, as well as the words in it, vary depending on the size of the dataset. In the experiment, three types of datasets will be tested, which consist of:

1. Data that was originally written in the target language (4,000 reviews).
2. Data that was originally written in the target language (10,000 reviews).

3. Data partially written in the target language (4000 reviews) and appended with data translated into the target language (6000 reviews).

For all experiments, we use sigmoid activation function in hidden layers. Several network configurations are used with different numbers of hidden layers and neurons in it to minimize the impact of the model to results. The number of training samples is 70% of the total quantity in all experiments.

Experimental results

Experiment No. 1 was carried out on data that were originally written in Russian. Dataset contains 4,000 reviews; the number of positive reviews is 2,000.

Table 1 shows the model parameters and resulting accuracy on a test set in several setups.

Table 1

The result of the MLP model using data that was originally written in the target language (4,000 reviews)

№	Dictionary Size	Number of hidden layers	The number of neurons in the hidden layers	Epochs	Test Accuracy, %
1	20000	2	160, 80	15	85.5
2	12000	2	80, 20	5	85.37
3	10000	2	80, 20	8	85
4	12000	2	60, 20	5	83.68
5	6000	2	60, 20	5	83.43
6	12000	2	60, 20	10	84.12
7	17000	3	120,60,20	15	82.56

Table 2

The result of the MLP model using data that was originally written in the target language (10,000 reviews)

№	Dictionary Size	Number of hidden layers	The number of neurons in the hidden layers	Epochs	Test Accuracy, %
1	20000	2	160, 80	15	91.83
2	12000	2	80, 20	5	92.68
3	10000	2	80, 20	8	90.92
4	12000	2	60, 20	5	90.45
5	6000	2	60, 20	5	90.15
6	12000	2	60, 20	10	88.57
7	17000	3	120, 60, 20	15	87.25

 ISSN 1560-8956

Based on the accuracy of the neural network in the test sample from the first table, the network is achieving acceptable results. However, the loss rate is 14.5% at best that is still quite large.

Second experiment was conducted on data that contains 10,000 reviews that were originally written in Russian. The number of positive reviews is 5000.

As expected, the results of the neural network from table 2 improved compared to the 1st experiment. The loss rate decreased to 7.32% at best. In this experiment, losses in the accuracy of the network are minimal.

Table 3

The result of the MLP model using data that was translated to the target language

№	Dictionary Size	Number of hidden layers	The number of neurons in the hidden layers	Epochs	Test Accuracy, %
1	20000	2	160, 80	15	90.07
2	12000	2	80, 20	5	91.22
3	10000	2	80, 20	8	90.65
4	12000	2	60, 20	5	89.12
5	6000	2	60, 20	5	88.69
6	12000	2	60, 20	10	87.16
7	17000	3	120,60,20	15	85.98

This experiment is necessary to compare the results of a neural network on data that were written in Russian and those that were augmented with translations.

Third experiment was conducted on data that contains 10,000 reviews. Of these, 4,000 are written in Russian, the rest were translated into the target language. The number of positive reviews is 5000.

The results of the neural network in experiment 3 are better than in 1 and slightly worse than in 2. The network results in the third experiment increased by 5.7% compared with the first experiment. This means that adding translated reviews to the main dataset leads to better performance.

In comparison with experiment 2, the results worsened slightly, the difference in the best results is 1.46%, which can be treated as a good indicator.

Next, we studied the dependency of neural network on a different amount of training data and the impact of data augmentation. The blue line in Fig. 3 shows the accuracy of the model on the data taken from experiment 2. And orange line shows the accuracy based on the data taken from experiment 3.

Not hard to notice that the orange line is always next to the blue. In other words, network results based on data written in the target language have approximate indicators with network results whose dataset has artificially increased. Which confirms the advisability of using this method of data augmentation.

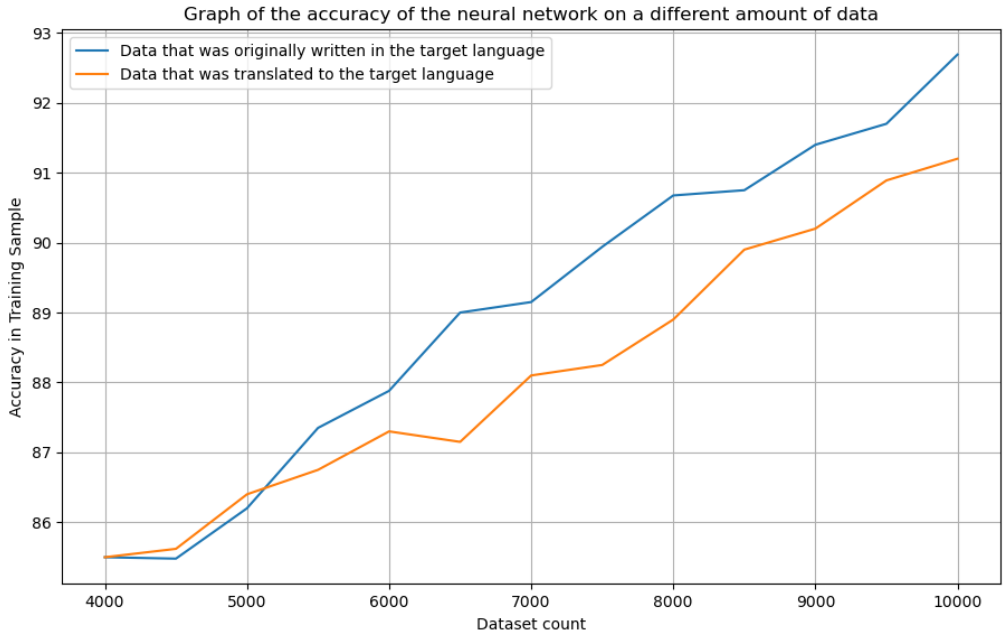

Figure. 3. Graph of the accuracy of the neural network on a different amount of data

The results on Fig. 3 shows that accuracy is rising almost linearly from 85.5% to 92% that clearly indicates that initial dataset was insufficient for the task and neural model. After that, we have typical saturation but the accuracy is still rising. Accuracy loss is between 0,7-1,8% and is rather constant in all range. It can be explained with additional static error added by translation. Still the behavior of the accuracy graph does not change and we manage to get much better results compared to limited dataset experiment. Such error may prevent using this method for state-of-the-art accuracy models but is acceptable in many real life applications and proves the concept of using general-purpose translations to augment text datasets.

Conclusions

In this paper, we demonstrated the benefits of augmenting text datasets with translated data. Sentiment analysis problem in the hotel business area was taken as an example and simple MLP neural network model was used as a classifier.

It has been experimentally proven that the quality of the classifier improves when source sample of text data is increased by data that is translated to the target

 ISSN 1560-8956

language. Moreover, gradual data augmentation with translated data showed similar to original language progression of classifier accuracy (with small static error). This fact allows making a hypothesis that text translations generated by state-of-the-art models, i.e. Google engine, can be widely used in text classification tasks based on ‘bag of words’ for dataset augmentation when dataset is limited.

Future work will be aimed to prove this concept as well as the fact that unknown context and errors of translations etc. can be minimized and neglected.

REFERENCES

1. Zhou X. Attention-based LSTM network for cross-lingual sentiment classification / Zhou X., Wan X., Xiao J. //Proceedings of the 2016 conference on empirical methods in natural language processing. – 2016. – P. 247-256.

2. Wan X. Co-training for cross-lingual sentiment classification // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1. – Association for Computational Linguistics, 2009. – P. 235-243.

3. Chen X. et al. Adversarial deep averaging networks for cross-lingual sentiment classification //Transactions of the Association for Computational Linguistics. – 2018. – Т. 6. – С. 557-570.

4. Kobayashi S. Contextual augmentation: Data augmentation by words with paradigmatic relations //arXiv preprint arXiv:1805.06201. – 2018.

5. Wei J. W., Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks //arXiv preprint arXiv:1901.11196. – 2019.

6. Романенко А.Ю. Узагальнена модель розпізнавання голосових команд / А.Ю. Романенко, В.В. Олійник // Міжвідомчий науково-технічний збірник «Адаптивні Системи Автоматичного Управління», К:Політехніка – 2017. – Т.1, №30 – С. 130-139.

7. Ямпольський Л.С. Нейротехнології та нейрокомп’ютерні ситеми / Л.С. Ямпольський, О.І. Лісовиченко, В.В. Олійник // Д К.: «Дорадо-Друк» – 2016, 571 с.