UDC 004.94 **doi: 10.26906/SUNZ.2020.3.094**

S. Olizarenko[1], V. Argunov[2]

[1] Kharkiv National University of Radio Electronics University, Kharkiv, Ukraine
[2] HIPSTO, Kharkiv, Ukraine

ON POSSIBILITIES OF MULTILINGUAL BERT MODEL FOR DETERMINING SEMANTIC SIMILARITIES OF THE NEWS CONTENT

Abstract. The results of implementation of modern achievements in the field of Natural Language Processing field based on the methods and models of Deep Learning technologies into the HIPSTO's system management of content (HIPSTO Publishing, AI Technology, Digital Media, Mobile Apps) are discussed and analyzed. In particular, the possibilities and ways of applying the multilingual BERT model to handle the problem of semantic likeness of news content have been investigated. An efficient method is proposed to define the semantic similarities of the multilingual news content in HIPSTO aggregated news feeds on the basis of the sentence embeddings using the first task of the pre-trained multilingual BERT model within the HIPSTO system of content management. The results of the research highlight the effectiveness and promise of this technology within the HIPSTO project. Below the data of its first implementation in HIPSTO are substantiated scientifically and experimentally.

Keywords: Natural Language Processing; BERT; semantic similarities; news content, Deep Learning.

Introduction

This study is a part of development of AI driven (mobile) information curation platform HIPSTO (hobbies, heroes, interests, people, personalities, subjects, topics, objects and organizations). The HIPSTO content management system consists of the three major blocks (Fig. 1):

- content scraping;
- content moderation and information discovery;
- content delivery.

The goal of this research is to investigate the best approach for building a related news and article detection service to accomplish the information discovery layer of HIPSTO.

The verification system should provide, among others, the solution for the problem of semantic likeness detection problem in analyzing the news multilingual content using the latest advances in Natural Language Processing (NLP).

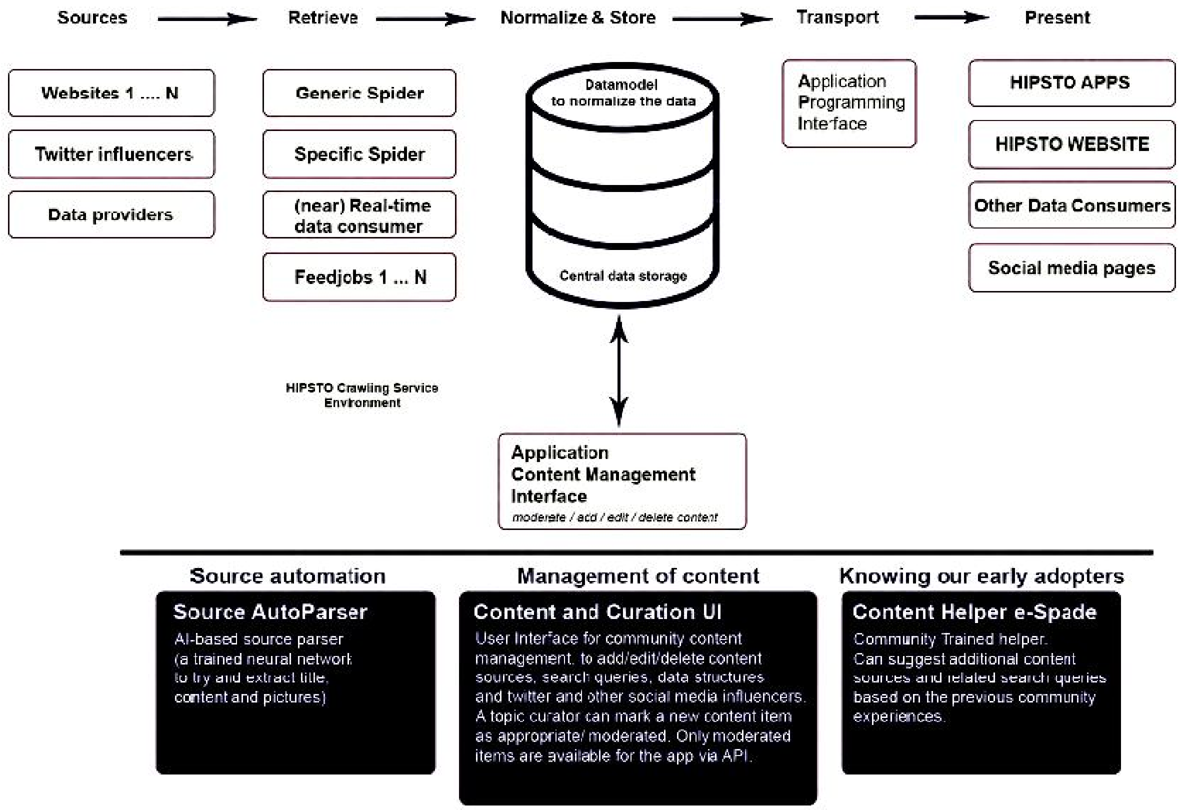

Sources
Retrieve
Normalize & Store
Transport
Present
Websites 1 .... N
Twitter influencers
Data providers
Generic Spider
Specific Spider
(near) Real-time data consumer
Feedjobs 1 ... N
Datamodel to normalize the data
Central data storage
Application Programming Inteface
HIPSTO APPS
HIPSTO WEBSITE
Other Data Consumers
Social media pages
HIPSTO Crawling Service Environment
Application Content Management Interface
moderate / add / edit / delete content
Source automation
Management of content
Knowing our early adopters
Source AutoParser
AI-based source parser (a trained neural network to try and extract title, content and pictures)
Content and Curation UI
User Interface for community content management, to add/edit/delete content sources, search queries, data structures and twitter and other social media influencers. A topic curator can mark a new content item as appropriate/ moderated. Only moderated items are available for the app via API.
Content Helper e-Spade
Community Trained helper. Can suggest additional content sources and related search queries based on the previous community experiences.

Fig. 1. The place of the HIPSTO's system management of content in the general HIPSTO technology

Analysis publications. A promising approach for finding semantic similarity in the analysis of multilingual news content is provided in the pre-trained multilingual model at the level of sentence embeddings.

Currently, almost all of these models are implemented using Deep Learning techniques. Let us consider the best-known ones. For example, the paper (Yang et al., 2019) [1] reviews the multilingual universal sentence encoder for semantic retrieval in 16 languages within the group of models embedding sentences of universal sentence coder (USE) (Cer et al., 2018) [2]. The models are implementations of the CNN (Kim, 2014) [3] and Transformer architectures (Vaswani et al., 2017) [4]. It is proposed (Lee, 2019) [5] to implement a multilingual similarity search using a bi-directional LSTM with preparatory training on the basis of LASER (Language-Agnostic SEntence Representations) [6]. Recently (2018) the development of a set of BERT (Bidirectional Encoder Representations from Transformers) models [7] has come as the main breakthrough in the of NLP field. Its possibilities are currently being studied by the scientific community, they have been used practically and as a boost to develop new NLP models in 2019 (see XLNet, RoBERTa (Robustly Optimized BERT Pretraining Approach), ERNIE (Enhanced Representation through kNowledge IntEgration) and others). BERT is a new method of pre-training language representations that obtains high quality results on a wide range of Natural Language Processing (NLP) tasks [8]. The BERT complex includes the multilingual "BERT-Base, Multilingual Cased" model, which currently includes the largest number of languages among the multilingual models of this class. The paper (Lee, 2019) [9] proposes the task of identifying similarities among news articles using the other task of the pre-trained multilingual BERT model (next sentence prediction (NSP)).

This paper researches a more effective way of defining the semantic similarities of news multilingual content based on sentence embeddings using the first task of the pre-trained multilingual BERT model (masked language model (MLM)).

The Generalized Scheme of Research

The generalized scheme of research includes the realization of the following main steps:

1) experiment execution:

- implementation of the NLP model for creating contextual word embeddings with the subsequent formation of sentence embeddings;

- determination of the degree of similarity of sentence embeddings using the selected metric;

2) reporting and analysis of the results of determination of the degree of similarity of sentence embeddings;

3) identification of the areas of difficulty and the ways to address them when defining the semantic similarity of the multilingual news content based on sentence embeddings using the appropriate BERT model.

Experiment Technique

The Implementation of the NLP Model. The multilingual NLP model can be implemented using the “BERT-Base, Multilingual Cased" model, which includes 104 languages, 12 layers, 768 hidden layers and 177M of parameters. Here we consider the multilingual NLP model for the determining of the semantic similarity of the generalized, non-thematic multilingual news content, omitting the issues of constructing a training sample for fine-tuning of the model on a specific topic and directly fine-tuning of the BERT model. The pre-trained multilingual BERT model is used here only to create contextual word embeddings. In this connection, the research investigates the possibilities of using output vectors of individual layers and the combination of the layers within the BERT model to determine the best embeddings. Firstly, this is because the different layers of BERT encode quite different types of information that can be used accordingly in a variety of tasks of NLP. Secondly, rather high values of the quality indicators of the model’s functioning can be obtained on different layers without fine-tuning (see the work (Alammar, 2019) [10]). Here the results of the formation of word vectors on the following layers of the model are analyzed:

1) 7 Embedding-Norm False (First Layer);

2) 79, 87, 95, 103 Encoder-12-FeedForward-Norm False (Sum Last Four Hidden);

3) 79, 87, 95, 103 Encoder-12-FeedForward-Norm False (Concat Last Four Hidden);

4) 103 Encoder-12-FeedForward-Norm False (Last Layer).

The sentence embeddings were created using two approaches:

1) each element of sentence embeddings is taken as the average of the corresponding elements of all word embeddings of news content (mean);

2) each element of sentence embeddings is transformed using MaskedGlobalMaxPool1D [11] layer, expanding the basic multilingual BERT model (pool).

The Degree of Similarity. Since the NLP deep neural network model presents multilingual content as object vectors, the degree of similarity of the content is determined by measuring the distance between these objects. There are several metrics to determine the degree of similarity, such as the Jaccard Similarity, K-means, Cosine Similarity, Jensen-Shannon distance, etc.

The Cosine Similarity was used in this research for defining the similarity of two contents, then the results were converted into angular distance as in this paper [2]

$$(u,v)=\left(1-\arccos\left(\frac{uv}{|u|\cdot|v|}\right)\right)\Big/\pi . \qquad (1)$$

where u,v are the sentence embeddings obtained from the outputs of the NLP deep neural network model layers.

Besides, according to the recommendations in [12], a relative rather than an absolute assessment of the results of the Cosine Similarity definition is made

$$\begin{aligned}&IF\ \ sim(u,v)>sim(u,c)\\&THEN\ \ u\ \ is\ \ more\ \ to\ \ v\ \ THAN\ \ c.\end{aligned} \qquad (2)$$

The results on similarity are visualized within [13]. It is based on constructing a scatter diagram and a dimension chart of sentence embeddings. The degree of similarity decreases when the scatter and the dimensions of vector parts on diagrams increase.

Analysis and results

The possibilities offered by the multilingual BERT model for determining the semantic similarity of news content are analyzed using the test data:

1) the marked-up SentEval set (STSbencmark) for estimating the correspondence results on the NLP deep neural network model to determining semantic similarity [14];

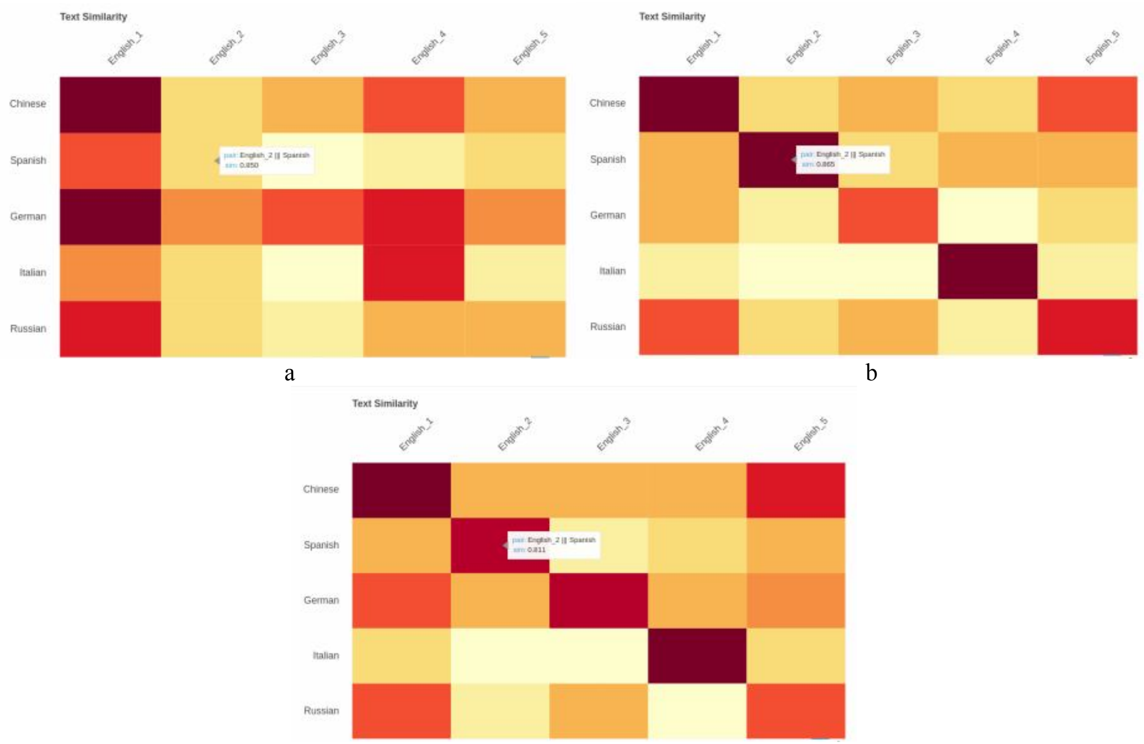
2) a data set including pairs of completely identical sentences, but in different languages ("English - Chinese", "English - Spanish", "English - German", "English - Italian", "English - Russian") for evaluating the results of the NLP neural network model for determining inter-language semantic similarity.

Table 1 shows the averaged data for the multilingual BERT model in various configurations on the SentEval set.

The diagrams Fig. 2, show the results of the Cosine Similarity obtained with the NLP deep neural network model on solving the problems of inter-language semantic similarity for the configuration of the First Layer, Concat Last Four Hidden, Last Layer model (the method of generation of sentence embeddings for all layers - pool).

Table 1 – **The averaged data for the multilingual BERT model in various configurations on the SentEval set**

№	Adapted Degree of Similarity	Model configuration		The similarity rating value (%)
		layer	way to form sentence embeddings	
1.	Full	First Layer	pool	90.6
2.		Sum Last Four Hidden	pool	88.53
3.		Concat Last Four Hidden	pool	**91.72**
4.		Concat Last Four Hidden	mean	90.16
5.		Last Layer	pool	82.6
6.		Last Layer	mean	52.6
7.	Partial	First Layer	pool	87.1
8.		Sum Last Four Hidden	pool	81.02
9.		Concat Last Four Hidden	pool	**86.97**
10.		Concat Last Four Hidden	mean	86.6
11.		Last Layer	pool	72.8
12.		Last Layer	mean	64.09
13.	Absent	First Layer	pool	84.77
14.		Sum Last Four Hidden	pool	80.54
15.		Concat Last Four Hidden	pool	**86.7**
16.		Concat Last Four Hidden	mean	86.3
17.		Last Layer	pool	69.1
18.		Last Layer	mean	52.6



a

b

c

Fig. 2. The results of the Cosine Similarity

Fig. 3–5 The examples of scattering diagrams and dimension values of vector elements in solving the problems of the inter-language semantic similarity for the configuration of the First Layer, Concat Last Four Hidden, Last Layer model (a method of generation sentence embeddings for all layers - pool) for a set of sentences “English - Chinese".

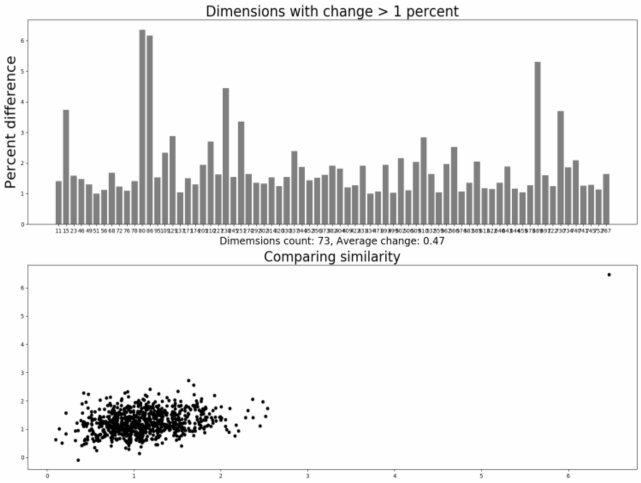


Fig. 3. Scatter diagrams and dimension values for the model configuration with First Layer

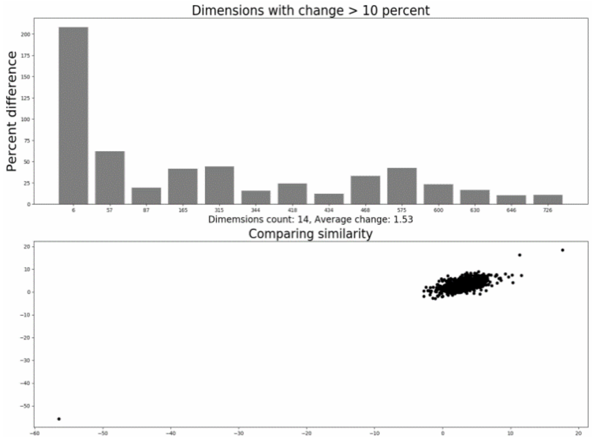


Fig. 4. Scatter diagrams and dimension values for the model configuration with Concat Last Four Hidden

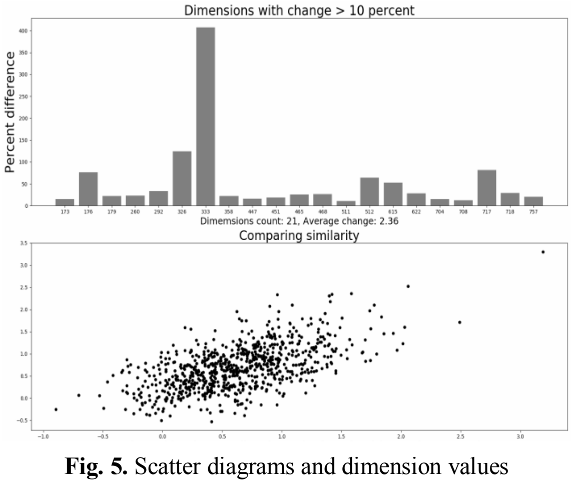


Fig. 5. Scatter diagrams and dimension values for the model configuration with Last Layer

The results obtained using the BERT multilingual model for determining the semantic similarity in the news content analysis show that the model with the Concat Last Four Hidden configuration employing MaskedGlobalMaxPool1D layer has the highest similarity rating values for different data sets.

However, it is important that the lower bound for the similarity score also rises significantly, and the interval between the upper and lower bounds for the similarity score in the model with the Concat Last Four Hidden configuration is one of the smallest.

Detection of the Problem Points and the Ways of Troubleshooting Them

The multilingual BERT model used to determine the semantic identity of news content in HIPSTO’s system management of content involves the following problems of a classic model restricting the BERT potentiality:

1) limitation of 512 elements on the input sequence of the classic multilingual model which calls for either the modification of certain layers of the model to increase the length of the input sequence and perform subsequent fine-tuning or splitting the input content longer than 512 elements into composite messages with their subsequent processing;

2) requirement of significant computing resources not only at the stage of training and setting up the model, but also during the process of its practical use due to the large number of BERT parameters which require scaling the BERT model and its subsequent effective use for both mobile devices and desktop computers and servers.

How can the use of the model limited to 512 elements determine the length of the sequence of tokens of the input content imposing even more restrictions on the length of the input content sequence. It is obviously necessary to split the input content of more than 512 elements into parts, but the following should be considered:

1) the content should be divided considering crossing of received sentence embeddings to save the contextual dependence of words in sentences at the vector borders as much as possible;

2) when combining vectors into a single sentence embedding, it is necessary to control double-counting of boundary words;

3) technologically, the token dictionary in the Bert model is built so that some words are divided into certain parts, which can appear on the boundaries of vectors where it is difficult to account for them correctly.

Currently, the tasks of pre-processing news content to represent input sequences of information longer than 512 characters are implemented within the HIPSTO technology to effectively process them in the BERT model. To solve the problem of scaling (compression) models, several concepts and approaches are currently used to optimize memory usage and increase the speed of the model.

For example, such approaches include the quantization of models during or after training, the removal of weight compounds (cutting off by the value of weight, neurons or weight matrices), etc. The studies show that small transformers and BERT sensors can be quantized for mobile devices using TensorFlow Lite without significant loss of accuracy however there is practically no acceleration of output. For stationary systems, post-training quantization using TensorFlow Lite slows down the BERT output by more than 5 times [15]. The HIPSTO technology for BERT scaling explores the approaches based on weight cutting and neuron cutting.

Conclusions

During the development of the AI driven (mobile) information curation platform HIPSTO the possibilities of using the multilingual model BERT for determining the semantic similarities of news content were considered.

The results obtained show that the HIPSTO technology can be supplemented with the topology of the BERT model forming word embeddings due to the concatenation of the last four layers and the subsequent formation of sentence embeddings using the MaskedGlobalMaxPool1D layer.

HIPSTO has now implemented preliminary stages of processing news content to effectively address the problem of limiting the input sequence of the modified BERT neural network model, the BERT is practiced based on weight trimming and pruning of neurons. Preliminary results show the efficiency of these approaches both for mobile devices and for fixed systems.

References

1. Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil Multilingual Universal Sentence Encoder for Sematic Retrieval. arXiv:1907.04307v1 [cs.CL] 9 Jul 2019.
2. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In Proceedings of the 2018 Conf. on Empirical Methods in Natural Language Proc.: System Demonstrations, pages 169–174.
3. Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751.
4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of NIPS, pages 6000–6010.
5. Ceshine Lee Multilingual Similarity Search Using Pretrained Bidirectional LSTM Encoder. Evaluating LASER (Language-Agnostic SEntence Representations)/ https://medium.com/the-artificial-impostor/multilingual-similarity-search-using-pretrained-bidirectional-lstm-encoder-e34fac5958b0.
6. Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library. POSTED ON JAN 22, 2019 TO AI RESEARCH/ https://engineering.fb.com/ai-research/laser-multilingual-sentence-embeddings/.
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL] 24 May 2019.
8. Join GitHub today, available at: https://github.com/google-research/bert.
9. Ceshine Lee News Topic Similarity Measure using Pretrained BERT Model. Utilizing Next Sentence Predictions, available at: https://medium.com/the-artificial-impostor/news-topic-similarity-measure-using-pretrained-bert-model-1dbfe6a66f1d.
10. Jay Alammar The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), available at: http://jalammar.github.io/illustrated-bert.
11. available at: https://github.com/ CyberZHG/keras-bert.
12. available at: https://bert-as-service.readthedocs.io.
13. Using NLP to Automate Customer Support, Part Two, available at: https://blog.floydhub.com/automate-customer-support-part-two.
14. available at: https://github.com/facebookresearch/SentEval.
15. Sam Sucik Compressing BERT for faster prediction, available at: https://blog.rasa.com/compressing-bert-for-faster-prediction-2.

Received (надійшла) 19.05.2020
Accepted for publication (прийнята до друку) 22.07.2020

Дослідження можливостей багатомовної моделі BERT для визначення семантичної подібності новинного контенту

С. А. Олізаренко, В. В. Аргунов

Анотація. Мета статті – провести дослідження більш ефективного способу визначення семантичної подібності багатомовного вмісту новин на основі вбудовування речень за допомогою першого завдання попередньо навченої багатомовної моделі BERT. **Результати.** В роботі представлені результати впровадження сучасних досягнень в області обробки природної мови на основі методів і моделей технологій глибокого навчання в систему управління новинним контентом HIPSTO (HIPSTO Publishing, AI Technology, Digital Media, Mobile Apps). Досліджено можливості та способи застосування багатомовної моделі BERT для вирішення проблеми семантичної подібності новинного контенту. Зокрема, пропонується ефективний метод визначення семантичної подібності багатомовного новинного контенту в агрегованих новинних стрічках HIPSTO. Даний підхід заснований на використанні в системі управління новинним контентом HIPSTO векторних уявлень речень з використанням першого завдання попередньо навченої багатомовної моделі BERT. Результати досліджень, представлених в роботі, підкреслюють ефективність розвитку даної технології в рамках проекту HIPSTO. Подальший викладення матеріалу надає науково-експериментальне обґрунтування отриманих результатів, що мають вже практичну реалізацію в перших версіях HIPSTO.

Ключові слова: обробка природної мови; BERT; семантична подібність; новинний контент, глибоке навчання.