Intelligent information systems

UDC 621.396.6 **doi:** https://doi.org/10.20998/2522-9052.2021.2.18

Serhii Olizarenko, Viacheslav Radchenko

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

METHOD FOR DETERMINING THE SEMANTIC SIMILARITY OF ARBITRARY LENGTH TEXTS USING THE TRANSFORMERS MODELS

Abstract. The paper considers the results of a method development for determining the semantic similarity of arbitrary length texts based on their vector representations. These vector representations are obtained via multilingual Transformers model usage, and direct problem of determining semantic similarity of arbitrary length texts is considered as the text sequence pairs classification problem using Transformers model. Comparative analysis of the most optimal Transformers model for solving such class of problems was performed. Considered in this case main stages of the method are: Transformers model fine-tuning stage in the framework of pretrained model second problem (sentence prediction), also selection and implementation stage of the summarizing method for text sequence more than 512 (1024) tokens long to solve the problem of determining the semantic similarity for arbitrary length texts.

Keywords: text; arbitrary length; semantic similarity; vector representation; Transformer model; fine-tuning.

Introduction

One promising approach for solving the problem of finding semantic similarity in text analysis is an approach based on using pretrained Transformers models in the Deep Learning methodology framework [1]. In this paper, following Transformers-based models are researched: the BERT model (Pre-training of Deep Bidirectional Transformers for Language Understanding) for 104 languages [2]; DistilBERT model for 104 languages, which is a lightweight version of BERT that works faster by 60% and retains more than 95% BERT characteristics, measured with GLUE test (General Language Understanding Evaluation) [3]; XML model for 100 languages [4]. One of the limitations of all presented models is the tokenized text input sequence length – no more than 512 (1024) tokens.

The most common approach for solving semantic similarity determination problem of arbitrary length texts using Transformers model is an approach based on sliding window arbitrary length vector representation with subsequent formation and similarity degree determination of the received compressed vector representations. This study proposes the problem solution implementation of multilingual texts semantic similarity determination using pretrained multilingual Transformers model second problem (sentence prediction problem).

The main approaches for solving directly the problem of overcoming the input sequence length are approaches based on truncation methods (selection of the first or last sequence fragments 512 (1024) tokens long, combining first and last sequence fragment, but no more than 512 (1024) tokens in total) or hierarchical methods (for example, with combining latent states of the all fragments from the sequence) [5]. However, application of given approaches can lead to the contextual dependency loss of most significant words (phrases and sentences, respectively) in text sequences, which in its own turn may drastically affect semantic similarity determination quality of analyzed texts. Thus, there is a Transformers model application problem for the texts longer than 512 (1024) tokens with provisioning of contextual dependency maximum preservation for most significant words in text sequences with the goal to effectively determine semantic similarity of arbitrary length texts. Given basic summarizing approach, extractive and abstract generalization approaches are analyzed in the paper.

Literature analysis. In this section papers are considered which have various research results presented regarding Transformers-based models' usage for solving semantic similarity determination problem of texts. So, in the papers (S. Olizarenko, V. Argunov, 2019) news content semantic similarity determination possibilities are researched with the usage of pretrained multilingual BERT model first problem (word masking problem) [1]. In the paper (Yang et.al., 2019) [6] multilingual universal sentence encoder for semantic retrieval is considered for 16 languages in the sentence embedding model family of universal sentence encoder (USE) (Cer et al., 2018) [7]. Given models represent CNN architecture implementation (Kim, 2014) and Transformer (Vaswani et al., 2017) [8, 9]. In paper (Lee, 2019) multilingual similarity search implementation is proposed with LTSM bidirectional coder usage and preliminary preparation based on LASER (Language-Agnostic SEntence Representations) [10, 11]. In paper (Chi Sun et al., 2019) experiment results are provided regarding various BERT fine-tuning methods for text classification problems, including ones in the text semantic similarity determination context [5]. The publication (Nils Reimers et al., 2019) presents Sentence-BERT (SBERT) model, which is a modification of pretrained BERT network, that uses conjoined and triplet network structures in order to obtain semantically meaningful vector sentence representations to be compared using similarity cosine [12]. In the paper (Manish Patel, 2019) semantic-oriented search system is developed, which uses BERT inclosures and additional neural network for similarity estimate finding with subsequent document ranking, in order from most meaningful to least meaningful document [13]. In the paper (Han Xiao, 2019) search system is developed, which uses BERT inclosures and cosine similarity to compute query and document similarity estimate [14]. At the same time, efficient processing questions of arbitrary length texts in these works were not considered.

Purpose of paper. Development of a method, essence of which is the preliminary generalization (automatic summarization) of the arbitrary length compared texts based on machine learning method and subsequent direct determination of their semantic similarity within the framework of the text sequence pairs classification problem solving (the predicting sentences problem) using a pretrained and fine-tuned Transformers model.

Main part

In this paper the semantic similarity determination problem of arbitrary length texts is considered as the problem of text sequence pairs classification problem. In accordance with given problem statement, method development procedure for semantic similarity determination of arbitrary length texts using Transformers models is presented, in the form of the following main stages:

1) Software module architecture definition for semantic similarity determination of arbitrary length texts;

2) Basic summarizing method of text sequence more than 512 (1024) tokens long for semantic similarity determination of arbitrary length texts problem solving;

3) Application possibilities analysis of various Transformers model types (BERT, DistilBERT and XML) for texts semantic similarity determination.

4) Transformers model fine-tuning (BERT, DistilBERT and XML) to solve the text sequence pairs classification problem;

5) Tuning results analysis and basic Transformers model choice for semantic similarity determination problem solution of arbitrary length texts based on their summarization results.

Schematically semantic similarity determination of arbitrary length texts software module is presented in Fig. 1. In accordance to given schema, software module comprises two main blocks:

1) Summarization block of text sequence no more than 254 tokens long (for considered in the paper BERT, DistilBERT and XML models);

2) Fine-tuned multilingual Transformer model for sequence pairs classification problems.

Generalization is the task of reducing the text to a shorter version, decreasing the size of the original text while retaining the key information elements and content meaning. The main summarization problem in this study was to generalize a text sequence of no more than 254 tokens long (for the BERT, DistilBERT and XML models considered in the paper) for subsequent representation as single sequence from input pair for corresponding multilingual model. Several models of summarizers were analyzed, among which following were selected:

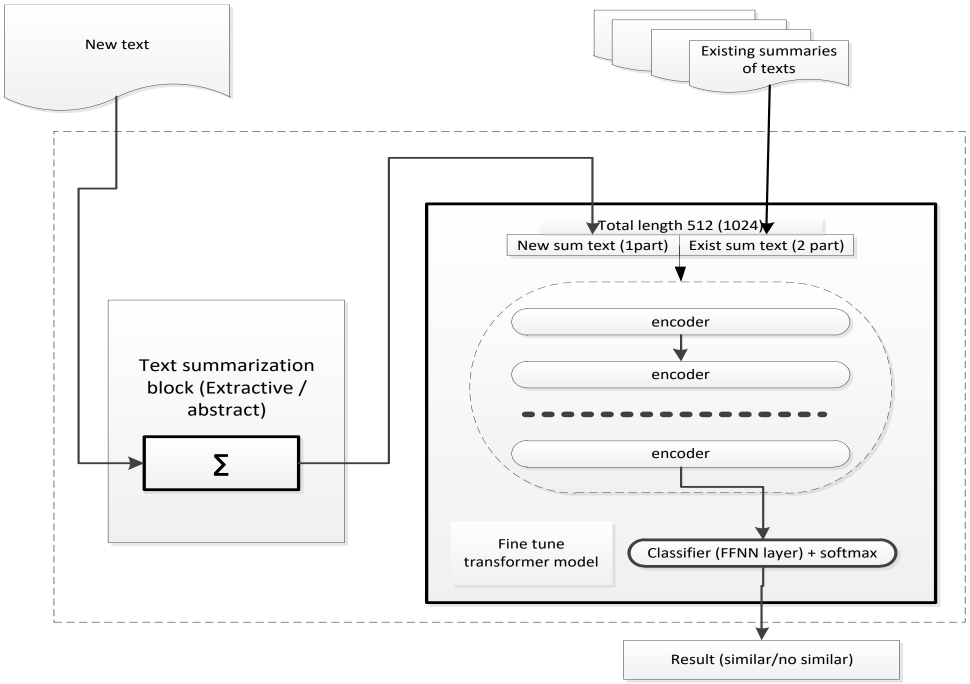
New text
Existing summaries of texts
Text summarization block (Extractive / abstract)
Σ
Total length 512 (1024)
New sum text (1part)
Exist sum text (2 part)
encoder
encoder
encoder
Fine tune transformer model
Classifier (FFNN layer) + softmax
Result (similar/no similar)

Fig. 1. Software module schematic representation for semantic similarity determination of arbitrary length texts

1) As extractive models – LSA, KLS, LexRankS and TextRank;

2) As abstract models – based on Transformers T5, mT5 and Pegasus, available in the Hugging Face library for TensorFlow [15].

In particular, mT5 model is a multilingual version of the English T5 model, trained on multilingual dataset. With the help of this multilingual model actually the abstract summarization (generalization) problem is solved for a multilingual text (at the moment the main limitation is the size of the input text).

Models were evaluated using English documents and their performance was compared by their ROUGE score. Based on the results, a decision was made to use the latent semantic analysis (LSA) model as an extractive model and the mT5 model (a massively multilingual pretrained text-to-text transformer).

In this paper three Transformers models are considered (BERT, DistilBERT and XML) for the semantic similarity determination of texts, available in Hugging Face library for TensorFlow [15]. Multilingual model BERT (trainable parameters 177,854,978) for 104 languages is a bidirectional converter, preliminarily trained using the combination of target simulation utilizing masked language modeling (MLM) and next sentence prediction (NSP) [2]. While training NSP in the BERT model, a specialized token (CLS) was used as a sequence for prediction results estimation. In given paper this token (first token in the sequence) is used to solve the classification problem of text sequence pairs for all Transformers models. Multilingual DistilBERT model (trainable parameters 135,326,210) for 104 languages is a lightweight BERT version, operating faster by 60% and retaining more than 95% of BERT performance, measured in GLUE test (General Language Understanding Evaluation) [3]. Multilingual XML model (trainable parameters 571,499,522) for 100 languages is a pretrained Transformer model using one of the following objectives [4]:

1) a causal language modeling (CLM) objective (next token prediction);

2) a masked language modeling (MLM) objective (BERT-like);

3) a Translation Language Modeling (TLM) object (BERT MLM extension for multiple language inputs).

Thus, XML model is not directly trained for NSP, unlike BERT and DistilBERT models. By the way, input data format for XML model, like in BERT model, ensures encoding of two different sequences in the equal input identifiers (token type IDs). At the same time input data format of DistilBERT model does not have token type IDs. That is, the given model does not indicate, which token belongs to which text sequence segment. To solve this problem, DistilBERT model segments are simply separated with the help of a special token (SEP) (same as in BERT model).

As part of the study, for the text sequence pairs classification problem Transformers model adopt final latent state $\boldsymbol{s}$.

As the activation function of the fully connected classifier layer, a $softmax$ function is used to predict the probability p towards the class label l [17]:

$$p(l|\boldsymbol{s}) = softmax(W_S),. \qquad (1)$$

where W_S – resulting tensor of the latent state $\boldsymbol{s}$.

Fine-tuning of Transformers models (BERT, DistilBERT, XML) for text sequence pairs classification problem solving was carried out with the MRPC dataset. Paraphrasing detection is a problem of studying two text objects and determining whether they are the same value. In the general case, for obtaining high accuracy while solving this problem, both syntactic and semantic thorough analysis of two text objects is required. Based on paraphrasing style, paraphrases can be distributed into five types [17]. In the framework of this study given the text semantic similarity determination context special meaning has fifth paraphrase type (complex paraphrase).

Fine-tuning results of multilingual Transformers models (BERT, DistilBERT, XML) are presented in the Fig. 2-4 and Table 1 [18].

Fig. 2 chart analysis shows that already at the third training epoch for all models discrepancy happens between “validate loss” and “validate accuracy” values. That is, for the given models fine-tuning process up to two training epochs is enough using corresponding dataset.

Table 1 data analysis demonstrates that the BERT model has the highest F1-score values according to the multilingual Transformers models fine-tuning results for solving the text sequence pairs classification problem using MRPC dataset. At the same time precision measure values for DistilBERT and XML models are very close.

This is in a view of the fact that XML model has 4 times more trainable parameters than DistilBERT.

Table 1 – **F1-score values based on the fine-tuning results**

Model		**F1-score**
BERT	no similar	0,82
	similar	0,92
	accuracy	0,88
DistilBERT	no similar	0,74
	similar	0,90
	accuracy	0,86
XLM	no similar	0,76
	similar	0,91
	accuracy	0,86

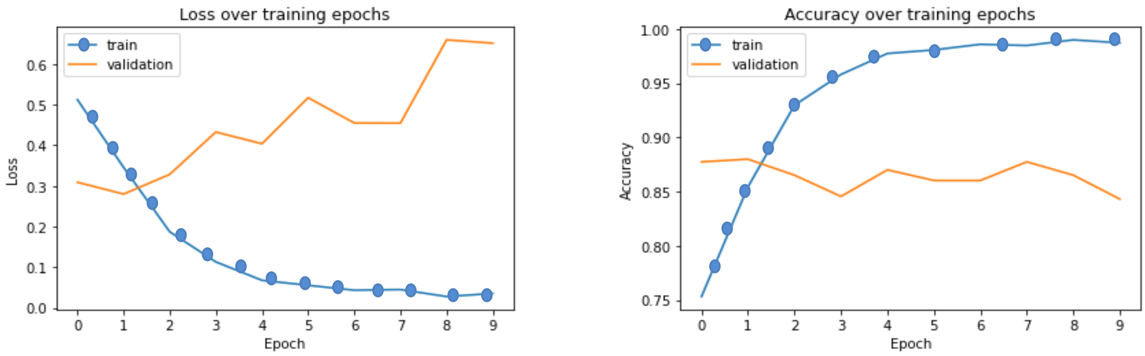

Fig. 2. Plots of error and accuracy of training for BERT

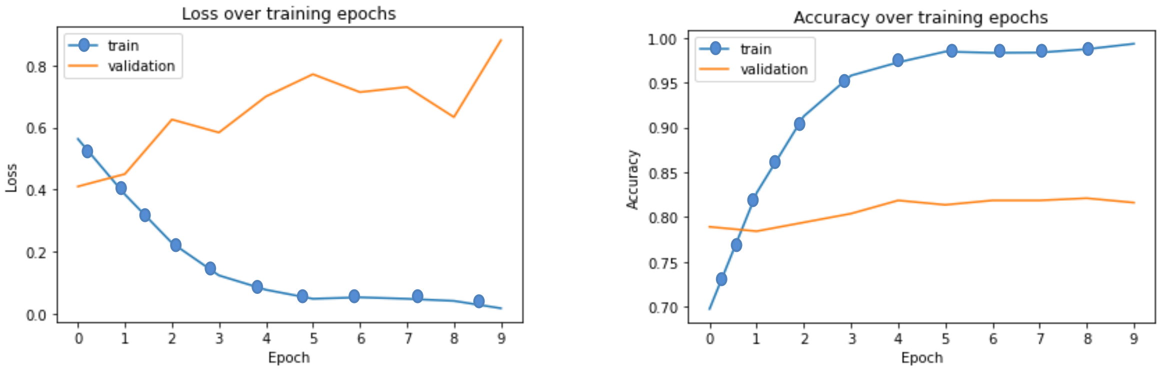

Fig. 3. Plots of error and accuracy of training for the DistilBERT model

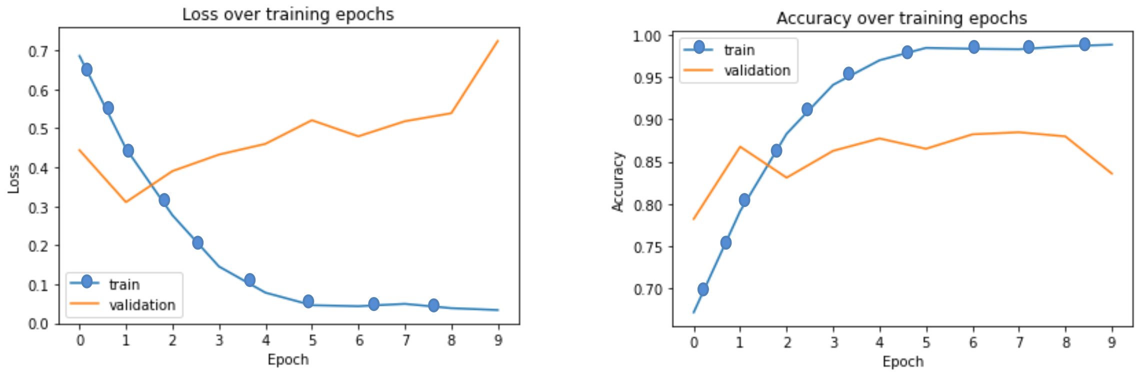

Fig. 4. Plots of error and accuracy of training for the XML model

So, when there are no hardware limitations for solving the multilingual news contents semantic similarity determination problem, the most effective way is to use multilingual BERT model. If there are restrictions, it is best to use multilingual DistilBERT model. The application of XML model for solving the given class of problems within the framework of the first approach is not very effective, since this basic model was not trained using the next sentence prediction target.

Conclusions

This paper describes a developed method of arbitrary length text semantic similarity determination based on text sequence pairs classification problem solving using Transformers model.

A feature of the given method implementation is the preliminary text processing more than 512 (1024) tokens long with the intellectual text summarizer application and subsequent usage of fine-tuned Transformers model for the text sequence pairs classification problems. Latent semantic analysis (LSA) model as extractive model and mT5 model (a massively multilingual pre-trained text-to-text transformer) as abstract model were used to perform summarization tasks.

Studies have shown that BERT model usage is the most effective as classification model when there are no hardware limitations.

When there are constraints, DistilBERT model usage will be the most effective.

The benefit of the given method is, above all, the possibility to overcome input sequence restrictions while determining semantic similarity of the texts in combination with the fine-tuned pretrained Transformers model advantages utilization.

References

1. Olizarenko, S. and Argunov, V. (2020), “On possibilities of multilingual Bert model for determining semantic similarities of the news content”, *Control, navigation and communication systems*, Poltava: NU PP, No. 3(61), pp. 94-99.
2. Devlin, J., Ming-Wei Chang, Lee, Ke and Toutanova, K. (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805v2 [cs.CL] 24 May 2019.
3. Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020), *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, arXiv:1910.01108v4 [cs.CL] 1 Mar 2020.
4. Guillaume, Lample and Alexis, Conneau (2019), *Cross-lingual Language Model Pretraining*, arXiv:1901.07291v1 [cs.CL] 22 Jan 2019.
5. Sun, C., Qiu, X., Xu, Y. and Huang X. (2020), *How to Fine-Tune BERT for Text Classification*, arXiv:1905.05583v3 [cs.CL].
6. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y., Strope, B. and Kurzweil R. (2019), *Multilingual Universal Sentence Encoder for Semantic Retrieval*, arXiv:1907.04307v1.
7. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.St., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B. and Kurzweil, R. (2018), “Universal sentence encoder for English”, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174.

8. Yoon, Kim (2014), “Convolutional neural networks for sentence classification”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751.
9. Ashish, Vaswani, Noam, Shazeer, Niki, Parmar, Jakob, Uszkoreit, Llion, Jones, Aidan, Gomez, Łukasz, Kaiser, and Illia, Polosukhin (2017), “Attention is all you need”, *Proceedings of NIPS*, pp. 6000–6010.
10. (2020), Multilingual *Similarity Search Using Pretrained Bidirectional LSTM Encoder. Evaluating LASER (Language-Agnostic SEntence Representations)*, available at: https://medium.com/the-artificial-impostor/multilingual-similarity-search-using-pretrained-bidirectional-lstm-encoder-e34fac5958b0.
11. (2019), *Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library*, POSTED ON JAN 22, 2019 TO AI RESEARCH, available at: https://engineering.fb.com/ai-research/laser-multilingual-sentence-embeddings.
12. Reimers, N. and Gurevych I. (2019), *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, arXiv:1908.10084v1 [cs.CL] 27 Aug 2019.
13. Patel, M. (2019), *TinySearch - Semantics-based Search Engine using Bert Embeddings*, available at: https://arxiv.org/ftp/arxiv/papers/1908/1908.02451.pdf.
14. Han, X. (2020), *Bert-as-service*, available at: https://github.com/hanxiao/bert-as-service..
15. (2020), *State of the art Natural Language Processing for Pytorch and TensorFlow 2.0*, available at: https://huggingface.co/transformers/index.html.
16. Goodfellow, I., Bengio, Y. and Courville, A. (2018), *Softmax Units for Multinoulli Output Distributions. Deep Learning*, MIT Press. pp. 180–184, ISBN 978-0-26203561-3.
17. Dolan, B. and Brockett, C. (2005), “Automatically Constructing a Corpus of Sentential Paraphrases*”, Proceedings of the 3rd International Workshop on Paraphrasing* (IWP 2005), Jeju Island, pp. 9–16.
18. Olizarenko, S. and Argunov, V. (2020), “Research on the specific features of determining the semantic similarity of arbitrary-length text content using multilingual Transformer-based models”, *Advanced Information Systems*, Vol. 4, No. 3, pp. 94-103, DOI: https://doi.org/10.20998/2522-9052.2020.3.13

Received (надійшла) 12.02.2021
Accepted for publication (прийнята до друку) 12.05.2021

About the authors / Відомості про авторів

Олізаренко Сергій Анатолійович – доктор технічних наук, старший науковий співробітник, професор кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Serhii Olizarenko – Doctor of Technical Sciences, Senior Researcher, Professor of the Electronic Computers Department, Kharkiv National University of Radio Electronics University, Kharkiv, Ukraine;
e-mail: sergejolizarenko5@gmail.com; ORCID ID: https://orcid.org/0000-0002-7762-6541.

Радченко В’ячеслав Олексійович – старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Viacheslav Radchenko – Senior Lecturer of the Electronic Computers Department, Kharkiv National University of Radio Electronics University, Kharkiv, Ukraine;
e-mail: viacheslav.radchenk@gmail.com; ORCID ID: https://orcid.org/0000-0002-2505-1969.

Метод визначення семантичної подібності текстів довільної довжини з використанням моделей Transformers

С. А. Олізаренко, В. О. Радченко

Анотація. В роботі розглянуті результати розробки методу визначення семантичної подібності текстів довільної довжини на основі їх векторних уявлень. При цьому векторні уявлення отримані з використанням мультимовної моделі Transformers, а безпосередньо завдання визначення семантичного подібності текстів довільної довжини розглядається як задача класифікації пар текстових послідовностей з використанням моделі Transformers. Виконано порівняльний аналіз найбільш оптимальної моделі Transformers для вирішення даного класу задач. Основними етапами методу при цьому розглядаються етап тонкої настройка моделі Transformers в рамках другого завдання преднавченрї моделі (завдання прогнозування пропозицій), а також етап вибору і реалізації методу суммарізації текстової послідовності довжиною понад 512 (1024) токенів для вирішення завдання визначення семантичного подібності текстів довільної довжини.

Ключові слова: текст; довільна довжина; семантична подібність; векторне подання; модель Transformer; тонке налагодження.

Метод определения семантического подобия текстов произвольной длины с использованием моделей Transformers

С. А. Олизаренко, В. А. Радченко

Аннотация. В работе рассмотрены результаты разработки метода определения семантического подобия текстов произвольной длины на основе их векторных представлений. При этом векторные представления получены с использованием мультиязычной модели Transformers, а непосредственно задача определения семантического подобия текстов произвольной длины рассматривается как задача классификации пар текстовых последовательностей с использованием модели Transformers**.** Выполнен сравнительный анализ наиболее оптимальной модели Transformers для решения данного класса задач. Основными этапами метода при этом рассматриваются этап тонкой настройка модели Transformers в рамках второй задачи предобученной модели (задачи прогнозирования предложений), а также этап выбора и реализации метода суммаризации текстовой последовательности длиной более 512 (1024) токенов для решения задачи определения семантического подобия текстов произвольной длины.

Ключевые слова: текст; произвольная длина; семантическое подобие; векторное представление; модель Transformer; тонкая настройка.