

Using BERT model to Identify Sentences Paraphrase in the News Corpus

Nina Khairova¹, Anastasiia Shapovalova¹, Orken Mamyrbayev², Nataliia Sharonova¹, Kuralay Mukhsina³

¹ National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., Kharkiv, 61002, Ukraine

² Institute of Information and Computational Technologies, 125, Pushkin str., Almaty, 050010, Republic of Kazakhstan

³ Al-Farabi Kazakh National University, 71 al-Farabi Ave., Almaty, Republic of Kazakhstan

Abstract

The paper analyzes the problems of semantic similarity identification of short text fragments and provides an overview of modern methods for solving them. The advantages of using the fine-tuned Sentence-BERT language model for paraphrased sentences classification over other modern models are proved. This paper presents the software implementation of the algorithm for automatic paraphrased sentences identification in a created corpus of English texts scraped from two news sites. The detailed information on designing and preprocessing of the news corpus on COVID-19 topic is provided. The operation of the created algorithm for automatic creation of the resulting corpus of semantically similar sentences is described in stages. The evaluation of the created algorithm based on the resulting corpus is calculated with the assistance of experts. The precision of the paraphrased sentences identification achieves 77 %.

Keywords

Automatic paraphrase identification, corpus of website news, semantic similarity, Sentence-BERT language model.

1. Introduction

Recently, the problem of paraphrase identification by technical means has attracted considerable interest among researchers working in the field of information retrieval [1], automatic plagiarism detection [2] and machine translation [3]. On the one hand, the automation of this process can optimize and accelerate the processes of classification, sorting and further processing of texts in natural language. On the other hand, this task is significantly complicated by a number of factors: the presence of errors, filler and jargon words in the texts, problems of program recognition of homonymy and polysemy, the lack of lexical intersection between the compared sentences. In other words, in natural language different phrases can have a similar meaning, even without intersections in words, and conversely, the same words or phrases may have different meanings in the contexts.

All this makes the task of identifying semantically similar sentences complicated even for a person, because during the work you need to be able to determine the criteria of similarity in given texts, cover and understand the discourse in general, and consider particular expressions in their context. Therefore, the question of automatic paraphrase identification in natural language texts remains open, and this task requires further study.

Nowadays, in order to solve the problem of automatic paraphrase identification, such state-of-the-art approaches as cosine similarity metric, overlap coefficient, knowledge-based methods, and sentence

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland.

EMAIL: khairova.nina@gmail.com (N. Khairova), randomnick675@gmail.com (A. Shapovalova), morkenj@mail.ru (O.Mamyrbayev), nvsharonova@ukr.net (N.Sharonova), kuka_ai@mail.ru (K. Mukhsina)

ORCID: 0000-0002-9826-0286 (N. Khairova); 0000-0002-5282-0209 (A. Shapovalova); 0000-0001-8318-3794 (O.Mamyrbayev), 0000-0002-8161-552X (N.Sharonova), 0000-0002-8627-1949 (K. Mukhsina)



© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

embedding are applied. However, in most cases, these approaches require large training corpora. Usually, training corpora are used by scientists in order to carry out successful work with natural language texts.

Text corpus itself represents a well-ordered and structured array stored (in most cases) electronically for the convenience of further processing of the contained information. Corpora are created for a wide range of users and for the purpose of solving various problems. Sometimes when designing a corpus some elements of the source text have to be removed and text should be preprocessed.

In this study, we create a corpus of single-topic articles collected from two news websites, consisting of two subcorpora: the first one contains the news texts related to the topic of COVID-19 and collected from CNN website, and the second one contains the news articles on the same topic from Yahoo News website. Our task is to apply a classification method in order to find a pair of paraphrased sentences that have the same meaning in both subcorpora using a fine-tuned Sentence-BERT language model.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related works, corresponding with methods and approaches to working with automatic paraphrase identification tasks. Section 3 describes the applied method for the task of paraphrase identification. Section 4 introduces our text corpus of news articles and describes its usage in our experiments. Section 5 gives the representation and interpretation of obtained experiment results. Section 6 contains comparisons of the outcomes with previous studies. In the last Section 7 conclusions are drawn about the work done and plans for future research are announced.

2. Related Works

Nowadays, there are many approaches to the automatic paraphrase identification task, among which are several based on: (1) description (calculation of distances between words using well-known semantic networks, for example, WordNet) [4], (2) search in web-sources [5], (3) paths (simpath) based on taxonomy [6], (4) constructing trees of two comparable sentences and their comparison (for example, MaltParser), (5) semantic distance (including TextTerra, Semanticus, S-Space, Semantic Vectors and their analogues) [7], (6) neural network model Word2Vec (Adaptive skip-gram, FastText, GloVe).[8]

Most of the above methods are suitable for determining semantic similarity at the level of separate words and phrases, considering them out of context. This is unacceptable when it comes to comparing sentences and paragraphs of texts. Therefore, the Sentence BERT [9] language model based on Bidirectional Encoder Representations of Transformers (BERT) was chosen to create the algorithm for the automatic paraphrase identification task in this study.

BERT itself is a machine learning method that was developed in 2018 by Jacob Devlin and other co-authors who worked at Google. This invention proved to be extremely convenient and productive; for this reason, two years later the company used it in almost all search queries. BERT is based on the Transformer [10], a focus mechanism that studies the contextual relationships between words (or parts of words) in a text. In its common configuration, the Transformer contains two particular mechanisms: an encoder that reads the entered text and a decoder that generates predictions to perform the task. The general operation of the Transformer (which was originally mentioned in the paper [10]) is shown in Figure 1.

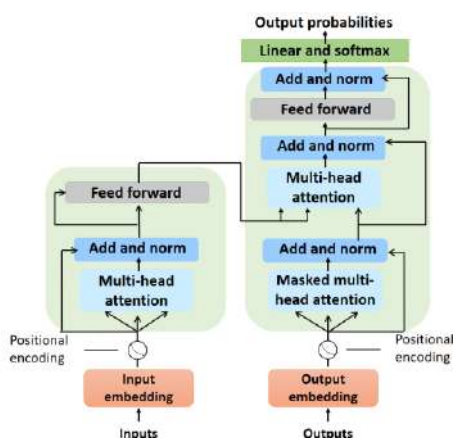


Figure 1: Transformer operation algorithm (from paper [10])

BERTBASE and BERTLARGE are built-in BERT models. They differ in the number of encoders and self-attention mechanisms – the former has 12 and 12, and the latter has 24 and 16, respectively. These models are pre-trained on word corpora with a total volume of 3300 million words. [11]

The BERT architecture can be applied to different types of tasks by adding specific markers related to the task and using word or sentence embedding. [9]

In order to determine the semantic similarity of two independent sentences, the authors of the BERT [9] developed the following principle: a token [SEP] is placed between these sentences and the sequence is processed through the model. The [CLS] marker is then stated to be used at the end of the sentence as input for simple classification or regression to find out whether the two sentences are related [12]. Although the proposed method of measuring the semantic similarity proves to be reliable enough, its operating algorithm becomes an obstacle when there is a necessity to fulfill the two following tasks:

1) Matching a pair of the most semantically similar sentences. When a classification is performed using BERT, then in order to accomplish this task, it is necessary to compare each specific pair of sentences and obtain an estimation of their semantic similarity, which would allow further comparisons. According to research by N. Reimers and I. Gurevych for n sentences, this leads to the formula $n(n-1)/2$ [13], which is a significant drawback of the model. According to researchers, such processing of even a small amount of data requires approximately 50 million iterations, which takes more than 2 days on a modern computer. Such large time costs make the use of BERT inconvenient and inefficient for such tasks.

2) Semantic search. The purpose of this task is to find the sentence that is most semantically similar to the entered query. One way to do this is to compare the input sentence with all the existing sentences in the dataset and perform n comparisons on the set of n sentences. According to the same group of scientists [13], due to the fact that BERT has to process each unique pair, a small set of sentences will take more than 40 seconds just to display the results of one query. In this case, the complexity of the calculations makes BERT unsuitable for the task at hand.

Due to the great difficulties of implementing these two scenarios by means of the BERT language model, there is a need to apply a new, more flexible and faster approach to solving the above issues. Most recently, N. Reimers and I. Gurevych [13] developed a twin network called Sentence-BERT model for processing two compared sentences simultaneously.

One of the tasks used by the authors of the model to evaluate the productivity of Sentence-BERT was SentEval [14]. This is a set of tasks that is usually used to assess the quality of sentence embedding. The classifier was fine-tuned over a set of vector representations of sentences, some of which are depicted in Figure 2, where multiple tasks (movie review, product review, binary sentiment analysis, etc.) and the sizes of the training and test sets are given. Also there is information about the presence in each dataset of “needs_train” (a model with its parameters that is trained on sentence embedding) and “set_classifier” (the possibility of determination of the parameters of a classifier).

Task	Type	#train	#test	needs_train	set_classifier
MR	movie review	11k	11k	1	1
CR	product review	4k	4k	1	1
SUBJ	subjectivity status	10k	10k	1	1
MPQA	opinion-polarity	11k	11k	1	1
SST	binary sentiment analysis	67k	1.8k	1	1
SST	fine-grained sentiment analysis	8.5k	2.2k	1	1
TREC	question-type classification	6k	0.5k	1	1
SICK-E	natural language inference	4.5k	4.9k	1	1

Figure 2: SentEval tasks for sentence embedding quality assessment

3. The Method

Our study follows the Sentence-BERT model [13], that allows two sentences to be processed contemporaneously and on the same basis. The fact is ensured by the interconnectedness of all weights and, therefore, by the possibility of repeated application of the model.

The model is based on BERT itself with a pooling layer added to it. This approach allows creating vector representations of a certain size for different sentences that are input. The main purpose of the approach is to encode the semantic component, so the network is configured with sources that contain information about semantic similarity. Including the Stanford Natural Language Inference Corpus (SNLI) and the Multi-Genre NLI (MG-NLI) makes it possible to train the model over more than a million pairs of sentences. The same set of data allows giving each pair of sentences a certain label that indicates whether the sentences correspond to “Entailment”, “Contradiction” or “Neutral” labels. Classification process is carried out according to the formula:

$$o = \text{softmax}(Wt(u, v, |u - v|)), \quad (1)$$

where u and v are the vectors for each pair of sentences, $|u - v|$ – the element-wise difference and W_t represents a trainable weight.

Subsequently, the authors [15] added the combination of the previously mentioned twin networks and the target function. Figure 3 shows the model of Sentence BERT dual architecture fine-tuned for the classification, which is basic for automatic paraphrase identification task.

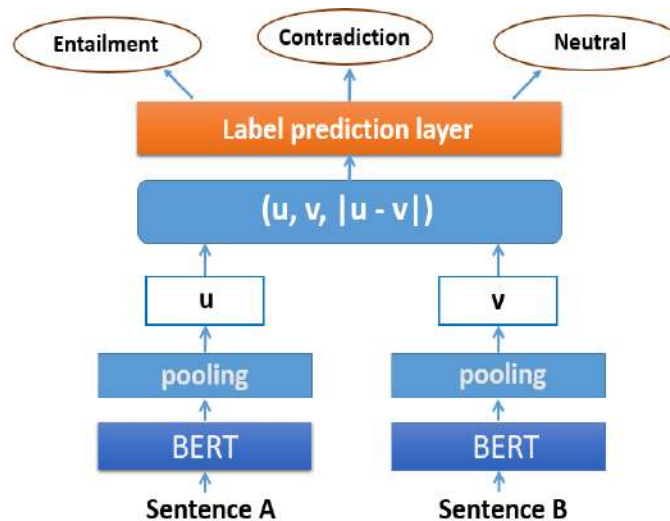


Figure 3: The Sentence-BERT dual architecture fine-tuned for the classification task

Following [15], we compared Sentence-BERT with other similar models such as sentiment prediction, sentence prediction and paraphrase identification from SentEval toolkit (sentiment prediction of movie reviews, consumer product reviews and newswire comments; sentence prediction from synopses and movie reviews; newswire comments classification; paraphrase identification in parallel news resources).

The results of these performance tests showed that the fine-tuned Sentence-BERT language model is superior to all other similar models in the tasks. The model shows excellent results even in those tasks for which it was not originally intended. Moreover, according to the data provided in the [15], Sentence-BERT outperformed all other embedding models in sentiment prediction tests.

Additionally, when choosing a model for paraphrase identification we should consider that the processing speed is of great importance, since sentence embedding is often performed on great amounts of text data. Comparison of the processing speed of various models shows that the standard BERT reveals unacceptably lower results in terms of the time spent on the process of embedding and outputting results (83 sentences per second), whereas Sentence-BERT model has no equal in speed when running on the Graphics Processing Unit (2042 sentences per second). In this way, what takes BERT more than 2 days, the Sentence-BERT language model has the opportunity to undertake within a few minutes.

In order to impartially assess the selected Sentence-BERT model we chose the SentEval approach, since it adapts the logistic regression classifier to sentences embedding. At the same time, this assumes that certain aspects still have a greater or lesser influence on the result of the classification.

Thus, we apply the Sentence-BERT language model that is an improved version of the BERT language model which allows us to achieve better results when embedding sentences and solve the problem of time spent on processing text information and outputting results.

4. Experiment

In order to conduct our experiments on the automatic paraphrase identification, it became necessary to create a special text corpus. Therefore, we carried out a number of specific steps to covert scrapped raw news texts related to the COVID-19 topic to a linguistic corpus.

The first stage was compliance checking. This stage was especially important since corpus texts should relate to a narrow topic, namely the COVID-19 problems. The second stage was processing and analysis of the texts. This stage includes edits and linguistic analysis of the received materials. The detected mistakes and typos in raw news texts were rectified. The third stage was transformation and cleaning of the texts, which is usually carried out at the operational level. During its implementation, such types of preprocessing as the removal, replacement of non-text elements and hyphens and ensuring the uniformity of spelling were carried out. The next step that we applied to create the corpus was graphematic analysis and segmentation of the text. At this stage, operations on the design of non-lexical characters and preprocessing of specific text elements were performed. At the fourth stage of corpus designing we marked up the text with several types of metadata. Table 1 shows the list of the metadata levels.

Table 1
Types of metadata

Type of metadata	Definition
Descriptive	Description of the information that is not directly related to the text content, including the author's name and title of the article.
Structural	Information about the place where the article was downloaded from and the date/time when the text file was downloaded. It provides text organization and corpus data structure. The text sets in our corpus are found and grouped by this type of metadata.

Additionally, in our case, the final step in creating the corpus was access providing. Different types of users were provided with rights and opportunities within the limits allowed by the administrator.

In order to implement an experimental research of the model that identifies semantically similar sentences, the two sources of news texts with the common subject matter were selected as a content of the corpus. The first source was CNN (edition.cnn.com), and the second source was Yahoo News (news.yahoo.com). The news articles were collected automatically by a scraping procedure that is based on the Python BeautifulSoup library from October 9, 2021 to October 31, 2021. Thereafter texts related to the COVID-19 pandemic were manually selected.

The CNN subcorpus contains 178 text files and the size of every article is approximately 5,000 characters, the Yahoo News subcorpus contains 204 text files and its typical size is about 4000-5000 characters per article. Figure 4 shows the same structure of both subcorpora. The first 4 digits of a file name, separated by a point, indicate the date of publication of the chosen article on the resource (CNN or Yahoo News), and the number in parentheses demonstrates a serial number of the news file in accordance with other files marked with the same date.

Some summary articles for the week contained redundant images, advertisements and labels that caused excessive "noise" that interfered with automatic word processing, and therefore the articles have been processed and corrected manually. After preprocessing, all the text files received a similar appearance to the one shown in Figure 4 (on the right).

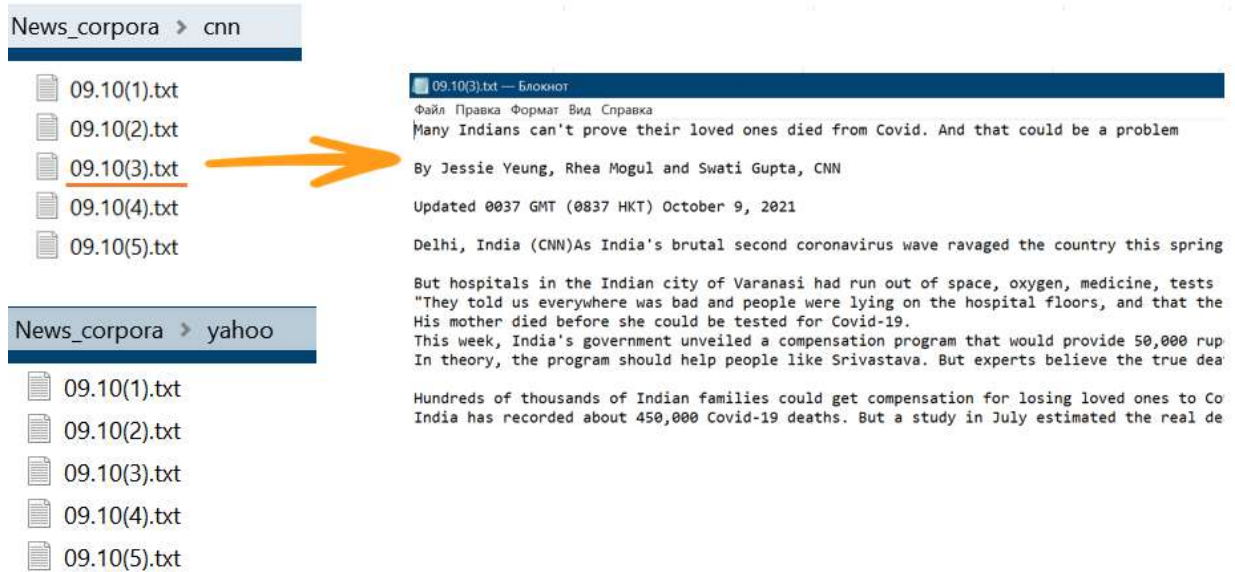


Figure 4: The fragment of our news corpus structure

To automatically create a corpus of semantically similar pairs of sentences, a software algorithm was proposed, the implementation of which can be represented in several following steps.

1. Ensuring the user enters the path to each of the two subcorpora for the purpose of providing the program with access to the textual information to be analyzed; implementation of the opportunity to create a path to a new empty corpus (using the methods of the *argparse* module for the command line). Creating a template for the final form of the resulting corpus.
2. Reading all sentences from both subcorpora, tokenizing them, removing "noise" (redundant information) automatically using templates of "garbage" (date, time of publication, etc.). Collecting all the sentences into one list.
3. Using the fine-tuned model 'all-MiniLM-L6-v2' based on Sentence Transformer, that displays sentences on 384-dimensional dense vector space and compares them. This model was configured on several dozen data sets, shown in Figure 5, and their total volume is more than a million pairs of sentences [16]. Recording the result to the new corpus. The code snippet looks as follows:

```
def mine(sentences1, sentences2):
    all_sentences = sentences1 + sentences2
    paraphrases = [
        (score, all_sentences[i], all_sentences[j])
        for score, i, j in paraphrase_mining(SentenceTransformer('all-MiniLM-L6-v2'),\
            all_sentences, show_progress_bar=True)]
```

4. Deleting pairs of sentences that belong to the same corpus, in order to prevent comparisons of the same sentence or similar sentences from a common news resource.

Dataset	Paper	Number of training tuples
<u>Reddit comments (2015-2018)</u>	<u>paper</u>	726,484,430
<u>S2ORC Citation pairs (Abstracts)</u>	<u>paper</u>	116,288,806
<u>WikiAnswers Duplicate question pairs</u>	<u>paper</u>	77,427,422
<u>PAQ (Question, Answer) pairs</u>	<u>paper</u>	64,371,441
<u>S2ORC Citation pairs (Titles)</u>	<u>paper</u>	52,603,982
<u>S2ORC (Title, Abstract)</u>	<u>paper</u>	41,769,185
<u>Stack Exchange (Title, Body) pairs</u>	-	25,316,456
<u>Stack Exchange (Title+Body, Answer) pairs</u>	-	21,396,559
<u>Stack Exchange (Title, Answer) pairs</u>	-	21,396,559
<u>MS MARCO triplets</u>	<u>paper</u>	9,144,553

Figure 5: List of several datasets used to fine-tune 'all-MiniLM-L6-v2'

5. Results

As an output of the program, an unmarked corpus was obtained with ten text files (subcorpora), the names of which are ranked from 0 to 10 (or from 0.0 (no matches) to 1.0 (one hundred percent match)), filled in depending on the level of semantic similarity between sentence pairs. Figure 6 shows the resulting corpus. We can observe that the files that include sentences with too low similarity rate (0.4 and below) are left blank. Absolute matches were also not found (there is almost no content in file 10). The largest number of sentences is contained in files with sentence similarity ranks from 0.5 to 0.8.

Имя	Дата изменения	Тип	Размер
0.txt	20.12.2021 7:03	Текстовый документ	0 КБ
1.txt	20.12.2021 7:03	Текстовый документ	0 КБ
2.txt	20.12.2021 7:03	Текстовый документ	0 КБ
3.txt	20.12.2021 7:03	Текстовый документ	0 КБ
4.txt	20.12.2021 7:03	Текстовый документ	0 КБ
5.txt	20.12.2021 7:03	Текстовый документ	20 836 КБ
6.txt	20.12.2021 7:03	Текстовый документ	19 666 КБ
7.txt	20.12.2021 7:03	Текстовый документ	3 127 КБ
8.txt	20.12.2021 7:03	Текстовый документ	310 КБ
9.txt	20.12.2021 7:03	Текстовый документ	22 КБ
10.txt	20.12.2021 7:03	Текстовый документ	1 КБ

Figure 6: The resulting corpus with the pairs of semantically similar sentences. The number in each file name corresponds to the semantic similarity rank of the paraphrases in the given file.

Due to the fact that, as a rule, single-topic texts contain the most pairs of sentences with general similarity level of 50% - 70%, the resulting corpus contains 20 836 kb of information in file number 5, 19 666 kb in file number 6, and 3 127 kb in file number 7.

All the files in the resulting corpus have the same structure. The similar sentences are in pairs, there is always a gap between the pairs. Figure 7 shows the fragment of one of the obtained corpus files.

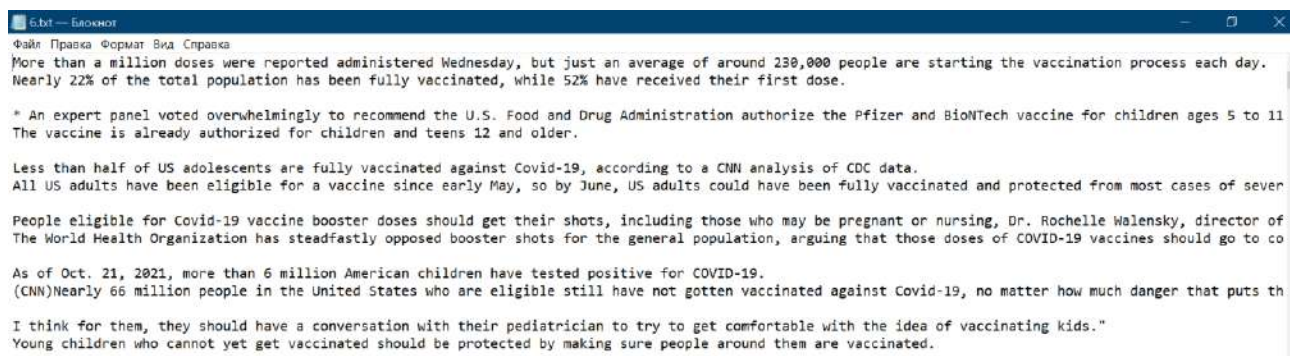


Figure 7: One of the resulting corpus files

Analysis of the results allows us to ensure that there are pairs of sentences of both the same length and different ones. This confirms the fact that the Sentence-BERT language model considers the similarity of sentences not only through synonyms and similar structure, but also through the common meaning of two sentences.

Figures 8 and 9 illustrate a fragment of results from each non-empty file of obtained paraphrase corpus. Figure 8 shows 3 random pairs of sentences that were selected from each 5, 6 and 7 ranks files. Here in files of rank 5 or 6, we can conclude that the similarity of sentences is not entirely obvious, but it is clear enough that each pair of sentences has a common subject or object, which is discussed in a particular discourse. For instance, in the first sentences of the fifth file it is the vaccination of a certain group of people, in the second pair of sentences it is the mortality rate among the seniors and so on. Also analyzing the obtained results, we conclude that one sentence can occur in one file many times, because it potentially has several semantically similar sentences (not a single one) in another input news subcorpus. In the file of seventh rank, sentences which are similar not only in common object of discussion, but also structurally (third pair of sentences) began to appear.

Subcorpus name	First sentence	Second sentence
5.txt	1. And the drop in cases comes as more Americans get vaccinated.	1. This is particularly true if all kids in a class are known to be vaccinated.
	2. Of those breakthrough cases resulting in death, 85% were among people age 65 and older and 57% were among men, according to the CDC.	2. Among the community cases were 531 seniors aged 60 and above.
	3. But when it comes to vaccinating children, you want to make sure the safety profile is solid," he said.	3. This is particularly true if all kids in a class are known to be vaccinated.
6.txt	1. Vaccines for kids under 5 may not come until next year.	1. With young children unable to even receive the vaccine at this point, this brings a level of unnecessary risk.
	2. Among them, six had been unvaccinated against COVID-19, two had been partially vaccinated and one had been fully vaccinated.	2. (CNN) Nearly 66 million people in the United States who are eligible still have not gotten vaccinated against Covid-19, no matter how much danger that puts them in and no matter how many incentives or mandates have been thrown at them.
	3. "Also, individuals who are hesitant to receive COVID-19 vaccinations may be influenced by vaccine misinformation."	3. "This is not about a vaccine, or medicine or a virus.
7.txt	1. CNN: What do we know about the effectiveness of the Covid-19 vaccine for children ages 5 to 11?	1. Watch: COVID vaccine rolled out to 12- to 15-year-olds
	2. Over the last week, an average of 87,676 people reported infections and 1,559 people died of Covid-19 a day, according to JHU data.	2. In September, "COVID-19 took the lives of 1,899 people per day on average," KFF writes.
	3. Children will also have the option to be vaccinated at school, according to Arwady.	3. Kids are already required to get a host of other vaccines to attend school.

Figure 8: Random pairs of sentences from files ranked as 5, 6 and 7

Figure 9 shows an example of the paraphrased sentences that were selected from both 8 and 9 ranks files. In these files there are the best results for the task of finding two most similar sentences. Here, sentence pairs

often have the same object of discourse, the same numbers and statistics (age, percentage) and almost the same sentence structure. We can see such level of similarity in the first pair of sentences of the eighth rank file.

Subcorpus name	First sentence	Second sentence
8.txt	1. As of Tuesday night, only 56.5% of the US population was fully vaccinated, according to CDC data.	1. Based on the latest data from the CDC, 56.5 percent of the country's population has been fully vaccinated so far.
	2. J&J vaccine recipients should get their second dose as soon as it's available, experts say.	2. The panel also recommended a second shot of the J&J vaccine for all recipients at least two months after receiving their first.
	3. The FDA's independent vaccine advisers will meet to discuss Pfizer's request for authorization for its Covid-19 vaccine for children ages 5 to 11.	3. Kids ages 5-11 could start getting Pfizer's COVID-19 vaccine next month.
9.txt	1. But it left in place a complex formula for who should get boosters and when, with officials saying they may simplify the framework as more safety data comes in.	1. But it left in place a complex formula for who should get boosters and when, so officials can simplify the framework as more safety data is gathered.
	2. About 56.4% of the US population is fully vaccinated against Covid-19, according to the US Centers for Disease Control and Prevention (CDC).	2. More than 188.2 million people in the U.S. are fully vaccinated against COVID-19 as of Oct. 14 — about 57% of the total population, a Centers for Disease Control and Prevention tracker shows.
	3. "If you are a Christian, or if you're anybody who has not yet gotten vaccinated, hit the reset button on whatever information you have that's causing you to be doubtful or hesitant or fearful, and look at the evidence."	3. "Let me make a plea right here that if you are a Christian, or if you're anybody who has not yet gotten vaccinated, hit the reset button on whatever information you have that's causing you to be doubtful or hesitant or fearful and look at the evidence," Collins said.

Figure 9: Random pairs of sentences from files ranked as 8 and 9

6. Discussions

In order to estimate the correctness and accuracy of the program, we analyzed the result of processing a random 200 pairs of sentences from the most efficient ranks of the files (50 pairs from each file). The selected sentences were placed in the table, where the first column denotes the name of the subcorpus, the second and the third columns contain the compared sentences of the file, and the fourth one reflects the expert's binary assessment. The digit 1 implies "correspondence" (the sentences are semantically the same or similar enough) and the digit 0 implies "contradiction" (the sentences have different meanings). Fig. 10 shows a fragment of the table for evaluating the result of comparing the semantic similarity of sentence pairs drawn from the ranking file 7.

Subcorpus name	First sentence	Second sentence	Expert evaluation
7.txt	1. The total number of identified Covid-19 cases in Russia has risen to 7,958,384 since the beginning of the pandemic.	1. The country has seen 377 cases of COVID-19 between October 17-29, according to National Health Commission (NHC) data.	0
	2. And other questions about Covid-19 boosters	2. You have COVID booster shot questions.	1
	3. By May 9, around 570,000 Covid-19 deaths had occurred in the United States.	3. In the week ending Saturday, there were less than 50,000 COVID deaths globally for the first time since the week ending Nov. 3, 2020.	0
	4. We may not appreciate that, because about 50% of the infections in children are asymptomatic," Fauci told a White House briefing.	4. Dr. Anthony Fauci recently said about 50% of the infections in children are such.	1
	5. Pfizer has submitted data and a formal request for authorization for its one-third dose vaccine for use in children 5-11.	5. - Pfizer has formally asked the FDA to authorize its COVID-19 vaccine for emergency use among kids ages 5 to 11.	1
	6. You can now get a Moderna and Johnson & Johnson Covid-19 booster at Walgreens nationwide	6. Walmart, Walgreens U.S. stores roll out Moderna, J&J COVID-19 booster shots	1

Figure 10: An example of a random sampling analysis for file 7

The precision of the resulting paraphrased sentences was calculated by the well-known formula:

$$Precision = TP / (TP + FP), \quad (2)$$

where TP represents “True Positive” (well-defined semantic proximity) and FP denotes “False Positive” (misidentified similarity). Table 2 shows precision coefficients obtained for each ranked file.

Table 2

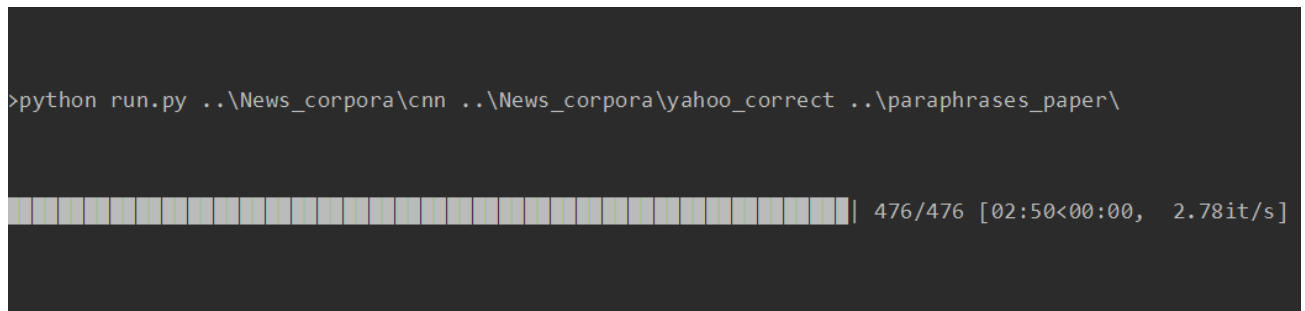
Expert evaluation results

Subcorpus name	Correspondence	Estimation results (precision)
6.txt	30 of 50	0,6
7.txt	35 of 50	0,7
8.txt	43 of 50	0,86
9.txt	46 of 50	0,92

As a result, we get the following values for the entire set of pairs: 154 sentence pairs were correctly identified as semantically similar, and 46 were wrongly placed in the resulting corpus. Then we calculate precision for the entire sample and get a result of 0.77, which is the corpus estimate.

It was mentioned earlier that N. Reimers and I. Gurevych [13] evaluated the SBERT embedding on the specific SentEval tasks, among which was the question-type classification and paraphrase identification in parallel news resources, which fundamentally intersects with the task of our study. In their work, SBERT-NLI-base scored 87.41 and SBERT-NLI-large scored 87.69 for this task. If we compare these results with the data obtained on the operation of our algorithm, we can conclude that the program has completed the task at a sufficiently high level (more than 75% correct).

Another important task was to obtain the output of the program in the shortest possible time. As we can see from Figure 11, the model has successfully processed a large amount of data (about 50 thousand kilobytes of information and 476 sentence embeddings) in almost 3 minutes.



```
>python run.py ..\News_corpora\cnn ..\News_corpora\yahoo_correct ..\paraphrases_paper\  
| 476/476 [02:50<00:00, 2.78it/s]
```

Figure 11: Screenshot of the console running with the specified execution time

This process was carried out on Intel(R) Core(TM) i5-8250U CPU 1.60Gz 1.80GHz and this fact indicates that the processing of data via this algorithm on the GPU will be even faster.

7. Conclusions

The problems of semantic similarity identification in texts in natural language are analyzed, and an overview of modern methods of solving them is provided. As a result of our research, a software implementation of the algorithm for detecting and processing semantically similar sentences and saving them in a text corpus is offered. The estimation of the results is given.

The created algorithm entirely fulfills the goal of paraphrase identification and writing them in the text corpus; analyzes the corpus of processed news texts and detects semantically similar pairs of sentences, writes in a separate text corpus and ranks them by the level of semantic similarity. The program implementation is suitable for analyzing other similar paired subcorpora, but is limited by only one language (English).

In addition, the algorithm is remarkably fast and performant in terms of sentence processing time, comparable to other leading semantic analysis algorithms. In future studies, it is planned to implement the possibility to use the algorithm to process and semantically analyze Ukrainian texts.

8. Acknowledgements

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP09259309).

9. References

- [1] D. Dinh and N. Le Thanh, English–Vietnamese Cross-Language Paraphrase Identification Using Hybrid Feature Classes, *Journal of Heuristics*, (2019). doi:10.1007/s10732-019-09411-2.
- [2] A. Altheneyan, M.E. Menai, Evaluation of State-of-the-Art Paraphrase Identification and its Application to Automatic Plagiarism Detection, *International Journal of Pattern Recognition and Artificial Intelligence* 34 (2020). doi:10.1142/S0218001420530043.
- [3] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, PAWS-X: A Cross-Lingual Adversarial Dataset for Paraphrase Identification, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3687–3692. Association for Computational Linguistics.
- [4] K. Orkphol, W. Yang, Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet, *Future Internet*, (2019), p. 6. doi:10.3390/fi11050114.
- [5] D. Coltuc, M. Datcu, and D. Coltuc, On the Use of Normalized Compression Distances for Image Similarity Detection, *Entropy*, vol. 20, no. 99, (2018) 1–15. doi:10.3390/e20020099.
- [6] M. Oussalah, M. Mohamed, Knowledge-Based Sentence Semantic Similarity: Algebraical Properties. *Prog Artif Intell* 11, (2022) 43–63. doi:10.1007/s13748-021-00248-0.
- [7] D. Bondarchuk, G. Timofeeva, Vector Space Model Based on Semantic Relatedness, in: *41st International Conference “Applications of Mathematics in Engineering and Economics” AMEE’15*, (2015) 3-4. doi:10.1063/1.4936683.
- [8] K. Orkphol, W. Yang, Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet, *Future Internet*, (2019), p. 10. doi:10.3390/fi11050114.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: *the 1th NAACL-HLT*, (2019). arXiv:1810.04805v2.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention Is All You Need, in *Proc. NIPS*, (2017), pp. 2-4. arXiv:1706.03762v5.
- [11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: *CoRR*, (2018). arXiv:1810.04805.
- [12] D. Viji, S. Revathy, A Hybrid Approach of Weighted Fine-Tuned BERT Extraction with Deep Siamese Bi – LSTM Model for Semantic Text Similarity Identification, *Multimedia Tools and Applications*, volume 81, 6131–6157, (2022). doi:10.1007/s11042-021-11771-6.
- [13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019, pp. 3982–3992. Association for Computational Linguistics.
- [14] A. Conneau and D. Kiela, SentEval: An Evaluation Toolkit for Universal Sentence Representations, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 1699 –1704. Facebook Artificial Intelligence Research.
- [15] N. Reimers, I. Gurevych, Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 4512–4525. arXiv:2004.09813.
- [16] Huggingface.co: all-MiniLM-L6-v2, 2021. URL: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>