http://dx.doi.org/10.35596/1729-7648-2026-24-2-85-91

УДК 004.021

СРАВНЕНИЕ МЕТОДОВ ОЦЕНКИ СЕМАНТИЧЕСКОГО СХОДСТВА ТЕКСТОВЫХ ФРАГМЕНТОВ

К. С. КРЕЗ, Е. Н. ШНЕЙДЕРОВ, П. А. ШИШ, Е. В. КОНДРАТЕНКО

Белорусский государственный университет информатики и радиоэлектроники (Минск, Республика Беларусь)

Аннотация. В условиях быстрого роста объема текстовых данных появляется потребность в методах, способных эффективно сравнивать фрагменты текста по смыслу, включая случаи перефразирования, синонимизации и перестройки структуры предложений. Одна из актуальных задач – сопоставление результатов методов семантического сравнения на основе различных моделей с человеческим восприятием смысловой близости. В статье рассматривается экспертный метод оценки семантического сходства текстовых фрагментов, основанный на оценках участников анкетирования. Суть метода заключается в формировании интерпретируемой шкалы семантической близости, полученной на основе человеческого восприятия содержания текстов и используемой для анализа согласованности различных методов. Для формирования «человеческой» оценки проведен опрос 138 участников. Сравнительный анализ показал, что различные методы оценки семантического сходства демонстрируют неодинаковую степень согласованности с человеческим восприятием смысловой близости текстов.

Ключевые слова: семантическое сходство, обработка естественного языка, экспертный метод, сравнение текстов, шкалирование оценок, анкетирование, корреляция Пирсона, корреляция Спирмена.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Сравнение методов оценки семантического сходства текстовых фрагментов / К. С. Крез [и др.] // Доклады БГУИР. 2026. Т. 24, № 2. С. 85–91. http://dx.doi.org/10.35596/1729-7648-2026-24-2-85-91.

COMPARISON OF METHODS FOR ASSESSING THE SEMANTIC SIMILARITY OF TEXT FRAGMENTS

KARINA KREZ, EVGENI SHNEIDEROV, POLINA SHISH, EKATERINA KONDRATENKO

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Abstract. With the rapid growth of text data, there is a need for methods capable of effectively comparing text fragments by meaning, including cases of paraphrasing, synonymization, and sentence restructuring. One of the pressing challenges is comparing the results of semantic comparison methods based on various models with the human perception of semantic similarity. This article discusses an expert method for assessing the semantic similarity of text fragments based on the assessments of survey participants. The method consists of creating an interpretable semantic similarity scale derived from human perception of text content and used to analyze the consistency of various methods. To develop a “human” assessment, a survey of 138 participants was conducted. A comparative analysis revealed that various semantic similarity assessment methods demonstrate varying degrees of consistency with the human perception of text semantic similarity.

Keywords: semantic similarity, natural language processing, expert method, text comparison, rating scaling, questionnaire, Pearson correlation, Spearman correlation.

Conflict of interests. The authors declare that there is no conflict of interests.

For citation. Krez K., Shneiderov E., Shish P., Kondratenko E. (2026) Comparison of Methods for Assessing the Semantic Similarity of Text Fragments. *Doklady BGUIR*. 24 (2), 85–91. http://dx.doi.org/10.35596/1729-7648-2026-24-2-85-91 (in Russian).

Введение

Значительный прогресс в области анализа текста связан с развитием методов машинного обучения. Особую роль в этом процессе сыграли модели трансформерного типа, способные учитывать контекст употребления слов и выявлять скрытые смысловые связи между элементами текста. Одна из базовых моделей данного класса – BERT (Bidirectional Encoder Representations from Transformers), основанная на предварительном обучении двунаправленных контекстных представлений и ставшая базой для создания современных методов семантического сопоставления текстов [1]. На этой базе были разработаны подходы, ориентированные на получение векторных представлений предложений и коротких текстовых фрагментов, что существенно расширило возможности автоматической оценки их смысловой близости.

Вместе с тем результаты автоматических методов целесообразно рассматривать не изолированно, а в сопоставлении с человеческим восприятием смысла. В этой связи особый интерес представляет экспертная оценка семантического сходства, формируемая на основе субъективных суждений участников анкетирования. Такая оценка позволяет использовать человеческое восприятие смысловой близости в качестве интерпретируемого ориентира для анализа качества автоматических методов и степени согласованности их результатов с интуитивным пониманием содержания текста.

В статье проведен сравнительный анализ методов оценки семантического сходства текстовых фрагментов при их сопоставлении с экспертной оценкой, полученной в ходе анкетирования. В процессе исследования рассмотрены методы семантического сравнения на основе моделей SBERT (Sentence-BERT), SimCSE (Simple Contrastive Learning of Sentence Embeddings), LaBSE (Language-agnostic BERT Sentence Embedding), Word2Vec, FastText и TF-IDF (Term Frequency-Inverse Document Frequency). Проанализирована степень согласованности их результатов с человеческим восприятием смысловой близости текстов.

Обзор и сравнение методов семантического сходства текстовых фрагментов

Экспертный метод. На основе этого метода степень смысловой близости текстов определяется с учетом набора эмпирических признаков и экспертных предположений. В рамках проводимого исследования роль такого экспертного механизма выполняли участники анкетирования, которые, опираясь на собственное понимание содержания, языковую интуицию и субъективное восприятие смысла, оценивали, насколько близки между собой представленные текстовые фрагменты. Экспертный метод применяется в тех случаях, когда точное формальное описание семантического сходства затруднено или использование строгого алгоритма не позволяет получить быстрый интерпретируемый результат. Его задача состоит в том, чтобы на основе лексических, структурных и смысловых характеристик текста дать приближенную, но практически полезную оценку того, насколько два фрагмента близки по содержанию.

Экспертный метод оценки семантического сходства удобно рассматривать как функцию

$$Sim(T_1, T_2) \rightarrow [0, 1], \quad (1)$$

где *Sim* – численная мера смысловой близости T_1, T_2; T_1, T_2 – сравниваемые текстовые фрагменты.

Результат вычисления $Sim(T_1, T_2)$ принадлежит отрезку [0, 1]: значение, близкое к 0, соответствует отсутствию семантической связи между фрагментами, тогда как близкое к 1 указывает на их максимальную смысловую эквивалентность. Таким образом, функция $Sim(T_1, T_2)$ формализует задачу семантического сравнения, позволяя представить качественные смысловые различия в виде количественного показателя, пригодного для дальнейшего анализа, ранжирования и принятия решений в автоматизированных системах обработки текста.

С помощью экспертного метода оценки семантического сходства текстовых фрагментов был проведен опрос 138 участников. Им предлагалось оценить степень семантической близости между текстами, представленными в виде десяти пар. Оценивание осуществлялось по шкале от 1 до 5, где минимальные значения соответствовали низкому уровню смыслового сходства, а максимальные – высокому. Результаты анкетирования участников опроса представлены в табл. 1.

Таблица 1. Результаты анкетирования участников опроса
Table 1. Results of the survey participants' questionnaire

Номер пары текстовых фрагментов	Количество голосов за оценку				
	1	2	3	4	5
1	1	2	66	48	21
2	3	9	30	59	37
3	1	10	39	60	28
4	2	12	44	55	25
5	0	9	34	63	32
6	4	15	54	45	20
7	5	16	63	37	17
8	4	6	54	48	26
9	1	6	34	52	45
10	4	14	53	48	19
Всего голосов	25	99	471	515	270

При проведении опроса текст менялся по различным критериям:

– замена слов на синонимические эквиваленты с изменением стилистической маркированности;

– переход между стилями (художественный, публицистический, научный, официально деловой);

– добавление уточняющих компонентов, терминов и пояснений для повышения информативности;

– замена частных формулировок на более абстрактные и обобщенные смысловые конструкции;

– изменение структуры предложения (порядок слов, типы конструкций, дробление или объединение фраз);

– смещение коммуникативной установки текста (оценочность, эмоциональность, степень категоричности);

– усиление или снижение образности и выразительных средств (метафоризация/деэкспрессивизация).

При анализе данных, приведенных в табл. 1, можно сделать следующие выводы:

– участники воспринимают предложенные пары как семантически близкие: большинство оценок сосредоточено в диапазоне 3–5;

– наиболее частыми стали оценки 4 (515 голосов, 37,3 %) и 3 (471 голос, 34,1 %). Оценка 5 также встречается достаточно часто (270 голосов, 19,6 %);

– низкие оценки 1 и 2 составляют лишь небольшую долю (в сумме 124 голоса из 1380, т. е. 9 %), что указывает на сохранение смысла в большинстве случаев, даже несмотря на изменения формулировок.

Средняя итоговая оценка семантической близости по всем парам составила примерно 3,66 из пяти, что соответствует умеренно высокой степени сходства. Результаты подтверждают, что изменения текста по критериям (синонимизация, смена стиля, перестройка структуры, добавление уточнений и т. д.) чаще всего не разрушали смысл, а приводили к частичному смещению интерпретации, что и отражается в преобладании оценок 3 и 4. Это показывает, что участники устойчиво распознают общий смысл даже при заметных лексических и стилистических преобразованиях, однако степень сходства может снижаться, когда меняются эмоциональность, коммуникативная установка или уровень обобщенности.

Метод на основе модели SBERT. SBERT представляет собой модификацию архитектуры BERT, предназначенную для формирования семантически информативных векторных представлений предложений и коротких текстовых фрагментов [2]. В задаче сравнения двух фрагментов каждый из них независимо преобразуется в эмбеддинг фиксированной размерности, после чего

степень их семантической близости определяется на основе сходства между соответствующими векторами, как правило, с использованием косинусной меры. В отличие от подходов, ориентированных преимущественно на анализ отдельных токенов, SBERT позволяет получать целостные представления фрагментов текста, учитывающие контекст, смысловые связи и общую структуру текста. Вследствие этого модель применяется в задачах оценки семантического сходства, поиска близких по смыслу фрагментов и кластеризации текстов.

Метод на основе модели TF-IDF. Лексико-статистические методы основаны на представлении текста в виде набора лексических единиц и на вычислении их числовых весов. Одним из базовых методов такого представления текста является мера TF-IDF. В рамках данного метода каждый сравниваемый текстовый фрагмент преобразуется в вектор признаков, где компонентами выступают слова, а их значения определяются соответствующими весами. Мера TF показывает частоту появления слова в конкретном фрагменте, а IDF – степень его информативности, т. е. редкость слова в фрагментах текстов. После построения таких векторов сходство между двумя фрагментами оценивается с помощью косинусной меры: чем ближе направления векторов, тем выше степень их лексико-статистического сходства. Эффективность данного метода взвешивания терминов подробно изучена в работах по автоматическому поиску, включая исследования Дж. Салтона и К. Бакли [3]. Несмотря на практическую полезность, данный метод слабо учитывает порядок слов, контекст употребления и смысловые отношения между понятиями.

Метод на основе модели Word2Vec. Это классический метод распределенных векторных представлений слов, основанный на статистическом анализе их контекстного окружения. В задаче сопоставления двух текстовых фрагментов данный метод требует перехода от уровня слов к уровню предложений, что обычно реализуется путем усреднения векторов слов, входящих в каждый фрагмент. После построения таких усредненных представлений степень близости между фрагментами определяется с помощью косинусной меры. Преимущество Word2Vec – способность выявлять содержательное сходство и частично семантическую близость текстов. Однако отсутствие контекстной дифференциации значений многозначных слов и недостаточная чувствительность к синтаксическим преобразованиям ограничивают его применимость в задачах, связанных с выявлением семантического сходства.

Метод на основе модели SimCSE. SimCSE относится к классу методов получения эмбеддингов предложений на основе контрастивного обучения [4]. При сравнении двух текстовых фрагментов каждый из них кодируется в векторное пространство признаков, после чего между полученными представлениями вычисляется мера семантической близости. Особенность данного подхода состоит в том, что в процессе обучения формируется пространство, в котором семантически близкие тексты располагаются на меньшем расстоянии друг от друга, а несвязанные по смыслу фрагменты на большем. Это позволяет модели эффективно выявлять скрытую смысловую близость между текстами даже при отсутствии прямого лексического совпадения.

Метод на основе модели LaBSE. LaBSE является многоязычной трансформерной моделью, предназначенной для формирования унифицированных векторных представлений предложений на различных языках [5]. При сравнении двух текстовых фрагментов каждый из них проецируется в общее семантическое пространство, после чего близость между ними определяется посредством вычисления косинусного сходства между соответствующими векторами. Основное преимущество данной модели заключается в обеспечении сопоставимости смысловых представлений при различиях не только в формулировках и синтаксической организации текста, но и в языке его выражения. По этой причине LaBSE может использоваться как в монолингвальных, так и в кросс-языковых задачах семантического сравнения текстов.

Метод на основе модели FastText. FastText представляет собой развитие распределительных моделей векторных представлений слов и отличается учетом символьных *n*-грамм при построении эмбеддингов. При сравнении двух текстовых фрагментов сначала формируются векторные представления входящих в них слов, после чего для каждого фрагмента вычисляется агрегированное представление, например, путем усреднения векторов слов. Далее между полученными векторами определяется мера сходства. Учет символьных *n*-грамм позволяет модели частично отражать морфологические особенности языка и повышает устойчивость к редким словам и словоформам. Вместе с тем FastText, как и другие модели статических словарных представлений, ограниченно учитывает контекст употребления слов и порядок их следования, что снижает точность метода при сравнении фрагментов со сложной семантической структурой.

Методика и результаты сравнения методов семантического сходства текстовых фрагментов

Под семантическим сходством текстовых фрагментов понимается степень близости их смыслового содержания, отражающая, в какой мере два текста выражают одну и ту же мысль, описывают идентичные факты либо обладают эквивалентной семантикой при возможных различиях в формулировках. Существенная характеристика данного явления – его независимость от прямого лексического совпадения: тексты могут демонстрировать высокую степень семантической близости даже при использовании разных словоформ, грамматических конструкций и различного порядка слов.

Для оценки эффективности методов результаты анкетирования сопоставлялись с результатами работы алгоритмов. В качестве основы для сравнения использовалась субъективная оценка участников. Экспертная оценка семантического сходства, выраженная в процентах для каждой пары текстов, вычислялась по формуле

$$\text{Пара текста } (\%) = \frac{\sum_{i=1}^{5} in_i}{5N} \cdot 100 \ \%, \tag{2}$$

где n_i – число голосов за оценку; $N = 138$ – общее число респондентов, т. е. средняя оценка по шкале 1–5, нормированная к диапазону 0–100 %.

Расчет для каждой пары текстов приведен в табл. 2.

Таблица 2. Результаты, полученные различными методами при оценке семантического сходства текстовых пар
Table 2. Results obtained by different methods in assessing the semantic similarity of text pairs

Номер пары текстовых фрагментов	Результат расчета для метода, %						
	Анкетирование	SBERT	TF-IDF	Word2Vec	SimCSE	LaBSE	FastText
1	72,46	82,60	66,66	48,54	78,60	72,61	51,29
2	77,10	93,12	78,83	59,12	89,25	70,41	52,12
3	75,07	94,61	72,61	60,34	90,24	70,51	49,78
4	72,90	91,95	79,91	55,67	89,87	73,12	50,56
5	77,10	92,01	67,01	63,89	94,06	72,01	53,76
6	68,99	91,25	79,22	45,10	91,00	70,89	56,51
7	66,52	83,43	71,41	37,21	78,29	68,01	37,70
8	72,46	89,57	75,51	48,45	81,12	73,78	45,00
9	79,42	91,18	81,12	52,23	84,22	74,16	51,87
10	69,28	90,48	78,45	48,78	85,29	75,12	48,29
Среднее значение	73,13	90,42	75,67	51,35	86,19	72,46	49,69

Использование только средних значений семантической близости не позволяет сделать корректные выводы о качестве методов, поскольку они могут демонстрировать различную степень систематического завышения или занижения показателей, а также по-разному воспроизводить относительный порядок близости текстовых пар. Поэтому были приняты дополнительные статистические характеристики, такие как: коэффициент корреляции Пирсона (r), отражающий степень линейного соответствия между оценками метода и анкетирования; коэффициент ранговой корреляции Спирмена (ρ), характеризующий совпадение ранжирования пар по уровню смысловой близости; показатель MAE, позволяющий оценить среднее абсолютное отклонение методов от экспертного метода оценки семантического сходства в исходной шкале от 1 до 5. Сопоставление результатов методов оценки семантического сходства с субъективным человеческим восприятием, выявленным в ходе анкетирования, приведено в табл. 3.

Таблица 3. Сопоставление результатов методов оценки семантического сходства
Table 3. Comparison of the results of semantic similarity assessment methods

Метод	Статистическая характеристика		
	r	ρ	MAE
SBERT	0,535	0,585	0,840
TF-IDF	0,073	0,244	0,281
Word2Vec	0,798	0,817	1,060
SimCSE	0,403	0,341	0,653
LaBSE	0,273	0,146	0,163
FastText	0,509	0,439	1,172

При анализе табл. 3 можно отметить следующее:

– Word2Vec показывает максимальную согласованность с анкетированием по корреляции ($r = 0{,}798$; $\rho = 0{,}817$), но характеризуется большой ошибкой (MAE = 1,060);

– SBERT демонстрирует наиболее сбалансированный результат: умеренно высокая корреляция и приемлемая ошибка ($r = 0{,}535$; $\rho = 0{,}585$; MAE = 0,840);

– LaBSE имеет минимальное отклонение от анкетирования (MAE = 0,163), однако низкая корреляция указывает на слабую взаимосвязь пар;

– TF-IDF практически не согласуется с анкетированием по корреляции ($r = 0{,}073$), что подтверждает ограниченность этого метода;

– FastText и SimCSE показывают промежуточные результаты, но уступают SBERT по общей согласованности.

Заключение

1. Предложенные пары текстовых фрагментов в большинстве случаев воспринимаются участниками анкетирования как семантически близкие. Средняя итоговая экспертная оценка по всему набору пар составила 3,66 из пяти, что соответствует 73,13 %.

2. Сопоставление экспертных оценок с результатами методов выявило различия как в среднем уровне оценивания, так и в степени согласованности методов с человеческим восприятием. Установлено, что SBERT и SimCSE склонны к завышению степени семантического сходства: их средние значения соответственно составили 90,42 и 86,19 %. Напротив, Word2Vec и FastText продемонстрировали занижение оценок – 51,35 и 49,69 % соответственно. Наиболее близкие к экспертному уровню средние значения показали TF-IDF (75,67 %) и LaBSE (72,46 %), однако при последующем анализе отмечено, что такие значения не являются достаточным основанием для вывода об эффективности методов. Для более полной оценки методов семантического сходства текстовых фрагментов использовались коэффициенты корреляции Пирсона (r) и Спирмена (ρ), а также метрика MAE, выраженная в баллах исходной шкалы от 1 до 5. Так, метод семантического сравнения на основе модели TF-IDF, несмотря на достаточно близкое среднее значение, продемонстрировал крайне низкую согласованность с результатами анкетирования: $r = 0{,}073$, $\rho = 0{,}244$ и MAE = 0,281, что указывает на его слабую способность выявлять семантическое сходство при перефразировании и замене лексики. Наиболее высокую согласованность по степени семантической близости продемонстрировал Word2Vec ($r = 0{,}798$, $\rho = 0{,}817$), однако при этом характеризовался высокой абсолютной ошибкой 1,060, что связано с систематическим занижением оценок. Сбалансированные результаты продемонстрировал метод на основе модели SBERT, для которого получены значения $r = 0{,}535$, $\rho = 0{,}585$ и MAE = 0,840. Это позволяет сделать вывод о том, что именно данный метод наиболее устойчиво воспроизводит человеческое восприятие относительной смысловой близости текстовых фрагментов.

3. Исследование показало, что классические лексико-статистические методы недостаточно эффективны для оценки смысловой близости текстов в условиях перефразирования. Наиболее перспективными являются современные методы класса векторных представлений предложений, среди которых наиболее сбалансированные результаты продемонстрировал метод семантического сравнения на основе модели SBERT. При этом экспертный метод, учитывающий человеческое оценивание, выступает интерпретируемой основой для сопоставления автоматических оценок с человеческим восприятием смысла.

Список литературы / References

1. Devlin J., Chang M.-W., Lee K., Toutanova K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. 4171–4186. DOI: 10.18653/v1/N19-1423.
2. Reimers N., Gurevych I. (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. 3982–3992. DOI: 10.18653/v1/D19-1410.
3. Salton G., Buckley C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*. 24 (5), 513–523. DOI: 10.1016/0306-4573(88)90021-0.
4. Gao T., Yao X., Chen D. (2021) SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.
5. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. (2022) Language-Agnostic BERT Sentence Embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 878–891. DOI: 10.18653/v1/2022.acl-long.62.

Поступила 26.01.2026 Принята в печать 03.04.2026

Received: 26 January 2026 Accepted: 3 April 2026

Вклад авторов

Крез К. С. определила цели и задачи исследования с общей координацией проекта, провела анализ и систематизацию алгоритмов сравнения текстовой информации, подготовила рукопись статьи.

Шнейдеров Е. Н. осуществил научное руководство с изучением методологии, участвовал в утверждении окончательного варианта рукописи.

Шиш П. А., Кондратенко Е. В. выполнили сбор и обработку данных, анализ алгоритмических подходов, составили и оформили список литературы.

Authors' contribution

Krez K. defined the goals and objectives of the study with overall coordination of the project, conducted an analysis and systematization of algorithms for comparing text information, and prepared the manuscript of the article.

Shneiderov E. provided scientific supervision with the study of methodology and participated in the approval of the final version of the manuscript.

Shish P., Kondratenko E. collected and processed data, analyzed algorithmic approaches, compiled and formatted a list of references.

Сведения об авторах

Крез К. С., асп., ассист. каф. проектирования информационно-компьютерных систем, Белорусский государственный университет информатики и радиоэлектроники (БГУИР)

Шнейдеров Е. Н., канд. техн. наук., доц., каф. проектирования информационно-компьютерных систем, проректор по учебной работе, БГУИР

Шиш П. А., студент, БГУИР

Кондратенко Е. В., студент, БГУИР

Information about the authors

Krez K., Postgraduate, Assistant at the Department of Information and Computer Systems Design, Belarusian State University of Informatics and Radioelectronics (BSUIR)

Shneiderov E., Cand. Sci. (Tech.), Associate Professor at the Department of Information and Computer Systems Design, Vice-Rector for Academic Affairs, BSUIR

Shish P., Student, BSUIR

Kondratenko E., Student, BSUIR

Адрес для корреспонденции

220013, Республика Беларусь,
Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 29 952-75-56
E-mail: karinakrez04@gmail.com
Крез Карина Сергеевна

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 29 952-75-56
E-mail: karinakrez04@gmail.com
Krez Karina