УДК 004.912
DOI: 10.26102/2310-6018/2025.50.3.017

Мера семантической близости текстов

В.И. Шиян✉

Кубанский государственный университет, Краснодар, Российская Федерация

Резюме. В статье рассматривается задача автоматического определения семантической близости текстов, направленная на выявление первоисточников и заимствований в новостных материалах. Представлен двухфазный алгоритм, который на первом этапе использует косинусную меру близости для предварительной фильтрации текстов, а на втором – рассчитывает несимметричную взвешенную меру семантической близости с применением моделей RuBERT. Алгоритм осуществляет комплексный анализ текстов, учитывая их морфологические, синтаксические и семантические особенности, и демонстрирует устойчивость к типичным ошибкам, встречающимся в новостных материалах. Разработанный алгоритм включает этапы лингвистической обработки текстов, построения инвертированных индексов и расчета мер близости с использованием различных лингвистических признаков. Особое внимание уделяется обработке предложений: взвешиванию по TF-IDF, удалению дубликатов и анализу пересечений. Для оценки семантической близости предложений применяется система взвешенных показателей, учитывающих лексические, морфологические, синтаксические и семантические особенности. Экспериментальная часть работы направлена на определение оптимальных параметров алгоритма, таких как пороговые значения и весовые коэффициенты для различных лингвистических признаков. Результаты эксперимента показывают, что предложенный алгоритм эффективно выявляет заимствования, включая случаи значительной переработки текстов, с высокой полнотой на этапе фильтрации и повышенной точностью после семантического анализа. Алгоритм особенно полезен для автоматического формирования новостных обзоров и мониторинга заимствований в региональных СМИ.

Ключевые слова: семантическая близость, обработка текстов, нейросети, RuBERT, морфологический анализ, синтаксический анализ, семантический анализ, заимствования, первоисточник.

Для цитирования: Шиян В.И. Мера семантической близости текстов. *Моделирование, оптимизация и информационные технологии.* 2025;13(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1940 DOI: 10.26102/2310-6018/2025.50.3.017

A measure of semantic text similarity

V.I. Shiyan✉

Kuban State University, Krasnodar, the Russian Federation

Abstract. The article explores the task of automatically determining the semantic similarity of texts, aimed at identifying original sources and instances of borrowing in news materials. A two-phase algorithm is presented: the first stage employs cosine similarity for preliminary text filtering, while the second stage calculates an asymmetric weighted measure of semantic similarity using RuBERT models. The algorithm conducts a comprehensive analysis of texts, taking into account their morphological, syntactic, and semantic features, and demonstrates robustness against typical errors found in news materials. The developed algorithm includes stages of linguistic text processing, inverted index construction, and similarity calculation using various linguistic features. Special attention is given to sentence processing: TF-IDF weighting, duplicate removal, and intersection analysis. To assess the semantic similarity of sentences, a weighted scoring system is applied, incorporating lexical, morphological, syntactic, and semantic characteristics. The experimental part of the study focuses on determining the algorithm's optimal parameters, such as threshold values and weight coefficients for different linguistic features. The results demonstrate that the proposed algorithm effectively detects

borrowings, including cases of substantial text modifications, achieving high recall at the filtering stage and improved precision after semantic analysis. The algorithm is particularly useful for automated news digest generation and monitoring text reuse in regional media.

Keywords: semantic similarity, text processing, neural networks, RuBERT, morphological analysis, syntactic analysis, semantic analysis, borrowings, original source.

For citation: Shiyan V.I. A Measure of Semantic Text Similarity. *Modeling, Optimization and Information Technology*. 2025;13(3). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1940 DOI: 10.26102/2310-6018/2025.50.3.017

Введение

В статье рассматривается задача автоматического определения количественной меры семантической близости текстов, представленных в электронном виде. Данная задача решается в рамках общей задачи автоматического формирования новостного обзора региональными СМИ на основе новостных рубрик центральных СМИ. Проблемными факторами получения такого обзора являются вероятные ошибки в текстах и множественные заимствования электронными СМИ содержательного контента у своих конкурентов, а также случайные пересечения содержания независимых друг от друга статей, освещающих одну тему.

Входными данными служат электронные новостные тексты различных СМИ. Нечувствительность к грамматическим ошибкам новостных текстов обеспечивает использование нейросетей [1] на этапе подготовки данных [2]. Предложенный алгоритм выполняет фильтрацию [3] коллекции новостных текстов и выявляет первоисточники и заимствования на основе вычисления меры семантической близости каждой пары текстов в коллекции.

Авторский вклад состоит в использовании моделей RuBERT для лингвистического разбора текстов и несимметричной взвешенной меры семантической близости пары текстов.

Материалы и методы

Алгоритмические примитивы и обозначения. В алгоритме вычисления меры семантической близости текстов использованы следующие отображения и алгоритмические примитивы:

1) $T \to W$ – разбиение текста T на множество слов W;

2) $T \to S$ – разбиение текста T на множество предложений S;

3) $\delta: W \to D$ – отображение множества W на множество нормальных форм D, заключающееся, по сути, в лемматизации текста;

4) $\rho: W \to POS$ – отображение множества W на множество частей речи POS;

5) $\mu: W \to Morph$ – отображение множества W на множество $Morph$, ставящее в соответствие каждому слову w_i его морфологические признаки $Morph_i$;

6) $syn: S \to (W^2 \to SR)$ – отображение, ставящее в соответствие каждому предложению $s_k \in S$ множество таких упорядоченных пар слов (w_i, w_j) предложения s_k, которым соответствует синтаксическое отношение из конечного множества синтаксических отношений SR. Первый компонент в паре (w_i, w_j) считается главным, второй – зависимым;

7) $sem: S \to ((W \times Roles)^2 \times SR \to \Omega)$ – отображение, ставящее в соответствие каждому предложению $s_k \in S$ множество таких упорядоченных пар слов (w_i, w_j) предложения s_k, которым соответствует семантическая связь из конечного множества семантических связей Ω. Каждый компонент из пары (w_i, w_j) характеризуется

семантической ролью из конечного множества семантических ролей $Roles$, а сами компоненты связаны синтаксическим отношением из конечного множества SR.

Подготовка данных для определения семантической близости текстов. Входные тексты преобразуются в числовые векторы с помощью моделей, основанных на архитектуре BERT (англ. Bidirectional Encoder Representations from Transformers) и адаптированных для работы с русскоязычными текстами: RuBERT-tiny, RuBERT-tiny2 и RuBERT-base-cased. Указанные варианты нейросетей RuBERT имеют собственные показатели точности и времени работы, связанные соотношением: чем выше точность, тем дольше время работы. Обоснование и выбор конкретной модели выходит за рамки данной статьи. Стоит отметить, что выбранные модели RuBERT имеют техническое ограничение на длину входного текста в 512 токенов, что составляет примерно 2500 символов. Для обработки новостных материалов, превышающих этот лимит, использовалось разделение на смысловые блоки, с последующей агрегацией эмбеддингов. Такой подход позволил минимизировать потерю информации при сохранении исходной архитектуры алгоритма.

Предварительный этап. На этом этапе выполняется однократное обучение модели RuBERT. В дальнейшей работе алгоритма дообучение модели не предполагается, но допускается. Для обучения нейросети использовались тексты с разметкой данных в формате CoNLL-U [4] с 17 частями речи, 15 морфологическими признаками, 41 синтаксическим отношением, 133 семантическими связями и 860 семантическими ролями. На этом наборе данных обучались модели преобразования текста в числовые векторы – эмбеддинги [5, 6].

Первый этап. Лингвистическая обработка текста обученной нейросетью одновременно выполняет:

1) морфологический анализ – определение нормальных форм D, частей речи POS и морфологических признаков слов $Morph$ [7];

2) синтаксический анализ – выявление синтаксических отношений SR между словами;

3) семантический анализ – определение семантических ролей $Roles$ и связей Ω между словами [8].

Результатом первого этапа является вектор $(D, POS, Morph, SR, Roles, \Omega)$, описывающий необходимые для дальнейшей работы параметры текста. Полезным свойством этого этапа является адаптация к вероятным ошибкам в новостных текстах и одновременное выполнение трех указанных видов анализа текста, что сокращает время работы нейросети в сравнении с последовательным их выполнением отдельными нейросетями.

Второй этап. На основе векторов текстов $(D, POS, Morph, SR, Roles, \Omega)$ лингвистический препроцессор создает инвертированный индекс [9], который хранит следующую информацию:

1) номер текста, в котором встречается слово;

2) номер предложения в тексте;

3) номер слова в предложении;

4) слово w_i;

5) нормальная форма слова D_i, часть речи POS_i, морфологические признаки $Morph_i$, синтаксическое отношение SR_i, семантическая связь Ω_i и семантическая роль слова $Roles_i$ [10];

6) частота леммы слова в тексте.

В целом подготовка данных заключается в двухэтапной обработке каждого текста из коллекции новостных текстов (Рисунок 1). Однократное обучение отмечено пунктирной стрелкой. На первом этапе модель RuBERT строит векторы входных

новостных текстов, на втором этапе лингвистический препроцессор создает инвертированный индекс каждого текста.

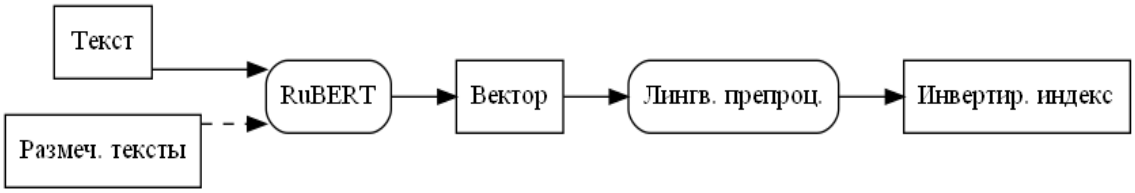

Рисунок 1 – Двухэтапная обработка текста
Figure 1 – Two-stage text processing

Общее описание алгоритма вычисления меры близости. Алгоритм содержит две последовательно выполняемые фазы (Рисунок 2):

1) фильтрация коллекции инвертированных индексов текстов на основании косинусной меры взаимной близости и отбор только тех пар, для которых значение этой меры превышает некоторый порог;

2) выявление первоисточника и заимствований в фильтрованной коллекции на основе несимметричной взвешенной меры близости каждой пары текстов [11].

Рисунок 2 – Двухфазный алгоритм вычисления меры близости
Figure 2 – Two-phase algorithm for similarity measure calculation

Фильтрация коллекции входных текстов. В фазе фильтрации каждый текст из коллекции входных текстов играет роль кандидата в первоисточник. Фильтрацию проходят те тексты, которые имеют непустой общий словарь ключевых слов.

Пусть задан текст $t_1 \in C$ – кандидат в первоисточники и $t_2 \in C$ – текст, близость которого к первоисточнику требуется оценить, C – коллекция текстов.

Для каждого текста t из коллекции C следует:

1) определить информационную значимость $u(w,t,C)$ слова w с помощью TF-IDF и выделить множество ключевых слов $\widetilde{W}(t)$, состоящее из слов с наибольшим TF-IDF весом. Текст t представляется набором слов, каждое из которых приводится к нормальной форме – лемме. Затем слова сортируются по убыванию TF-IDF веса и из них формируется множество ключевых слов. Число ключевых слов n определяется как 45 % от общего числа уникальных слов в тексте, но не более 120.

Информационная значимость слова вычисляется по формуле:

$$u(w,t,C) = TF(w,t) \cdot IDF(w,C), \quad (1)$$

в которой

$$TF(w,t) = log_{|t|+1}(k(w,t)+1), \quad (2)$$

$$IDF(w,C) = \begin{cases} log_{|C|}\left(\frac{|C|}{m(w,C)}\right), m(w,C) \neq 0 \\ 1, \text{ в противном случае} \end{cases}, \quad (3)$$

где $TF(w,t)$ – числовая функция, определяющая вес каждого слова, $0 \leq TF(w,t) \leq 1$; $k(w,t)$ – количество слов w в тексте t [12, 13]; $IDF(w,C)$ – числовая функция,

определяющая вес каждого слова, $0 \le IDF(w,C) \le 1$; $m(w,C) = |\{t \in C | w \in t\}|$ – количество текстов t, содержащих слово w [14].

Таким образом, каждому тексту соответствует вектор

$$\vec{u}(t) = \big(u(w_1,t,C), u(w_2,t,C), \dots, u(w_n,t,C)\big), \quad (4)$$

определяющий информационную значимость слова;

2) отправить запрос в инвертированный индекс для поиска текстов, содержащих максимальное количество таких ключевых слов;

3) рассчитать меру близости текстов t_1 и t_2 по косинусной мере:

$$SIM_{cos}(t_1,t_2) = \frac{\sum_{w \in \widetilde{W}(t_1) \cup \widetilde{W}(t_2)} u(w,t_1,C) \cdot u(w,t_2,C)}{\sqrt{\sum_{w \in \widetilde{W}(t_1) \cup \widetilde{W}(t_2)} \big(u(w,t_1,C)\big)^2} \cdot \sqrt{\sum_{w \in \widetilde{W}(t_1) \cup \widetilde{W}(t_2)} \big(u(w,t_2,C)\big)^2}} =$$

$$= \frac{\sum_{w \in \widetilde{W}(t_1) \cap \widetilde{W}(t_2)} u(w,t_1,C) \cdot u(w,t_2,C)}{\sqrt{\sum_{w \in \widetilde{W}(t_1)} \big(u(w,t_1,C)\big)^2} \cdot \sqrt{\sum_{w \in \widetilde{W}(t_2)} \big(u(w,t_2,C)\big)^2}}, \quad (5)$$

где $SIM_{cos}(t_1,t_2)$ – числовая функция, определяющая меру близости текстов, $0 \le SIM_{cos}(t_1,t_2) \le 1$;

4) текст t_1 становится первоисточником, а t_2 – зависимым, если косинусная мера взаимной близости превышает 0,135. Данное пороговое значение было получено экспериментальным путем.

Вычислительная сложность процедуры фильтрации $O(|C|^2 \cdot n)$.

Определение меры семантической близости. Для расчета близости текстов t_1 и t_2 используются множества слов W_1 и W_2, а также множества предложений S_1 и S_2, соответствующие этим текстам.

Для каждого предложения из текста t_2 следует:

1) рассчитать вес предложения на основе TF-IDF весов слов [15];

2) исключить предложения с весом ниже 0,01, а также слишком короткие предложения, содержащие менее K слов. Пороговые значения веса и минимальной длины предложений $K = 3$ были определены экспериментально;

3) удалить дубликаты предложений;

4) для каждого оставшегося предложения из текста t_2:

а) представить предложение как вектор уникальных чисел, соответствующих леммам слов;

б) найти пересечение этого вектора с векторами предложений из текста t_1;

в) оставить только те пары предложений $s_1 \in S_1$ и $s_2 \in S_2$, которые имеют не менее $M\%$ слов. Экспериментальным путем было установлено оптимальное значение $M = 30$, то есть пара предложений включается в дальнейший анализ только в том случае, если они имеют не менее 30 % общих слов.

5) для оставшихся пар предложений следует:

а) рассчитать *меру близости IDF весов слов*

$$\widetilde{w}_1(s_1,s_2) = \sum_{(w_1,w_2) \in N(s_1,s_2)} v(w_1,C), \quad (6)$$

где $N(s_1,s_2) = \{(w_1,w_2) \in W_1 \times W_2 | \exists w_1 \in s_1 \wedge \exists w_2 \in s_2 : \delta(w_1) = \delta(w_2)\}$ – множество пар слов с одинаковой леммой, где первое слово $w_1 \in s_1$ взято из $s_1 \in S_1$, а второе $w_2 \in s_2$ – из $s_2 \in S_2$; $v(w_1,C)$ – вес слова w_1.

Из условия нормирования весов слов текста t

$$\sum_{w \in W_1} v(w,C) = 1, \quad (7)$$

определить нормированный вес $v(w_1,C)$:

$$v(w_1, C) = \frac{1}{\sum_{w \in W_1} IDF(w,C)} \cdot IDF(w_1, C). \quad (8)$$

Очевидно, что выполняется следующее утверждение: $0 \leq \widetilde{w}_1(s_1, s_2) \leq 1$.

б) рассчитать *меру близости TF-IDF весов слов*

$$\widetilde{w}_2(s_1, s_2) = \sum_{(w_1,w_2) \in N(s_1,s_2)} f(w_1, w_2) \cdot v(w_1, C) \cdot TF(w_2, t_2), \quad (9)$$

где $TF(w_2, t_2)$ – TF вес слова w_2, $f(w_1, w_2)$ – своего рода «штраф» за несовпадение частей речи и морфологических признаков слов w_1, w_2, основанный на мере Жаккара:

$$f(w_1, w_2) = \frac{f_1(w_1,w_2) + |\mu(w_1) \cap \mu(w_2)|}{f_2(w_1,w_2) + |\mu(w_1) \cup \mu(w_2)|}, \quad (10)$$

где $\mu(w)$ – множество пар, где каждая пара состоит из названия морфологического признака и соответствующего значения этого признака для слова w, а

$$f_1(w_1, w_2) = \begin{cases} 1, \rho(w_1) = \rho(w_2) \\ 0, \text{в противном случае'} \end{cases} \quad (11)$$

$$f_2(w_1, w_2) = \begin{cases} 1, \rho(w_1) = \rho(w_2) \\ 2, \text{в противном случае'} \end{cases} \quad (12)$$

где $\rho(w)$ – часть речи слова w.

Если знаменатель равен нулю, то $f(w_1, w_2) = 0$.

Очевидно, что выполняется следующее утверждение: $0 \leq \widetilde{w}_2(s_1, s_2) \leq 1$.

в) Рассчитать *меру близости синтаксических отношений слов*

$$\widetilde{w}_3(s_1, s_2) = \frac{\sum_{(w_h,\sigma,w_d) \in Synt(s_1) \cap Synt(s_2)} v(w_h, C)}{\sum_{s \in S_1} \left(\sum_{(w_h,\sigma,w_d) \in Synt(s)} v(w_h, C) \right)}, \quad (13)$$

где $Synt(s)$ – множество троек (w_h, σ, w_d), w_h – главное слово, w_d – зависимое слово, σ – тип синтаксического отношения между ними.

Очевидно, что выполняется следующее утверждение: $0 \leq \widetilde{w}_3(s_1, s_2) \leq 1$.

Если знаменатель равен нулю, то $\widetilde{w}_3(s_1, s_2) = 0$.

г) рассчитать *меру близости семантических связей слов*

$$\widetilde{w}_4(s_1, s_2) = \frac{\sum_{(w_h,\eta,w_d) \in SemRels(s_1) \cap SemRels(s_2)} v(w_h, C)}{\sum_{s \in S_1} \left(\sum_{(w_h,\eta,w_d) \in SemRels(s)} v(w_h, C) \right)}, \quad (14)$$

где $SemRels(s)$ – множество троек (w_h, η, w_d), w_h – главное слово, w_d – зависимое слово, η – тип семантической связи между ними.

Очевидно, что выполняется следующее утверждение: $0 \leq \widetilde{w}_4(s_1, s_2) \leq 1$.

Если знаменатель равен нулю, то $\widetilde{w}_4(s_1, s_2) = 0$.

д) рассчитать *меру близости семантических ролей слов*

$$\widetilde{w}_5(s_1, s_2) = \frac{|SemRoles(s_1) \cap SemRoles(s_2)|}{\sum_{s \in S_1} |SemRoles(s)|}, \quad (15)$$

где $SemRoles(s)$ – множество, содержащее пары (w, ρ), w – лемма слова из предложения s с определенной семантической ролью ρ.

Очевидно, что выполняется следующее утверждение: $0 \leq \widetilde{w}_5(s_1, s_2) \leq 1$.

Если знаменатель равен нулю, то $\widetilde{w}_5(s_1, s_2) = 0$.

е) рассчитать несимметричную *меру семантической близости предложений* в виде взвешенной суммы

$$Prox(s_1, s_2) = \sum_{i=1}^{5} p_i \cdot \widetilde{w}_i(s_1, s_2), \quad (16)$$

$$\sum_{i=1}^{5} p_i = 1, \quad (17)$$

$$0 \leq p_i \leq 1, \quad (18)$$

где $p_i, i = \overline{1,5}$ определяют относительный вклад каждой из вычисленных мер близости [16].

Очевидно, что выполняется следующее утверждение: $0 \leq Prox(s_1, s_2) \leq 1$.

Весовой коэффициент $p_i = 0$ означает, что *i*-я мера близости не учитывается при расчете близости предложений. Иными словами, вклад этой меры полностью исключается из рассмотрения.

В случае $p_i = 1$ только *i*-я мера близости используется для расчета близости предложений, а все остальные меры не влияют на результат, поскольку их веса равны нулю ($p_j = 0$ для $j \neq i$).

В большинстве практических ситуаций веса p_i принимают значения на отрезке $[0;\ 1]$, отражая относительную важность каждой меры близости. Настраивая значения p_i, можно увеличить или уменьшить влияние конкретных мер на итоговый результат.

Экспериментальные исследования позволили определить оптимальные значения весовых коэффициентов для расчета меры семантической близости: $p_1 = 0{,}294$, $p_2 = 0{,}06$, $p_3 = 0{,}198$, $p_4 = 0{,}15$, $p_5 = 0{,}298$. Эти коэффициенты отражают относительный вклад различных лингвистических характеристик – от лексических до семантических – в итоговую оценку близости текстов. Они были получены путем многократного тестирования на корпусе с разным уровнем переработки заимствованных фрагментов.

Величина, определяющая наибольшую меру близости предложений сопоставляемого текста к предложению эталонного текста s_1, задается следующей формулой:

$$Y(s_1, t_2) = \max_{s_2 \in S_2} \{Prox(s_1, s_2)\}, \quad (19)$$

где S_2 – множество предложений текста t_2.

Очевидно, что выполняется следующее утверждение: $0 \leq Y(s_1, t_2) \leq 1$.

Несимметричная *мера семантической близости текста* t_2 к первоисточнику t_1 выражается следующей формулой:

$$X(t_1, t_2) = \sum_{s_1 \in S_1} Y(s_1, t_2), \quad (20)$$

где S_1 – множество предложений первоисточника t_1.

Очевидно, что выполняется следующее утверждение: $0 \leq X(t_1, t_2) \leq 1$.

На основе проведенных экспериментов было установлено, что тексты целесообразно считать семантически близкими, то есть содержащими признаки заимствования, если итоговая мера семантического расстояния между ними превышает пороговое значение 0,07. Это значение порога было определено эмпирически как наилучшее для обеспечения баланса между полнотой и точностью при выявлении зависимых текстов в условиях различной степени их переработки.

Вычислительная сложность определения меры семантической близости $O\big(|S_1| \cdot |S_2| \cdot (w \cdot f + r)\big)$, где $|S_i|$ – количество предложений в соответствующем тексте, w – количество слов в предложении, f – количество морфологических признаков, r – количество синтаксических отношений и семантических связей.

Результаты и обсуждение

Экспериментальная часть исследования была направлена на проверку эффективности предложенного алгоритма выявления заимствований на основе меры семантической близости текстов. Для этого использовался корпус ParaPlag v2, специально предназначенный для оценки качества методов обнаружения текстовых заимствований. Этот корпус включает тексты с заимствованиями различной степени

сложности обнаружения и охватывает широкий спектр способов переработки заимствованного материала. Подобный выбор корпуса позволил обеспечить надежную проверку алгоритма в условиях, приближенных к реальным задачам анализа новостных текстов (Рисунок 3).

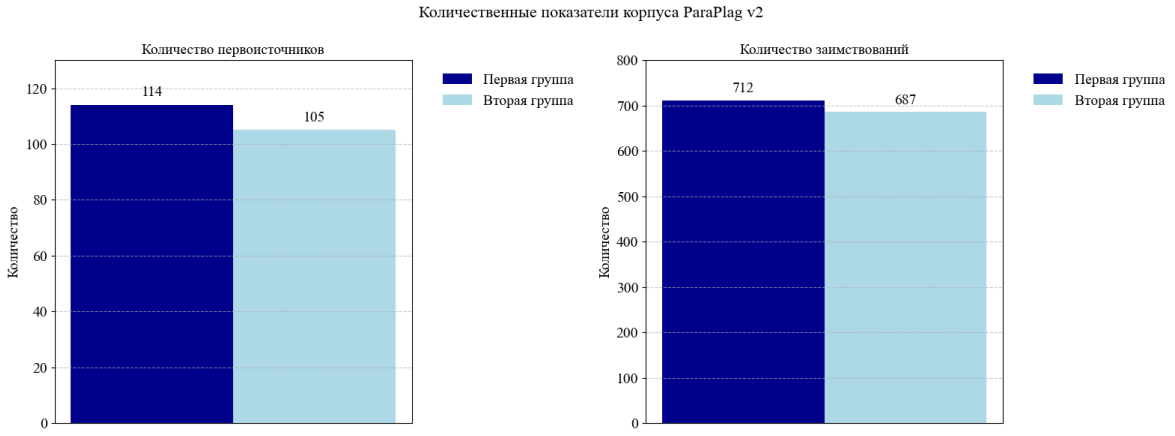

Рисунок 3 – Количественные показатели
Figure 3 – Quantitative indicators

Тексты корпуса были разделены на две основные группы в зависимости от особенностей заимствования и правил их создания. Первая группа содержала тексты, в которых допускалось использование дословных заимствований, что обеспечивало относительно легкое выявление фрагментов заимствованного материала. Эти тексты в среднем состояли из 150 предложений и характеризовались высоким процентом прямых или слабо измененных заимствований. Вторая группа включала тексты с исключительно модифицированными заимствованиями, где использовались сложные техники скрытия заимствованного материала, такие как объединение или разделение предложений, перестановка и замена слов, а также введение дополнительных слов. Средний размер текстов этой группы составлял около 100 предложений, что дополнительно усложняло задачу выявления заимствований (Рисунок 4).

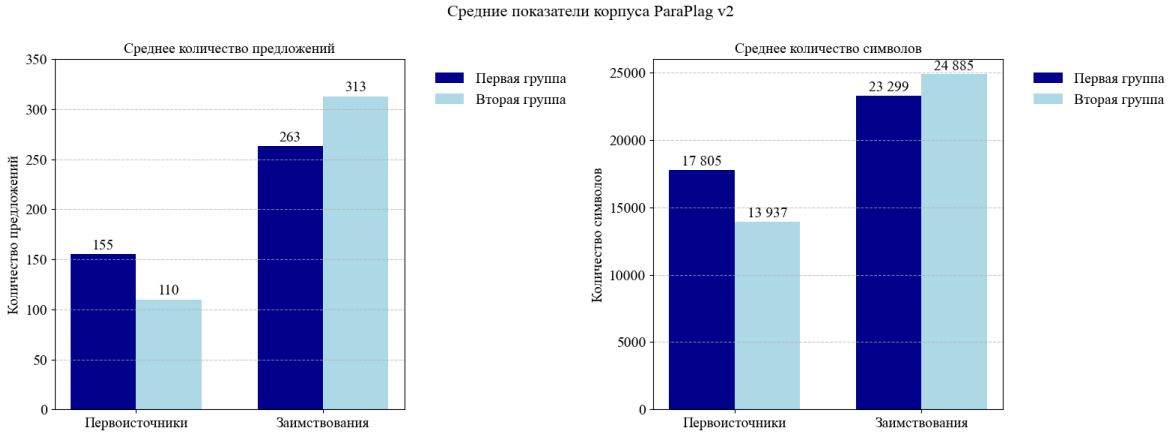

Рисунок 4 – Средние показатели
Figure 4 – Average indicators

На первом этапе эксперимента выполнялась фаза фильтрации, целью которой было максимально полное выявление возможных заимствований. Результаты показали

высокую полноту как для первой, так и для второй группы текстов: 0,983 и 0,969 соответственно. Это свидетельствует о том, что алгоритм на данном этапе практически не упустил ни одного действительно заимствованного фрагмента. Однако точность фильтрации оказалась сравнительно низкой: 0,149 для первой группы и 0,118 для второй. Такой результат объясняется тем, что на этапе фильтрации сознательно не применялись механизмы отсечения ложных совпадений, поскольку задача заключалась в том, чтобы на этом этапе не пропустить ни одного потенциального заимствования. Для улучшения точности на данном этапе можно рассмотреть возможность интеграции более жестких критериев предварительной оценки значимости совпадений, например, учета контекста ключевых слов или начального анализа синтаксических структур.

Второй этап эксперимента был направлен на проведение детального семантического анализа с использованием несимметричной взвешенной меры близости, учитывающей морфологические, синтаксические и семантические признаки. Для первой группы текстов полнота составила 0,97, а точность повысилась до 0,754, что отражает высокую способность алгоритма правильно выявлять заимствования при относительно небольшом числе ложных срабатываний. Во второй группе полнота составила 0,82, а точность достигла 0,709. Снижение полноты по сравнению с первой группой объясняется большей степенью переработки заимствованного материала, что затрудняет идентификацию фрагментов даже при учете сложных лингвистических признаков. Результаты также указывают на то, что второму этапу удалось существенно сократить количество ошибочно определенных пар заимствований за счет использования глубокой лингвистической обработки текстов (Рисунок 5).

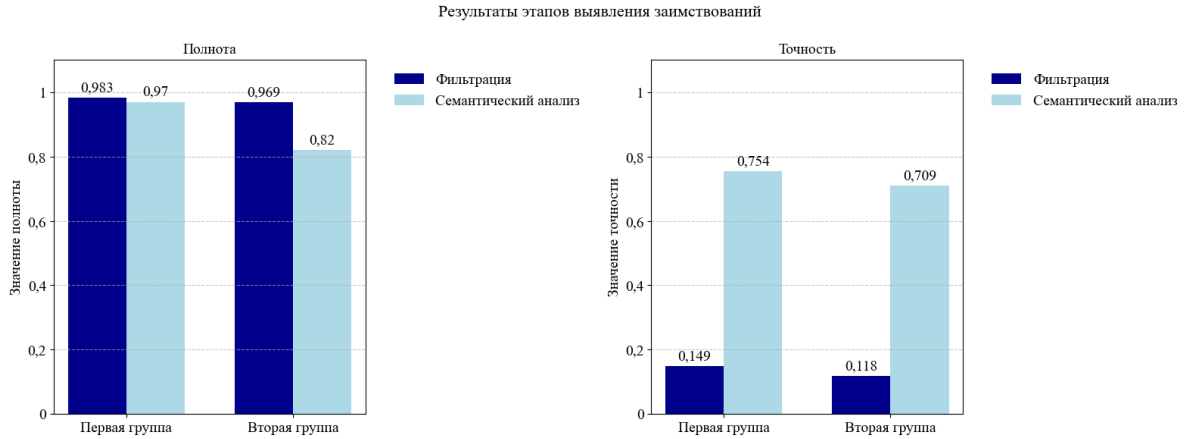

Рисунок 5 – Результаты этапов
Figure 5 – Stage results

Анализ полученных данных позволяет сделать вывод о том, что предложенный алгоритм особенно эффективен при работе с текстами, содержащими явно выраженные или умеренно измененные заимствования. В случае текстов с сильно модифицированными заимствованиями качество работы алгоритма остается удовлетворительным, однако для дальнейшего повышения полноты рекомендуется усилить вклад синтаксических и семантических мер за счет увеличения соответствующих весовых коэффициентов. Кроме того, целесообразно рассмотреть возможность включения дополнительных признаков, таких как анализ метафорических конструкций или устойчивых выражений, а также расширение обучающей выборки для дообучения используемых моделей RuBERT на примерах сложных заимствований.

Общее улучшение качества работы алгоритма при переходе от этапа фильтрации к этапу семантического анализа подтверждается увеличением точности и снижением числа ложных срабатываний. Этот результат подчеркивает значимость использования именно комплексной меры семантической близости для задач выявления заимствований в текстах с различной степенью переработки исходного материала. Итоги эксперимента демонстрируют практическую применимость разработанного алгоритма для автоматического мониторинга текстовых заимствований в электронных СМИ и его потенциал для адаптации к другим типам текстов.

Заключение

В статье представлен двухфазный алгоритм автоматического определения семантической близости текстов, состоящий из предварительной фильтрации на основе косинусной меры и последующего применения несимметричной взвешенной меры с использованием моделей RuBERT. Предложенный алгоритм демонстрирует устойчивость к типичным ошибкам в текстах новостного характера благодаря комплексному учету морфологических, синтаксических и семантических признаков. В ходе экспериментов были найдены оптимальные пороговые значения и весовые коэффициенты, учитывающие вклад различных лингвистических характеристик. Проведенные эксперименты подтвердили, что использование несимметричной меры семантической близости существенно повышает точность обнаружения заимствований и позволяет эффективно фильтровать ложные совпадения, что особенно важно для практических задач мониторинга текстов.

Практическая значимость алгоритма подтверждается экспериментальным тестированием на корпусе текстов с разными уровнями заимствований. Использование семантического анализа на второй фазе позволило снизить число ложных срабатываний и повысить точность выявления зависимых текстовых пар, что делает предложенный алгоритм эффективным инструментом для автоматического формирования новостных обзоров и выявления заимствований в региональных СМИ.

Перспективные направления дальнейших исследований включают расширение функционала алгоритма для работы с другими типами текстов, например, научными или художественными, оптимизацию вычислительных процессов и интеграцию дополнительных лингвистических признаков. Также актуальной задачей остается увеличение объема тестовых данных для более точной настройки параметров и повышения общей надежности предлагаемого алгоритма.

Кроме того, актуальным направлением дальнейших исследований является сравнительный анализ эффективности RuBERT с другими современными моделями, включая XLM-R, LaBSE и Sentence-BERT. Особое значение такой анализ имеет для задач обработки длинных текстов и выявления сложных заимствований. Проведение подобного сравнения поможет уточнить оптимальный выбор архитектуры модели и достичь лучшего баланса между точностью и скоростью работы алгоритма.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Николенко С., Кадурин А., Архангельская Е. *Глубокое обучение*. Санкт-Петербург: Питер; 2018. 480 с.
2. Feldman R., Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press; 2007. 410 p.
3. Tseng Yu.-H., Lin Ch.-J., Lin Yu-I. Text Mining Techniques for Patent Analysis. *Information Processing & Management*. 2007;43(5):1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

4. Ефименко И.В. Обработка естественно-языковых текстов: онтологичность в лингвистике и дискурсивность в извлечении знаний. В сборнике: *КИИ-2006: десятая национальная конференция по искусственному интеллекту с международным участием: труды конференции: Том 2, 25–28 сентября 2006 года, Обнинск, Россия*. Москва: Физматлит; 2006. С. 230–234.
5. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv. URL: https://arxiv.org/abs/1301.3781 [Accessed 28th March 2025].
6. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: *NIPS 2013: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, 05–08 December 2013, Lake Tahoe, NV, USA*. 2013. P. 3111–3119.
7. Dobrov B.V., Loukachevitch N.V. Multiple Evidence for Term Extraction in Broad Domains. In: *RANLP 2011: Recent Advances in Natural Language Processing, 12–14 September 2011, Hissar, Bulgaria*. Association for Computational Linguistics; 2011. P. 710–715.
8. Delgado M., Martín-Bautista M.J., Sánchez D., Vila M.A. Mining Text Data: Special Features and Patterns. In: *Pattern Detection and Discovery: ESF Exploratory Workshop, 16–19 September 2002, London, UK*. Berlin, Heidelberg: Springer; 2002. P. 140–153. https://doi.org/10.1007/3-540-45728-3_11
9. Hu K., Wu H., Qi K., et al. A Domain Keyword Analysis Approach Extending Term Frequency-Keyword Active Index with Google Word2vec Model. *Scientometrics*. 2018;114(3):1031–1068. https://doi.org/10.1007/s11192-017-2574-9
10. Cruse D.A. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press; 2011. 497 p.
11. Соченков И.В. Метод сравнения текстов для решения поисково-аналитических задач. *Искусственный интеллект и принятие решений*. 2013;(2):32–43.
Sochenkov I.V. Text Comparison Method for a Search and Analytical Engine. *Artificial Intelligence and Decision Making*. 2013;(2):32–43. (In Russ.).
12. Salton G., Buckley Ch. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*. 1988;24(5):513–523. https://doi.org/10.1016/0306-4573(88)90021-0
13. Luhn H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*. 1957;1(4):309–317. https://doi.org/10.1147/rd.14.0309
14. Jones K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*. 1972;28(1):11–21. https://doi.org/10.1108/eb026526
15. Jing L.-P., Huang H.-K., Shi H.-B. Improved Feature Selection Approach TFIDF in Text Mining. In: *2002 International Conference on Machine Learning and Cybernetics, 04–05 November 2002, Beijing, China*. IEEE; 2002. P. 944–946. https://doi.org/10.1109/ICMLC.2002.1174522
16. Zubarev D.V., Sochenkov I.V. Paraphrased Plagiarism Detection Using Sentence Similarity. In: *Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог», 31 May – 03 June 2017, Moscow, Russia*. Moscow: Russian State University for the Humanities; 2017. P. 399–408.

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Шиян Валерий Игоревич, старший преподаватель, Кубанский государственный университет, Краснодар, Российская Федерация.
e-mail: shiyanvi@yandex.ru

Valery I. Shiyan, Senior Lecturer, Kuban State University, Krasnodar, the Russian Federation.

Статья поступила в редакцию 01.05.2025; одобрена после рецензирования 10.07.2025; принята к публикации 14.07.2025.

The article was submitted 01.05.2025; approved after reviewing 10.07.2025; accepted for publication 14.07.2025.