



DOI: <https://doi.org/10.15688/mpcm.jvolsu.2026.1.4>

УДК 004.8
ББК 32.973

Дата поступления статьи: 21.12.2025
Дата принятия статьи: 20.02.2026

ИНТЕРПРЕТАЦИЯ ВНУТРЕННИХ ПРЕДСТАВЛЕНИЙ BERT ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ ТЕКСТОВ, СОДЕРЖАЩИХ РАССУЖДЕНИЕ

Александр Владимирович Каныгин

Аспирант кафедры компьютерных наук и экспериментальной математики,
Волгоградский государственный университет
a.kanygin99@gmail.com
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Цитирование: Каныгин А. В. Интерпретация внутренних представлений BERT для задачи классификации текстов, содержащих рассуждение // Математическая физика и компьютерное моделирование. 2026. Т. 29, № 1. С. 38–57.

DOI: <https://doi.org/10.15688/mpcm.jvolsu.2026.1.4>

Citation: Kanygin A.V. Interpretation of BERT Internal Representations for the Classification of Texts Containing Reasoning. *Mathematical Physics and Computer Simulation*, 2026, vol. 29, no. 1, pp. 38-57.

DOI: <https://doi.org/10.15688/mpcm.jvolsu.2026.1.4>

Аннотация. В статье исследуется задача бинарной классификации текстов на наличие в них рассуждений (логических связей, аргументации, причинно-следственных отношений). Данная задача была решена при помощи fine-tuning модели ruBert с более высоким качеством, по сравнению с классическими ML-подходами (TF-IDF, стэкинг моделей, статические векторы вложения + LSTM-классификатор). Полученные результаты послужили мотивацией для анализа внутренних представлений модели, направленных на выявление признаков и структурных элементов текста, которые определяют высокую эффективность BERT. В статье для анализа применена совокупность методик: визуализация [CLS]-векторов с использованием снижения размерности до 2D через UMAP, вычисление Silhouette score для каждого из 12 слоев BERT-модели, отдельные классификаторы для всех скрытых слоев модели, а также ablation study, которое оценивает, как удаление ключевых структур, определяющих рассуждение в тексте, влияет на уверенность модели. Результаты показывают существенное падение уверенности модели при удалении ключевых структур / маркеров причинно-следственных связей и логических связей. Кроме того, при помощи метрики silhouette score демонстрируется повышение способности модели различать классы по мере продвижения к высшим слоям модели. В совокупности

данные наблюдения косвенно доказывают, что BERT формирует представления текста, опирающиеся на логическую структуру текста, а не только на важность отдельных токенов. В работе представлены косвенные, но последовательно согласованные между собой свидетельства того, что [CLS]-вектора, сформированные моделью BERT, определяют логическую структуру рассуждения в тексте, делая ее эффективным инструментом в частных, нетривиальных задачах анализа текста. Кроме того, результаты подчеркивают необходимость дальнейшего изучения русскоязычных языковых моделей.

Ключевые слова: машинное обучение, интерпретируемость трансформеров, BERT, классификация текстов, аргументация, скрытые представления, fine-tuning, NLP.

Введение

Бинарная классификация текстов на наличие рассуждений (логических связок, аргументации, причинно-следственных отношений) представляет практический интерес в ряде сценариев. Она позволяет фильтровать тексты, выделяя материалы с рассуждениями среди описательных или повествовательных, а также может использоваться для предварительной разметки данных при сборе специализированных датасетов, ориентированных на анализ логических связей, например, аналогичных SNLI на русском языке [3]. Наличие рассуждений помогает аннотаторам сосредоточиться на информативных примерах и ускоряет процесс разметки. Автоматизация выявления текстов с рассуждениями имеет потенциал для применения в системах обработки пользовательского контента, образовательных платформах и системах анализа юридических документов.

Существующие методы классификации, такие как TF-IDF с ансамблями моделей или LSTM с предобученными статическими векторами слов, обладают ограничениями: они используют статические признаки и плохо учитывают глобальный контекст текста. Экспериментальные результаты показывают, что подход на основе fine-tuning модели ruBERT демонстрирует более высокое качество бинарной классификации (см. табл. 2), что подтверждается и рядом исследований [10]. Особая сложность связана с русским языком, который характеризуется высокой морфологической разнообразностью, свободным порядком слов и богатством средств выражения логических и семантических связей. Большинство исследований по интерпретации BERT ориентированы на английский язык, что делает изучение русскоязычных моделей актуальным.

Таким образом, актуальность работы определяется двумя факторами:

- 1) практическая значимость задачи — фильтрация и разметка текстов с рассуждениями, применение в NLP-приложениях;
- 2) научная новизна — изучение внутренних представлений русскоязычной модели BERT и механизмов извлечения логических и семантических структур текста.

Цель данной работы — исследовать внутренние представления русскоязычной модели BERT (DeepPavlov / rubert-base-cased) для выявления признаков рассуждения в тексте и понять, какие лингвистические и логические структуры модель использует при классификации текстов на «содержащие рассуждение» и «не содержащие рассуждение».

В рамках достижения этой цели решались следующие задачи:

- 1) разработка бинарного классификатора на основе fine-tuning модели BERT;

- 2) экспериментальное сравнение полученного решения с другими подходами векторизации и классификации текстов;
- 3) анализ внутренних представлений модели BERT на специфичной задаче классификации текстов, содержащих рассуждение;
- 4) исследование способности предложенного подхода улавливать семантические и структурные признаки рассуждения в текстах с использованием методов интерпретации нейросетевых моделей.

Новизна исследования заключается в систематическом анализе внутренних представлений русскоязычной модели BERT с использованием совокупности методов: визуализации, probing-классификаторов и ablation study. Такой комбинированный подход позволяет объяснить высокое качество классификации и выявить, какие уровни и внутренние структуры трансформера отвечают за распознавание логически организованного текста.

1. Математическая формализация задачи

Пусть задано множество текстов:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, \quad \text{пары (текст, метка) рассматриваются без учета порядка,}$$

где x_i — i -й текст множества, а $y_i \in \{0, 1\}$ — бинарная метка, показывающая наличие рассуждения (1 — текст содержит рассуждение, 0 — нет).

Модель трансформера f_θ предсказывает вероятность наличия рассуждения $\hat{y}_i$ для каждого текста x_i :

$$\hat{y}_i = f_\theta(x_i), \quad \hat{y}_i \in (0, 1).$$

Обучение модели производится минимизацией бинарной кросс-энтропии:

$$L(\theta) = - \sum_{i=1}^N \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

Ниже кратко описаны математические формулы для использованных методов исследования интерпретируемости модели BERT.

1) Метод интерпретации Ablation Study [4]. Для каждого текста x_i можно рассматривать удаление подмножеств токенов $S \subset \text{tokens}(x_i)$. Пусть $x_i \setminus S$ — это текст x_i без токенов S . Под уверенностью модели понимается числовая оценка вероятности, присваиваемая моделью конкретному классу при предсказании.

Пусть модель BERT выдает логиты $z = [z_1, z_2, \dots, z_C]$ для C классов. Уверенность модели в принадлежности текста x к классу c вычисляется с помощью функции softmax:

$$p(c | x) = \frac{\exp(z_c)}{\sum_{i=1}^C \exp(z_i)},$$

где z_c — логит модели для класса c , а $p(c | x)$ отражает уверенность модели в том, что текст принадлежит классу c .

Изменение уверенности модели при исключении токенов S из текста можно формализовать следующим образом:

$$\Delta \hat{y}_i(S) = \hat{y}_i - f_{\theta}(x_i \setminus S),$$

где $\hat{y}_i = f_{\theta}(x_i)$. Если значение $\Delta \hat{y}_i(S)$ велико, это означает, что токены S оказывают сильное влияние на классификацию текста как рассуждения. Соответственно, можно определить функцию важности токенов:

$$Importance(S) = \Delta \hat{y}_i(S).$$

2) Интерпретация внутренних представлений. В модели BERT входной текст предварительно преобразуется в последовательность токенов — элементарных единиц представления текста, получаемых с помощью подсловной токенизации (WordPiece) [8]. Токены могут соответствовать целым словам или их фрагментам, а специальные служебные токены, такие как [CLS], [SEP] и [PAD], добавляются для обозначения начала, конца и выравнивания последовательности.

Каждый токен после токенизации преобразуется в числовой идентификатор и далее в вектор вложения фиксированной размерности. Эти векторы подаются в трансформер, и на каждом слое BERT формируются скрытые представления токенов, содержащие контекстную информацию о последовательности.

Одним из служебных токенов является токен [CLS] (буквальное обозначение токена в модели BERT), который добавляется в начало каждой входной последовательности. Вектор, соответствующий токenu [CLS] на выходе последнего слоя модели, называется [CLS]-вектором. Этот вектор используется как агрегированное представление всей последовательности и применяется для задач классификации. Размерность CLS-вектора в используемой модели составляет 768.

Формально, для входной последовательности токенов

$$X = ([CLS], x_1, x_2, \dots, x_n),$$

где x_i — i -й токен исходного текста, BERT формирует матрицу скрытых состояний на последнем слое:

$$H^{(L)} = [h_{\text{CLS}}^{(L)}, h_1^{(L)}, \dots, h_n^{(L)}],$$

где $h_i^{(L)} \in \mathbb{R}^d$ — вектор вложения i -го токена. CLS-вектор определяется как

$$\mathbf{c} = h_{\text{CLS}}^{(L)} \in \mathbb{R}^d.$$

Для задачи классификации используется линейный слой:

$$\hat{y} = \text{softmax}(W\mathbf{c} + b),$$

где $W \in \mathbb{R}^{k \times d}$ — матрица весов, $b \in \mathbb{R}^k$ — смещение, k — число классов.

Для оценки того, насколько информация о рассуждении закодирована на каждом уровне трансформера, вводятся probing-классификаторы g_l , обучаемые на CLS-векторах c_i^l каждого слоя l :

$$c_i^l := h_{\text{CLS},i}^{(l)} \in \mathbb{R}^d,$$

$$\hat{y}_i^l = g_l(c_i^l),$$

где $\hat{y}_i^l$ — предсказание probing-классификатора для текста x_i на слое l .

На основе этих предсказаний вычисляются метрики качества $F1_l$, accuracy и Silhouette Score для каждого слоя:

$$\text{Quality}(D) = F1(\{(d_i, s_i)\}_{i=1}^n),$$

что позволяет формально оценить, на каком слое трансформера закодирована информация о рассуждении.

2. Описание данных и процесса fine-tuning BERT модели для решения задачи

1) Датасет, размер, источники, разметка. Для обучения модели использовались 2 969 русскоязычных текстов, собранных вручную из различных источников (новостные сайты, Википедия, художественная литература). В каждом тексте была определена двоичная метка «есть рассуждение» (1) или «нет рассуждения» (0). Разметка выполнена вручную, по заранее определенным критериям: наличие причинно-следственных конструкций, выводных сегментов, логических переходов, объяснений или интерпретаций явлений. Все пограничные случаи, содержащие лишь частичные признаки рассуждения или неоднозначные логические связи, систематически относились к классу 0 («нет рассуждения»). Тексты разнообразны, варьируются как по стилю так и по длине.

Набор включает 1 499 примеров, которые отнесены к классу «есть рассуждение» и 1 470 примеров — «нет рассуждения».

2) Подготовка данных для BERT. Для подачи текстов в модель BERT используется подсловная токенизация WordPiece [8], при которой подразумевается разбиение текста на последовательность токенов, представляющих собой целые слова или их фрагменты. Такой подход позволяет модели эффективно обрабатывать морфологически сложные и редкие формы слов.

В начале каждой последовательности добавляется специальный токен '[CLS]', скрытое состояние которого используется как агрегированное семантическое представление всего текста. Для разделения последовательностей и обозначения конца любого текста применяется специальный токен '[SEP]'. При необходимости выравнивать длину последовательности используется специальный токен '[PAD]' (в случае, если текущая последовательность меньше максимальной длины).

После процесса токенизации каждый токен преобразуется в числовой идентификатор в соответствии со словарем модели, таким образом формируя входную последовательность, которая в дальнейшем используется для получения контекстных векторов вложения.

Например, фраза «Интерпретация текста» может быть токенизирована моделью RuBERT следующим образом:

'[CLS] Интер ##претация текста [SEP]'

Здесь 'Интер' и '##претация' — подсловные токены, показывающие, что слово «Интерпретация» было разбито на два фрагмента. После этого каждому токenu присваивается числовой идентификатор (*token ID*) из словаря модели, формируя последовательность чисел, которая подается на вход модели для получения контекстных векторов вложения.

3) Описание процесса дообучения модели RuBERT. Поверх базовой архитектуры BERT модель RuBERT (DeepPavlov/rubert-base-cased) была загружена через класс AutoModelForSequenceClassification, который автоматически добавляет выходной полносвязный классификационный слой с softmax-активацией. Выходной классификационный

слой — это дополнительный слой, добавляемый поверх базовой архитектуры BERT. Он принимает на вход агрегированное представление текста, обычно [CLS]-вектор, и преобразует его в вектор вероятностей для каждого класса. В данном случае это полносвязный слой с softmax-активацией, который выдает вероятность того, что текст относится к классу «есть рассуждение» или «нет рассуждения». Количество выходных нейронов было установлено равным числу классов (`num_labels=2`).

Чтобы избежать переобучения на небольшой обучающей выборке и сохранить устойчивыми полезные языковые представления, уже выученные моделью BERT на большом корпусе текстов, первые 8 из 12 слоев были заморожены во время дообучения. Под устойчивостью понимается сохранение этих весов без изменений, чтобы они продолжали предоставлять надежное общее представление текста, тогда как обучение проводилось только в верхних слоях (9–12), адаптирующих модель к конкретной задаче классификации рассуждений [11].

Дообучение выполнялось с использованием библиотеки HuggingFace Transformers и класса Trainer.

Таблица 1

Гиперпараметры модели при дообучении

Параметр	Значение	Описание
Epochs	8	количество эпох
Learning rate	2e-5	скорость обучения (типичная для BERT)
Weight decay	0.05	L2-регуляризация
Eval strategy	«epoch»	оценка после каждой эпохи
Early stopping patience	2	остановка при отсутствии улучшения
per_device_train_batch_size	32	размер батча при обучении
per_device_eval_batch_size	32	размер батча при валидации
save_strategy	«epoch»	сохранение лучшей модели
load_best_model_at_end	True	загрузка лучшей по метрике

Резюмируя параметры в таблице 1: модель RuBERT дообучалась в течение 8 эпох с небольшой скоростью обучения (2e-5) и коэффициентом регуляризации (weight decay), равным 0,05. Для стабильности использовался разогрев (warmup), равный 0,1, а оценка модели производилась после каждой эпохи с возможностью ранней остановки через 2 эпохи при отсутствии улучшений. Размер батча составлял 32 для обучения и валидации. Лучшие веса автоматически сохранялись и загружались в конце обучения. Лучшие веса модели в данном случае – это сохраненные параметры, которые показали наивысшее качество на валидационной выборке во время обучения.

Для оценки эффективности предложенного в данной статье подхода, предварительно были проведены эксперименты с использованием метода стекинга моделей и рекуррентной нейронной сетью.

В методе ансамблевого стекинга (stacking [20]) несколько сильных базовых классификаторов (XGBoost, CatBoost и Random Forest) обучаются на одних и тех же признаках, представленных через TF-IDF-векторизацию текстов. На выходах этих моделей строится новый вектор признаков, который подается на мета-модель, выполняющую окончательную классификацию; в качестве мета-модели использовалась логистическая регрессия. Такой подход позволяет объединить преимущества разных алгоритмов, улуч-

шая общую точность и устойчивость классификации. На контрольной выборке данный метод показал $F1 = 0,905$ и $ROC\ AUC = 0,887$ [1]. Несмотря на высокий Recall (0,930), ограничение данного метода в том, что используются статические признаки (TF-IDF) и не учитывается глобальный контекст, что снижает точность при распознавании сложных конструкций, содержащих рассуждение.

Метод RusVectores + LSTM использует предобученные статические векторы вложения слов из модели RusVectores [6]. Тексты предварительно токенизируются и преобразуются в последовательности индексов по словарю векторов вложения, а слова вне словаря заменяются на индекс по умолчанию. Эти последовательности подаются на слой Embedding TensorFlow с фиксированными весами ($trainable = False$), после чего проходят обработку двунаправленной LSTM с регуляризацией и Dropout. Финальный Dense слой с softmax выполняет классификацию. Такой подход учитывает порядок слов и локальный контекст, но ограничен статическими представлениями слов. На контрольной выборке метод показал $F1 = 0,86$ и $ROC\ AUC = 0,84$, где $F1$ — гармоническое среднее точности (precision) и полноты (recall) модели [19]:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

a ROC AUC (area under the receiver operating characteristic curve) характеризует способность модели различать положительные и отрицательные классы [9].

Таблица 2

Сравнение производительности различных моделей на задаче классификации текстов, с наличием рассуждения

Модель	Accuracy	Precision	Recall	F1-score	ROC AUC	PR AUC
Fine-tuned ruBert	0,943	0,952	0,933	0,943	0,988	0,989
Stacking моделей	0,855	0,881	0,930	0,905	0,887	0,952
Rusvectors LSTM	0,890	0,880	0,840	0,860	0,840	0,910

В таблице 2 приведены результаты трех подходов к классификации текстов на рассуждения и не рассуждения.

BERT (RuBERT) с дообучением (fine-tuning) использует предобученную трансформерную архитектуру с двунаправленным вниманием, где специальный токен [CLS] агрегирует информацию о всей последовательности и подается на выходной классификационный слой для предсказания класса. Fine-tuning означает дообучение предобученной модели на конкретной задаче с адаптацией весов к целевой выборке [5], что позволяет эффективно использовать контекстуальные представления для выявления сложных конструкций, содержащих рассуждение. Данный метод показал наивысшие результаты по всем метрикам: precision достигла 0,952, F1-score 0,943, ROC AUC 0,988, PR AUC 0,989. Модель эффективно различает тексты с рассуждениями и без, используя контекстуальные векторные представления токена и улавливая синтаксические и семантические связи, что позволяет корректно обрабатывать длинные и разнообразные тексты.

В ряде исследований (Joshi et al., 2019 [17]; Ethayarajh, 2019 [7]; Wang et al., 2020 [18]; Zhou & Bloem, 2021 [21]) показано, что контекстные векторы вложения (ELMo, BERT) стабильно превосходят статические (Word2Vec, GloVe) на задачах, требующих глубокого анализа смысла и структурных отношений в тексте. На их основе можно полагать, что для задачи классификации рассуждений, в которой важно выявить логические

и причинно-следственные конструкции, контекстные векторы вложения могут дать преимущество перед статичными. Это согласуется с результатами текущей работы: методы, использующие статические представления (TF-IDF или RusVectores), демонстрируют более низкие показатели качества, тогда как fine-tuning трансформерной модели BERT приводит к значительному росту всех метрик.

Таким образом, превосходство BERT над другими подходами объясняется способностью предобученной трансформерной архитектуры в сочетании с fine-tuning формировать контекстно-зависимые представления, отражающие не только локальные связи между словами, но и более сложные структуры, содержащие рассуждение, такие как логические переходы, причинно-следственные отношения и глобальную организацию текста. Именно эта особенность позволяет модели демонстрировать наилучшие показатели качества среди всех рассмотренных методов.

Высокая эффективность fine-tuned BERT поднимает вопрос: какие элементы текста и какие скрытые закономерности модель использует для столь точного распознавания рассуждений? В последующих разделах рассматриваются методы интерпретации и анализа, позволяющие выяснить, какие фрагменты текста оказывают наибольшее влияние на уверенность модели при классификации текстов, содержащих рассуждение.

3. Визуализация [CLS]-векторов

Для анализа внутренних представлений модели были извлечены CLS-векторы всех слоев BERT-модели. Для их визуализации применялся UMAP (Uniform Manifold Approximation and Projection) — алгоритм снижения размерности, который проецирует высокоразмерные данные в пространство низкой размерности (например, 2D) так, чтобы сохранить локальную и глобальную структуру данных [12]. На визуализациях видно, как представления текстов постепенно разделяются по классам: на ранних слоях классы почти не различимы, тогда как на более высоких слоях проявляется явная кластеризация, что косвенно подтверждает, что модель учится выделять признаки рассуждений.

Резкий рост различимости кластеров на последнем слое объясним тем, что CLS-вектор на этом слое собирает в себя наиболее информативные признаки, кроме того, данный эффект усиливается fine-tuning.

4. Исследование характеристики Silhouette score

Для количественной оценки разделимости классов в пространстве CLS-векторов была вычислена Silhouette score для каждого слоя модели BERT.

Silhouette score является метрикой, позволяющей оценить качество кластеризации объектов на основе того, насколько схожи объекты в текущем кластере по сравнению с объектами в других кластерах [16].

Пусть есть набор объектов $\mathcal{X} = (x_1, \dots, x_N)$, разделенных на кластеры $G_1, \dots, G_K$, и выбрана метрика расстояния $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$.

- 1) Для каждого объекта x_i определим среднее расстояние до объектов другого кластера $|G \neq G_{i_1}|$:

$$d(x_i, G) = \frac{1}{|G|} \sum_{x_j \in G} d(x_i, x_j).$$

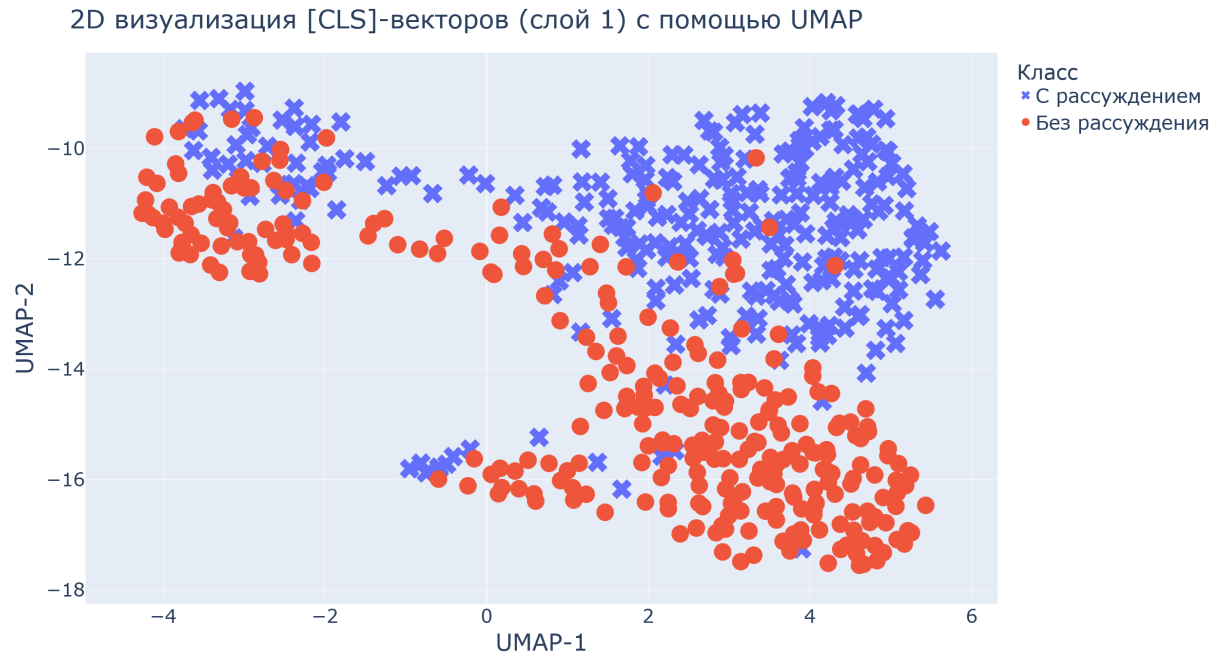


Рис. 1. UMAP-визуализация CLS-векторов на 1-м внутреннем слое модели

2D визуализация [CLS]-векторов (слой 7) с помощью UMAP

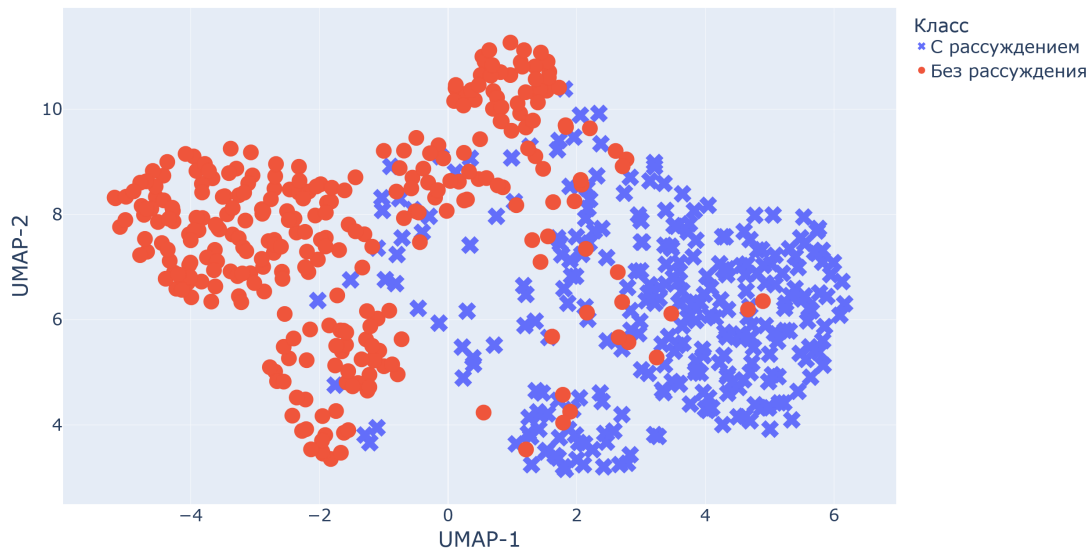


Рис. 2. UMAP-визуализация CLS-векторов на 7-м внутреннем слое модели

2) Пусть $b(x_i)$ — это минимальное среднее расстояние до других кластеров:

$$b(x_i) = \min_{G \neq G_{i_1}} d(x_i, G).$$

3) Silhouette score для объекта x_i определяется как:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \in [-1, 1].$$

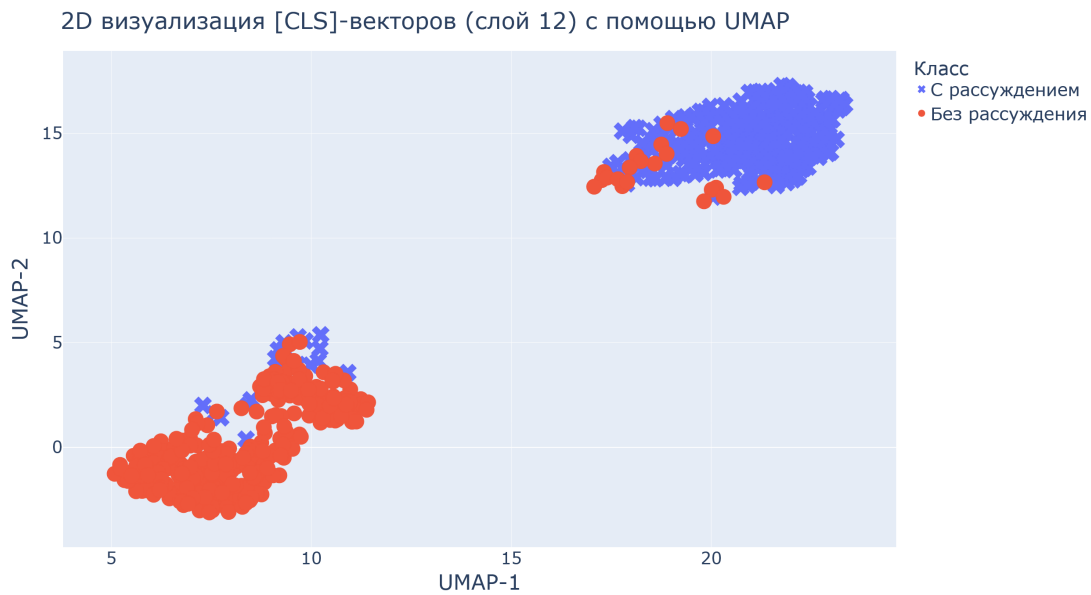


Рис. 3. UMAP-визуализация CLS-векторов на 12-м внутреннем слое модели

4) Средний Silhouette score для всего набора:

$$S = \frac{1}{N} \sum_{i=1}^N s(x_i).$$

На рисунке 4 видно, что по мере продвижения к более глубоким слоям значение Silhouette score растет, что подтверждает визуальные наблюдения из UMAP: модель постепенно формирует более чистые и разделенные представления для текстов с рассуждениями и без них.

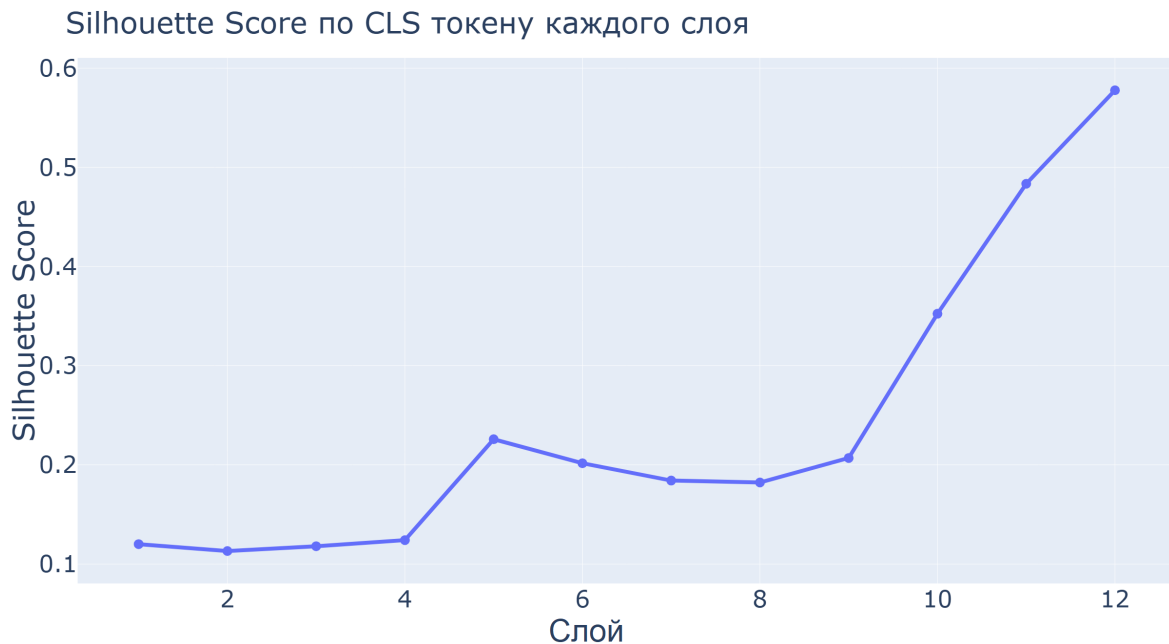


Рис. 4. Silhouette score для CLS-векторов на каждом слое модели

На графике Silhouette score по слоям модели виден резкий рост начиная с девятого слоя, что соответствует началу размороженных слоев для fine-tuning. Первые восемь слоев были заморожены, что объясняет относительно низкие значения метрики на этих слоях, тогда как обучаемые слои (9–12) начинают лучше различать представления текстов с наличием рассуждения и без. На рисунке 5 обучаемые слои обозначены синим цветом.

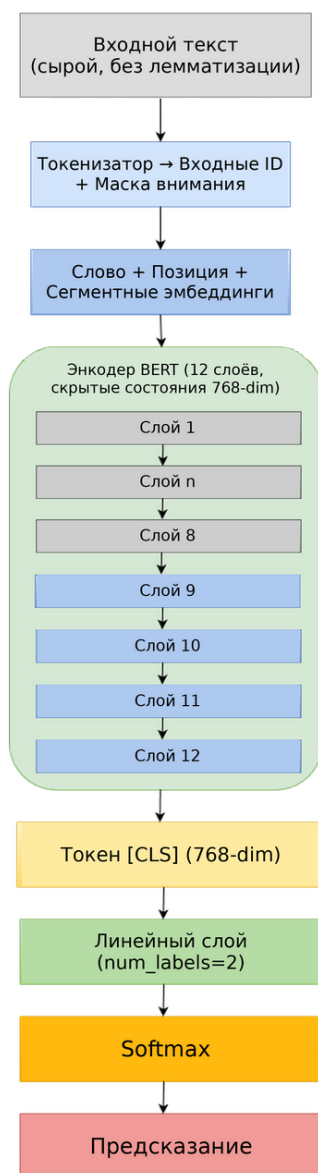


Рис. 5. Архитектура модели RuBERT для задачи классификации: входной текст токенизируется и преобразуется в эмбединги, далее проходит 12 слоев BERT, после чего вектор скрытого состояния токена [CLS] подается на линейный слой с Softmax для предсказания класса

5. Анализ скрытых представлений по слоям

Для оценки того, на каком слое модели BERT появляется информация, полезная для классификации рассуждений, был проведен анализ скрытых представлений по слоям, который является методом интерпретации и анализа внутренних представлений трансформерных моделей, таких как BERT, с целью выявить, на каком уровне архитектуры модель формирует признаки, полезные для конкретной задачи. Основная идея заключается в том, что каждый слой модели может кодировать разную информацию: нижние слои чаще содержат синтаксическую информацию (например, части речи, зависимости), а верхние слои — семантическую и контекстуально зависимую информацию (например, смысловые связи, аргументацию, логические конструкции) [15].

Для этого из каждого слоя извлекались [CLS]-векторы, и на их основе обучался простой линейный классификатор (логистическая регрессия) для предсказания наличия рассуждения [13].

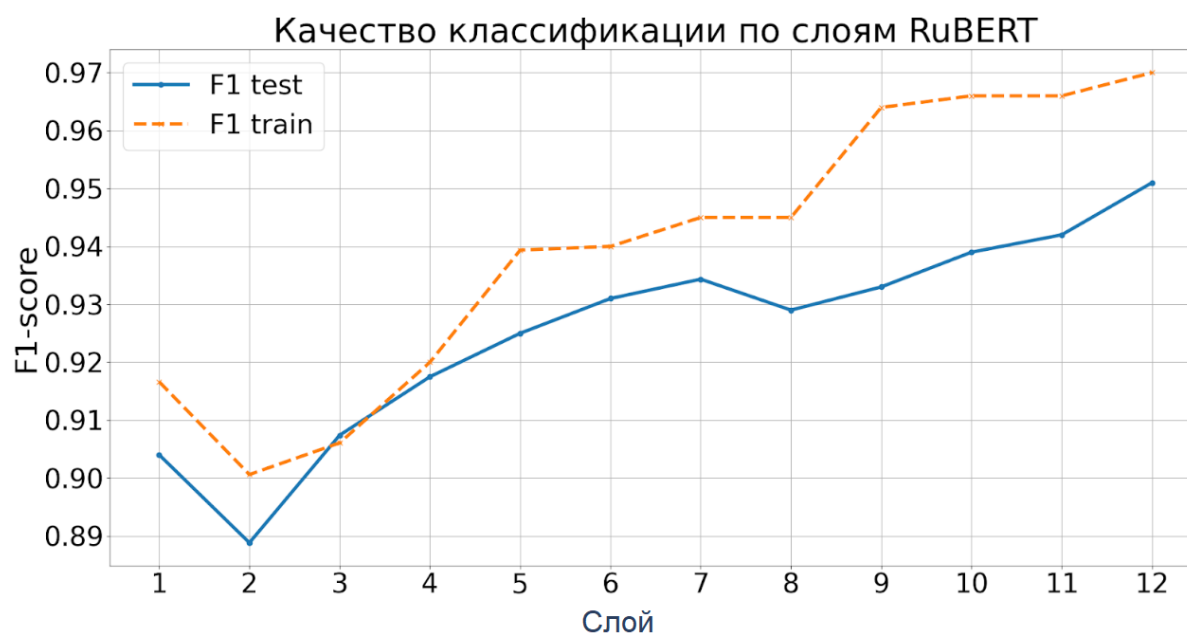


Рис. 6. F1 score для LR на [CLS]-токенах каждого слоя

Результаты показали, что даже на первых слоях модель уже демонстрирует неплохую точность: F1-score на тестовой выборке начинается с 0,90. С ростом номера слоя точность постепенно увеличивается, достигая максимума на последних слоях (F1-score на тестовой выборке 0,95 на 12-м слое), что совпадает с улучшением метрик Precision, Recall и F1. Это подтверждает, что BERT уже на ранних слоях выделяет базовую информацию о тексте, а последние слои концентрируют признаки, важные именно для задачи классификации рассуждений.

Таким образом, даже без дообучения (fine-tuning) BERT способен различать тексты с рассуждениями и без них на основе предобученных языковых представлений. Fine-tuning дополнительно усиливает эти признаки, особенно на верхних слоях, что объясняет высокие показатели итоговой модели.

6. Применение метода Ablation Study для анализа предсказаний модели

Для более глубокого понимания того, какие элементы текста наиболее значимы для предсказаний модели, был использован метод Ablation Study. Суть метода заключается в поэтапном удалении отдельных слов и конструкций из текста и оценке влияния этих изменений на уверенность модели. Для каждого текста генерировались комбинации токенов, которые затем исключались, и вычислялась вероятность класса «есть рассуждение» после их удаления.

Пусть текст представлен как последовательность токенов:

$$x = [t_1 \ t_2 \ \dots \ t_n]$$

Модель классификации f_{θ_1} возвращает вероятность класса «есть рассуждение»:

$$p_{\theta}(x) = f_{\theta}(x) \in [0, 1]$$

Для Ablation Study выбирается множество комбинаций токенов $C = \{c \mid c \subseteq x\}$, где каждая комбинация $c = \{t_{i_1}, \dots, t_{i_k}\}$. Операция удаления токенов из текста может быть определена так:

$$x \setminus c = x \text{ без токенов из } c$$

Тогда вероятность класса после удаления комбинации c вычисляется как:

$$p_{\theta}(x \mid c) = f_{\theta}(x \mid c)$$

Задача Ablation Study заключается в поиске комбинаций, которые минимизируют уверенность модели:

$$c^* = \arg \min_{c \in C} p_{\theta}(x \setminus c)$$

То есть c^* это те токены, удаление которых снижает уверенность модели максимальным образом, что формально отражает вклад этих токенов в распознавание рассуждений.

В таблице 3 и 4 первая колонка содержит отдельные слова или комбинации слов, которые модель BERT использует как маркеры рассуждений. Вторая колонка («Разность уверенности модели») показывает, насколько наличие данного маркера увеличивает уверенность модели в предсказании класса «Есть рассуждение». Для отдельных слов значения обычно меньше (около 0,1–0,2), для комбинаций слов выше (0,65–0,75), что отражает их более сильное влияние на итоговую уверенность модели.

Результаты показали, что наибольшее снижение уверенности модели происходит при удалении слов и конструкций, которые формируют структуру рассуждения, например, причинно-следственные связки («значит», «поэтому», «следовательно») и ключевые слова, несущие смысловую нагрузку. В частности, для текста «На улице лужи, хотя дождь прекратился несколько часов назад. Значит, дождь был очень сильным, так как земля не успела просохнуть. Это объяснение кажется мне логичным.» удаление слов «значит», «это», «логичным» снижало уверенность модели до 0,03, что свидетельствует о критической роли этих маркеров в идентификации рассуждений.

Для других текстов аналогичная закономерность сохраняется: комбинации слов, определяющих причинно-следственные отношения или выводы, оказывают максимальное влияние на предсказание модели. Например, для текста о научно-техническом прогрессе удаление «неразрывно», «связан» и «хотя» снижало вероятность до 0,80. Это

Таблица 3

Отдельные слова, при удалении которых из текста максимально снижается уверенность модели

Слово	Разность уверенности модели
думается	0,18
понимаю	0,17
следует	0,19
почему	0,16
следовательно	0,19
возможно	0,14
должно	0,18
видимо	0,15
несомненно	0,19
поскольку	0,18
рассуждение	0,20
понятие	0,15
исследуют	0,14
знание	0,17
философия	0,16
разумность	0,18
осознанность	0,17
мысль	0,14
идея	0,14
значит	0,18
поэтому	0,20
однако	0,18
таким образом	0,20
в итоге	0,17
вызывает	0,13
приводит	0,18
возникает	0,15
связано	0,12
согласно	0,14
доказательство	0,19
анализ	0,16

подтверждает, что модель не просто учитывает отдельные ключевые слова, но действительно распознает структурные элементы рассуждений.

Таким образом, Ablation Study позволяет формально выявить, какие токены и конструкции определяют уверенность модели, и показывает, что BERT способен выделять и использовать важные логические и причинно-следственные элементы текста для классификации рассуждений.

7. Результаты и обсуждение

Модель BERT, дообученная на корпусе русскоязычных текстов с разметкой наличия рассуждений, показала наилучшие результаты среди протестированных подходов. По метрикам Accuracy, F1-score, ROC AUC и PR AUC она превосходит как стекинг

Таблица 4

**Комбинации слов, при удалении которых из текста
максимально снижается уверенность модели**

Комбинация	Разность уверенности модели
поскольку, видимо, существует	0,74
поскольку, пристойность, важно	0,74
следует, принято, хвалят	0,73
причине, думается, трудно	0,71
согласно, возвращайся, философии	0,70
поэтому, выбирают, дают	0,75
однако, часто, большего	0,73
однако, итоге, достигают	0,74
таким, превращают, неизвестным	0,75
образом, превращают, страх	0,72
идешь, поэтому, значит	0,75
развивается, поэтому, значит	0,75
эпистемон, рассуждении, несомненно	0,75
эпистемон, можешь, несомненно	0,74
добродетель, ум, понятие	0,73
добродетель, ум, считается	0,70
представлениями, пропитывай, вещь	0,67
разумности, понять, знание	0,75
разумности, понять, исследуют	0,71
например, формируются, осознанность	0,72
учатся, формируются, осознанность	0,71
человек, опыт, помогает	0,67
это, помогает, развивать	0,66
переосмыслить, результате, будущем	0,73
это, поэтому, важно	0,75
человек, это, поэтому	0,74
человек, это, легче	0,63
это, получается, удивительно	0,70
получается, это, удивительно	0,69

древовидных моделей (CatBoost, XGBoost, Random Forest с логистической перцепцией в качестве мета-модели), так и комбинацию векторов вложения RusVecores с LSTM. Данные показатели демонстрируют, что трансформер эффективно различает тексты с рассуждениями и без них, даже без дополнительной предобработки или лемматизации текста.

Для анализа того, каким образом BERT различает структуры рассуждений, были проведены анализ скрытых представлений по слоям и визуализация [CLS]-векторов. С помощью линейного классификатора на CLS-векторах, извлеченных из каждого слоя, удалось оценить информативность представлений. Результаты показали, что уже на средних слоях модель улавливает полезные признаки, однако резкий рост точности наблюдается с девятого слоя, что совпадает с размороженными слоями при fine-tuning. Этот эффект подтверждается и Silhouette score: разделимость классов на CLS-векторах значительно улучшается на последних слоях, что отражает способность модели выделять структурные особенности текста.

Дополнительно была проведена Ablation Study: из текстов поочередно удалялись ключевые слова и конструкции, ответственные за причинно-следственные связи, после чего оценивалась уверенность модели в предсказании. Результаты показали, что удаление слов вроде «значит», «поэтому», «следовательно» и «это» приводит к заметному снижению вероятности корректного распознавания рассуждения. Это свидетельствует о том, что BERT учитывает не только отдельные слова, но и структурные элементы текста при классификации.

В совокупности полученные результаты подтверждают, что BERT способен эффективно различать тексты с рассуждениями и без них благодаря глубокому языковому представлению и тонкой настройке на задачу. Анализ скрытых представлений по слоям и ablation study показывают, что модель учится выделять ключевые структурные элементы текста, а не только отдельные токены, что делает ее подход особенно полезным для анализа структуры рассуждений в русскоязычных текстах. Особый интерес представляют возможности BERT и моделей трансформеров в области задач естественного языкового вывода (NLI), в том числе поиска логических связей и противоречий в тексте. В других подходах используются графовые методы для извлечения признаков на основе структуры текста, например, применение индекса Винера для вычисления текстовых характеристик и построения признаков для классификации текстов на естественном языке [2]. Использование трансформеров в данной области также может быть эффективным, поскольку их контекстуальные представления способны учитывать семантические и структурные связи между элементами текста, что позволяет интегрировать информацию о логике и связности текста в задачи NLI.

Заключение

В работе показано, что трансформер BERT, дообученный на корпусе русскоязычных текстов с метками наличия рассуждений, превосходит альтернативные методы, включая стекинг деревоподобных моделей и LSTM на векторах вложения. Модель демонстрирует высокую точность и уверенность в классификации текстов, эффективно улавливая логические и причинно-следственные структуры.

Анализ [CLS]-векторов по слоям и Ablation Study подтвердил, что последние слои BERT наиболее информативны для выделения признаков рассуждения, а удаление ключевых токенов существенно снижает уверенность предсказаний. Это говорит о том, что модель учитывает как отдельные слова, так и их структурные комбинации.

Полученные результаты подтверждают потенциал использования BERT для задач анализа рассуждений и открывают возможность дальнейшего изучения внутренних представлений модели, что может быть полезно для разработки инструментов анализа текста и автоматической разметки логических конструкций в документах на русском языке.

СПИСОК ЛИТЕРАТУРЫ

1. Каныгин, А. В. Построение модели для решения задачи классификации рассудительного текста / А. В. Каныгин // Математическая физика и компьютерное моделирование. — 2025. — № 1. — С. 1–12. — DOI: 10.15688/mpcm.jvolsu.2025.1.3
2. Клячин, В. А. О возможности использования индекса Винера для вычисления признаков текстов на естественном языке / В. А. Клячин, Е. В. Хижнякова // Математическая физика и компьютерное моделирование. — 2025. — № 3. —

C. 24–36. — DOI: 10.15688/mpcm.jvolsu.2025.3.3

3. A Large Annotated Corpus for Learning Natural Language Inference / S. R. Bowman, G. Angeli, C. Potts, C. D. Manning // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 632–642. — DOI: 10.18653/v1/D15-1075

4. Ablation Studies to Uncover Structure of Learned Representations in Artificial Neural Networks / R. Meyes, M. Lu, C. Waubert de Puiseau, T. Meisen // Proceedings of the 2019 International Conference on Artificial Intelligence (ICAI). — 2019.

5. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). — 2019. — P. 4171–4186. — DOI: 10.18653/v1/N19-1423

6. Bolshina, A. All-Words Word Sense Disambiguation for Russian Using Automatically Generated Text Collection / A. Bolshina, N. Loukachevitch // Cybernetics and Information Technologies. — 2020. — P. 90–107. — DOI: 10.2478/cait-2020-0049

7. Ethayarajh, K. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings / K. Ethayarajh // 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. — P. 55–65. — DOI: 10.18653/v1/D19-1006

8. Fast WordPiece Tokenization / X. Song, A. Salcianu, Y. Song, D. Dopson, D. Zhou // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2021.

9. Fawcett, T. An Introduction to ROC Analysis / T. Fawcett // Pattern Recognition Letters. — 2006. — Vol. 27, № 8 — DOI: 10.1016/j.patrec.2005.10.010

10. Gonzaez-Carvajal, S. Comparing BERT Against Traditional Machine Learning Models in Text Classification / S. Gonzaez-Carvajal, E. C. Garrido-Merchan, R. Gozalo-Brizuela // Journal of Computational and Cognitive Engineering. — 2023.

11. Griesshaber, D. Fine-Tuning BERT for Low-Resource Natural Language Understanding via Active Learning / D. Griesshaber, J. Maucher, N. T. Vu // Proceedings of the 28th International Conference on Computational Linguistics (COLING). — 2020. — P. 1158–1171. — DOI: 10.18653/v1/2020.coling-main.100

12. Healy, J. Uniform Manifold Approximation and Projection / J. Healy, L. McInnes // Nature Reviews Methods Primers. — 2024. — P. 82. — DOI: 10.1038/s43586-024-00363-x

13. Jawahar, G. What Does BERT Learn About the Structure of Language? / G. Jawahar, B. Sagot, D. Seddah // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). — 2019. — P. 3651–3657. — DOI: 10.18653/v1/P19-1356

14. Pre-Trained Models for Natural Language Processing: A Survey / X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang // Science China Technological Sciences. — 2020. — DOI: 10.1007/s11431-020-1647-3

15. Rogers, A. A Primer in BERTology: What We Know About How BERT Works / A. Rogers, O. Kovaleva, A. Rumshisky // Transactions of the Association for Computational Linguistics. — 2020. — P. 417–444. — DOI: 10.1162/tacl_a_00349

16. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis / P. J. Rousseeuw // Journal of Computational and Applied Mathematics. — 1987. — P. 53–65. — DOI: 10.1016/0377-0427(87)90125-7

17. SpanBERT: Improving Pre-Training by Representing and Predicting Spans / M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy // Transactions of the Association for Computational Linguistics. — 2020. — P. 64–77. — DOI: 10.1162/tacl_a_00300

18. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems / A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman // Proceedings of the Thirty-Third Conference on Neural Information Processing Systems (NeurIPS). — 2019.

19. van Rijsbergen, C. J. Information Retrieval / C. J. van Rijsbergen. — London : Butterworths, 1979. — 208 p.
20. Wolpert, D. H. Stacked Generalization / D. H. Wolpert // Neural Networks. — 1992. — P. 241–259. — DOI: 10.1016/S0893-6080(05)80023-1
21. Zhou, W. Comparing Contextual and Static Word Embeddings with Small Philosophical Data / W. Zhou, J. Bloem // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). — 2021. — P. 26–29. — DOI: 10.18653/v1/2021.acl-srw.26

REFERENCES

1. Kanygin A.V. Building a Model for Argumentative Text Classification. *Matematicheskaya fizika i kompyuternoe modelirovanie* [Mathematical Physics and Computer Simulation], 2025, no. 1, pp. 1-12. DOI: 10.15688/mpcm.jvolsu.2025.1.3
2. Klyachin V.A., Khizhnyakova E.V. On the Applicability of the Wiener Index for Computing Features of Natural Language Texts. *Matematicheskaya fizika i kompyuternoe modelirovanie* [Mathematical Physics and Computer Simulation], 2025, no. 3, pp. 24-36. DOI: 10.15688/mpcm.jvolsu.2025.3.3
3. Bowman S.R., Angeli G., Potts C., Manning C.D. A Large Annotated Corpus for Learning Natural Language Inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 632-642. DOI: 10.18653/v1/D15-1075
4. Meyes R., Lu M., Waubert de Puiseau C., Meisen T. Ablation Studies to Uncover Structure of Learned Representations in Artificial Neural Networks. *Proceedings of the 2019 International Conference on Artificial Intelligence (ICAI)*, 2019.
5. Devlin J., Chang M.-W., Lee K., Toutanova K. 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 (NAACL)*, 2019, pp. 4171-4186. DOI: 10.18653/v1/N19-1423
6. Bolshina A., Loukachevitch N. All-Words Word Sense Disambiguation for Russian Using Automatically Generated Text Collection. *Cybernetics and Information Technologies*, 2020, pp. 90-107. DOI: 10.2478/cait-2020-0049
7. Ethayarajh K. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *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, pp. 55-65. DOI: 10.18653/v1/D19-1006
8. Song X., Salcianu A., Song Y., Dopson D., Zhou D. Fast WordPiece Tokenization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
9. Fawcett T. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 2006, vol. 27, no. 8 DOI: 10.1016/j.patrec.2005.10.010
10. Gonzales-Carvajal S., Garrido-Merchan E.C., Gozalo-Brizuela R. Comparing BERT Against Traditional Machine Learning Models in Text Classification. *Journal of Computational and Cognitive Engineering*, 2023.
11. Griesshaber D., Maucher J., Vu N.T. Fine-Tuning BERT for Low-Resource Natural Language Understanding Via Active Learning. *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020, pp. 1158-1171. DOI: 10.18653/v1/2020.coling-main.100
12. Healy J., McInnes L. Uniform Manifold Approximation and Projection. *Nature Reviews Methods Primers*, 2024, pp. 82. DOI: 10.1038/s43586-024-00363-x
13. Jawahar G., Sagot B., Seddah D. What Does BERT Learn About the Structure of Language?. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 3651-3657. DOI: 10.18653/v1/P19-1356

14. Qiu X., Sun T., Xu Y., Shao Y., Dai N., Huang X. Pre-Trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, 2020 DOI: 10.1007/s11431-020-1647-3
15. Rogers A., Kovaleva O., Rumshisky A. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 2020, pp. 417-444. DOI: 10.1162/tacl_a_00349
16. Rousseeuw P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 1987, pp. 53-65. DOI: 10.1016/0377-0427(87)90125-7
17. Joshi M., Chen D., Liu Y., Weld D.S., Zettlemoyer L., Levy O. SpanBERT: Improving Pre-Training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 2020, pp. 64-77. DOI: 10.1162/tacl_a_00300
18. Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S.R. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Proceedings of the Thirty-Third Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
19. van Rijsbergen C.J. *Information Retrieval*. London, Butterworths, 1979. 208 p.
20. Wolpert D.H. Stacked Generalization. *Neural Networks*, 1992, pp. 241-259. DOI: 10.1016/S0893-6080(05)80023-1
21. Zhou W., Bloem J. Comparing Contextual and Static Word Embeddings with Small Philosophical Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 26-29. DOI: 10.18653/v1/2021.acl-srw.26

INTERPRETATION OF BERT INTERNAL REPRESENTATIONS FOR THE CLASSIFICATION OF TEXTS CONTAINING REASONING

Alexander V. Kanygin

Postgraduate Student, Department of Computer Sciences
and Experimental Mathematics,
Volgograd State University
a.kanygin99@gmail.com
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. This article examines the problem of binary text classification for the presence of reasoning (logical connections, argumentation, and cause-and-effect relationships). This problem was solved using a fine-tuned ruBert model, achieving significantly higher performance than classical ML approaches (TF-IDF, model stacking, static embedding vectors, and an LSTM classifier). The goal of the article is not so much to demonstrate the quality of the solution using vectorization via the BERT model, but rather to interpret BERT's internal representations, allowing to understand which features and/or structural elements of the text determine the model's high performance. The article utilizes a combination of methods for analysis: visualization of [CLS] vectors using 2D dimensionality reduction via UMAP, silhouette score calculation for each of the BERT model's 12 layers, separate classifiers for all hidden layers of the model, and an ablation study, which evaluates how removing key structures that define reasoning in the text affects the model's confidence. The results show a significant drop in model confidence when removing key structures/markers of causal relationships and logical connectives. Furthermore, using the silhouette score metric, the model's ability to distinguish between classes improves as it moves toward higher layers of the model. Taken together, these observations indirectly demonstrate that BERT generates text representations based on the logical structure of the text, not just the importance of individual tokens. This paper presents indirect but consistent evidence that the [CLS] vectors generated by the BERT model capture the logical structure of reasoning in the text, making it an effective tool for specific, non-trivial text analysis tasks. Furthermore, the results highlight the need for further study of Russian-language language models.

Key words: machine learning, interpretability of transformers, BERT, text classification, argumentation, hidden representations, fine-tuning, natural language processing.