УДК 004.8+004.912 **DOI** 10.24412/2413-7383-2025-2-37-79-90

К. А. Никитенко, А. В. Звягинцева
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Донецкий государственный университет»
283001, г. Донецк, ул. Университетская, 24

ИНТЕРПРЕТИРУЕМОСТЬ НЕЙРОСЕМАНТИЧЕСКИХ МОДЕЛЕЙ ПРИ ИХ ПРИМЕНЕНИИ В ПРИКЛАДНЫХ ОБЛАСТЯХ

K. A. Nikitenko, A. V. Zviagintseva
Federal State Educational Institution of Higher Education «Donetsk State University»
283001, Donetsk, University str, 24

INTERPRETABILITY OF NEUROSEMANTIC MODELS IN THEIR APPLICATION IN APPLIED FIELDS

К. А. Нікітенко, Г. В. Звягінцева
Федеральна державна бюджетна освітня установа
вищої освіти «Донецький державний університет»
283001, м. Донецьк, вул. Університетська, 24

ІНТЕРПРЕТОВАНІСТЬ НЕЙРОСЕМАНТИЧНИХ МОДЕЛЕЙ ПРИ ЇХНЬОМУ ЗАСТОСУВАННІ У ПРИКЛАДНИХ ОБЛАСТЯХ

В статье рассматриваются вопросы интерпретируемости нейросемантических моделей при их применении в прикладных областях, таких как медицина, право, финансы, образование и промышленная автоматизация. Обсуждаются ключевые сложности интерпретации высокоразмерных векторных представлений, контекстно-зависимых признаков и скрытых слоев трансформерных архитектур в реальных сценариях. Представлены современные встроенные и пост-хок методы объяснения, адаптированные к требованиям практических задач, а также анализ баланса между точностью и прозрачностью моделей. Статья завершается представлением перспективных направлений исследования, направленных на повышение доверия и безопасности ИИ-систем в прикладных приложениях.
Ключевые слова: интерпретируемость, нейросемантические модели, прикладные области, трансформер, объяснимость, практическое применение.

The article addresses the problem of interpretability in neurosemantic models used for natural language processing tasks. It discusses the main challenges related to the interpretation of high-dimensional vector representations, context-dependent features, and deep transformer architectures. The paper outlines current approaches to interpretability, including both intrinsic and post-hoc methods. Special attention is given to the trade-off between accuracy and interpretability, as well as to the prospects for further development in this area.
Keywords: interpretability, neurosemantic models, applied areas, transformer, explainability, practical application.

У статті розглядаються питання інтерпретованості нейросемантичних моделей при їхньому застосуванні в прикладних областях, таких як медицина, право, фінанси, освіта та промислова автоматизація. Обговорюються ключові складності інтерпретації високорозмірних векторних уявлень, контекстно-залежних ознак і прихованих шарів трансформерних архітектур в реальних сценаріях. Представлено сучасні вбудовані та пост-хок методи пояснення, які адаптовано до вимог практичних завдань, а також аналіз балансу між точністю та прозорістю моделей. Стаття завершується поданням перспективних напрямків дослідження, які спрямовано на підвищення довіри та безпеки ШІ-систем у прикладних додатках.
Ключові слова: інтерпретованість, нейросемантичні моделі, прикладні області, трансформер, пояснюваність, практичне застосування.

Введение

Современные нейросемантические модели обработки естественного языка (NLP), включая архитектуры Transformer (BERT, GPT и их производные), находят активное применение не только в академических исследованиях, но и в широком спектре прикладных областей.

Нейросемантические модели – это подкласс нейросетевых архитектур, ориентированных на представление, обработку и интерпретацию семантической информации в языковых данных [1]. Сегодня применение нейросемантических моделей охватывает широкий спектр задач и отраслей: автоматическая обработка и анализ текстов, машинный перевод, извлечение знаний и построение онтологий, интеллектуальные диалоговые системы, автоматизация документооборота, цифровая гуманитаристика и анализ общественного мнения, биомедицинские исследования, образовательные технологии, правоприменение и обеспечение нормативного соответствия [2].

С учётом широкого спектра задач, в которых применяются нейросемантические модели, представляется целесообразным проанализировать их характерные особенности с точки зрения архитектуры, объёмов обучающих данных и прикладной направленности. Основные модели, используемые в различных отраслях, приведены в табл. 1.

Таблица 1 – Нейросемантические модели, область их применения и характеристики

Модель	Область применения	Характеристики и параметры	Обучающие данные
1	2	3	4
BERT [3]	Обработка естественного языка, лингвистика	12-24 слоёв, 110 млн параметров; двухнаправленные трансформеры; высокая интерпретируемость через attention-механизмы	Книги, википедия (англоязычные тексты) (~16 гб)
GPT-2/3/4 [4]	Генерация текста, диалоговые системы	96 слоёв, 175 млрд параметров; автогенерация текста; ограниченная интерпретируемость	Интернет-корпусы, википедия, цифровые книги, статьи, журналы и т.п. (~570 гб для GPT-3)
BioBERT [5]	Биомедицина, извлечение информации из научно-медицинских текстов	Основана на BERT; дообучена на биомедицинских текстах; улучшена интерпретируемость в медицинских задачах	PubMed, PMC (биомедицинские статьи) (~18 гб)
ClinicalBERT [6]	Клинические заметки, прогнозирование госпитализаций	Основана на BERT; дообучена на клинических записях; адаптирована для медицинских терминов	MIMIC-3 (корпус с клиническими записями) (~600мб)
RuBERT [7]	Обработка русского языка, анализ тональности	12 слоёв; 180 млн параметров; дообучена на русскоязычных текстах; высокая интерпретируемость для рус. языка	Википедия, новостные статьи на русском языке (~12 гб)
T5 [7]	Универсальный NLP (перевод, математика, анализ данных)	Однонаправленный Transformer; около 11 млрд параметров	Colossal Clean Crawled Corpus (~750 гб)
DistilBERT [8]	Мобильные и быстрые приложения	Упрощённая версия BERT; до 6 слоёв, быстрее в 2 раза при сохранении 95% качества	Книги, википедия, статьи (~17 гб)
XLM-RoBERTa [9]	Перевод, многоязычные классификации	До 550 тыс. токенов, 100 языков, большая глубина обработки	CommonCrawl (~2,5 тб)

Продолжение таблицы 1

1	2	3	4
SciBert [10]	Научные публикации	Дообучен на научных статьях	1,14 млн статей Semantic Scholar (~3,2 гб)
mBERT [11]	Многоязычные задачи	104 языка; 12 слоёв	Wikipedia на 104 языках (~110 гб)
Reformer [12]	Работа с длинными документами	Transformer с долгой памятью	Enwik8 и текстовые выборки (~10 гб)
LongFormer [13]	Обработка длинных текстов (более 100.000 символов)	Использованы механизмы global attention; упрощённая интерпретируемость	ArXiv, Wikipedia (~40 гб)
SqueezeBERT [14]	NLP для мобильных устройств	Свёрточные attention-блоки; низкое ожидание ответа	Wikipedia, BooksCorpus (~16 гб)
MiniLM [15]	Лёгкие NLP задачи	22 млн параметров, высокая точность, дистилляция	Wikipedia, BooksCorpus (~16 гб)
T5 1.1 [16]	Улучшенный T5	Улучшен препроцессинг, оптимизация обучения	C4 (~750 гб)
PalM [17]	Универсальный ИИ, высокая интерпретируемость	540 млрд параметров	C4, Wikipedia, BooksCorpus (~780 ГБ)
YaLM (SberAI) [18]	Русскоязычные генеративные задачи	Архитектура GPT, оптимизирована под русский язык, 100 млрд параметров	Large Russian Text Courpus (~100 гб)
GShard [19]	Мультиязычные и масштабируемые системы	Обучение параллельно на 600+ языках	GShard corpus (~2 тб)
Godel [20]	Универсальный генератор диалогов	Использует T5 как основу, углублён в на-учные знания	WebGPT, Wizrd of Wikipedia (~60 гб)
DialoGPT [21]	Диалоговые системы с разными персонажами из игр\фильмов	GPT-2, специализирован для диалогов	Reddit conversations dataset (~147 млн. Диалогов, ~40 гб)

Как видно из табл. 1, модели демонстрируют разнообразие в архитектуре, объёмах обучающих данных и областях применения.

Проблема интерпретируемости (explainability) нейросемантических моделей приобретает особую значимость. Под интерпретируемостью в данной работе понимается способность модели предоставлять понятное объяснение своих решений, структуры и внутренних представлений (эмбеддингов), достаточное для их оценки человеком-экспертом. В отличие от традиционных алгоритмов машинного обучения (например, решающих деревьев или логистической регрессии), поведение глубоких нейросетевых моделей трудно поддаётся интуитивному осмыслению без дополнительных инструментов [22].

Проблемы интерпретации нейросемантических моделей

Современные нейросемантические модели обладают высокой производительностью, однако их применение сопряжено с рядом затруднений, связанных с недостаточной прозрачностью и воспроизводимостью их решений. К числу ключевых проблем интерпретируемости относятся следующие [23].

1. Непрозрачность векторных представлений при анализе смысла. Пути решения: разложение (PCA, ICA), кластеризация, выделение семантических осей.
2. Недостаточная прослеживаемость выводов для эксперта. Пути решения: трассировка активаций, использование attention, интерпретируемые промежуточные слои.
3. Контекстная изменчивость. Методы совершенствования: отслеживание семантических сдвигов, контекстные аннотации.

4. Уязвимость к небольшим искажениям. Методы совершенствования: регуляризация, симметричные архитектуры, контрастивное обучение.
5. Расхождения с логикой мышления эксперта. Пути решения: обучение с учётом пользовательских оценок, интеграция онтологических знаний.
6. Уязвимость к искажениям и атакам. Методы защиты: adversarial training, устойчивые эмбеддинги, фильтрация.
7. Отсутствие универсальных метрик интерпретируемости. Решение проблемы: разработка формальных метрик, стандартизация пользовательских опросов.

Нейросемантические модели, как правило, основаны на трансформерных архитектурах и обучаются на больших неразмеченных корпусах, кодируя знания в виде плотных эмбеддингов. Их многослойная структура и нелинейность усложняют понимание логики вывода, что снижает прозрачность и доверие. В свою очередь, интерпретируемость позволяет обеспечивать доверие пользователей, соблюдение правовых и этических норм, отладку и верификацию модели, снижение ошибок в критически важных задачах, а также анализ и улучшение архитектур. Интерпретируемые модели обеспечивают лучшую диагностику, контроль, адаптацию и повышение надёжности при внедрении в реальные приложения.

Современные методы интерпретации нейросемантических моделей

В условиях увеличивающейся сложности архитектур, задача интерпретации работы нейросемантических моделей становится всё более актуальной. Современные подходы к объяснению решений таких моделей принято делить на две основные группы**:** встроенные (intrinsic) и постобучающие (post hoc) методы [24].

Встроенные методы предполагают закладывание интерпретируемости непосредственно в архитектуру модели на этапе проектирования **(**Concept Bottleneck Models (CBM); модульные архитектуры; интерпретируемые эмбеддинги; механизмы внимания (Attention Mechanisms))**.**

Постобучающие методы применяются к уже обученным моделям и не требуют модификации их внутренней структуры:

- Attention-based интерпретация [25];
- LIME (Local Interpretable Model-Agnostic Explanations), **м**етод локального приближения;
- SHAP (SHapley Additive exPlanations) – метод, использующий теорию игр [26];
- Embedding Projection – визуализация векторных представлений;
- Rationales – методы построения минимального подмножества входа, достаточного для сохранения прежнего предсказания [27].

В таблице 2 представлена сравнительная характеристика отдельных методов.

Таблица 2 – Краткий обзор актуальных решений

Метод	Интерпрети-руемость	Применимость к NLP	Основные ограничения
Attention maps	Умеренная	Высокая	Не всегда объясняют причинность
SHAP	Высокая	Средняя	Высокая вычислительная сложность
LIME	Средняя	Высокая	Чувствительность к параметрам
Rationales	Высокая	Высокая	Требует обучения специальных дополнительных моделей
CBM	Очень высокая	Средняя	Для обучения требуются тщательно аннотированные данные

Глобальная и локальная интерпретируемость

Интерпретируемость моделей машинного обучения в зависимости от масштаба анализа и целей применения делится на глобальную и локальную. Глобальная интерпретируемость позволяет понять общую логику функционирования модели: какие признаки вносят наибольший вклад в предсказания, как распределяются веса или важности признаков, как модель принимает решения. Такой уровень анализа полезен, например, при проверке соответствия модели нормативным требованиям, оценке устойчивости поведения модели или при передаче результатов заинтересованным сторонам. К моделям с высокой глобальной интерпретируемостью относятся линейные классификаторы, решающие деревья и обобщённые аддитивные модели (GAM).

Локальная интерпретируемость, напротив, фокусируется на объяснении конкретного решения, принятого моделью для одного входного примера. Это позволяет ответить на вопрос: *почему модель приняла именно это решение в данной ситуации?* Такой подход особенно востребован в юридических приложениях, в онлайн-образовании. Для сложных моделей вроде трансформеров или нейросетей, методы локальной интерпретации позволяют «подсветить» важные фрагменты входных данных, например, слова, которые повлияли на классификацию текста как токсичного или нейтрального.

На практике используются следующие методы:

- LIME, позволяющий аппроксимировать поведение модели линейной моделью в окрестности конкретного примера;
- SHAP, вычисляющий вклад каждого признака в предсказание;
- Анализ внимания (attention weights) в трансформерах – визуализация какие токены модель считает более значимыми при генерации или классификации.

Таким образом, глобальная интерпретируемость необходима для построения доверия, анализа стабильности модели и объяснения её работы в целом. В свою очередь, локальная интерпретируемость критична в прикладных задачах, где каждое отдельное решение должно быть прозрачным, обоснованным и проверяемым.

Компромисс между интерпретируемостью и точностью в прикладных задачах

При разработке нейросемантических моделей ключевой вызов заключается в поиске баланса между интерпретируемостью (способностью объяснить логику вывода) и предсказательной точностью**.** В разных прикладных областях оптимальные акценты могут сильно различаться:

1. **Медицина**

○ *Ситуация:* В системе поддержки принятия клинических решений важно не только вовремя обнаружить отклонение, но и обосновать выбор диагноза.

○ *Подход:* используют гибридные схемы: сначала быстрый интерпретируемый классификатор (логистическая регрессия или дерево решений) отбирает подозрительные случаи, а затем для уточнения применяют глубокую нейросеть с визуализацией активаций (Grad-CAM) или SHAP-анализом локальных вкладов признаков.

2. **Финансовая экспертиза**

○ *Ситуация:* при оценке кредитного риска или выявлении мошенничества нормативы требуют объяснять, почему клиенту отказано в кредите или почему операция помечена как подозрительная.

○ *Подход:* здесь часто комбинируют обобщённые аддитивные модели (GAM) для глобального анализа факторов риска с LIME/SHAP для локального обоснования

каждого решения. Для усиления точности к схемам иногда добавляют «чёрный ящик» (например, градиентный бустинг) и поверх него натягивают объяснитель.

3. **Юриспруденция и комплаенс**

○ *Ситуация:* Системы семантического анализа правовых документов должны обосновать, на каких положениях закона основано их заключение.

○ *Подход:* применяют модульные архитектуры**,** где один модуль («семантический движок») извлекает эмбеддинги нормативных актов, а второй – лёгкая логическая система на базе правил или решающих деревьев, формирующая объяснение на естественном языке.

4. **Образование**

○ *Ситуация:* Рекомендательные системы для онлайн-курсов предлагают студенту следующий шаг в обучении. Здесь важно понять, какие ответы или ошибки самого студента повлияли на рекомендацию.

○ *Подход:* используют **двухэтапную модель**: сначала нейросеть предсказывает потребность в повторении темы, затем к ней привязывают LIME-анализ, чтобы показать преподавателю ключевые вопросы или фрагменты теста, где у студента возникли затруднения.

5. **Промышленность и IoT**

○ *Ситуация:* При обслуживании оборудования нужно своевременно обнаружить аномалию и объяснить, какие параметры вышли за рамки нормы.

○ *Подход:* часто строят **ансамбли** из простых моделей (например, одноранговая кластеризация для обнаружения выбросов) и RNN/трансформеров для прогноза. Вершинный класс кластеризатора служит объяснителем, указывая, какие датчики «отклонились» от кластера нормального режима.

Подходы к интерпретации нейросемантических моделей

Интерпретируемость нейросемантических моделей может быть достигнута за счёт различных стратегий, которые условно можно разделить на структурные**,** постхоковые **и** гибридные подходы. Ниже представлены ключевые направления, сжатое описание которых соответствует научной задаче выделения объяснимых компонентов в сложных языковых моделях [29].

1. Функционально-дистрибутивные модели (Functional Distributional Semantics). Преимущества: возможность формализованного вывода, структурной декомпозиции и семантической композиций. Ограничения: высокая вычислительная сложность и трудоёмкость построения аннотированных корпусов.
2. Методы локальной пост-хок интерпретации. Преимущества: модель-агностичность, гибкость применения. Недостатки: локальность, нестабильность при повторных запусках, сложность интерпретации в высокоразмерных пространствах.
3. Интерпретируемые архитектуры. Преимущества*:* высокая совместимость с современными трансформерами. Недостатки: не всегда внимание коррелирует с причинностью, а sparsity может снижать точность.
4. Методы декодирования скрытых представлений. Преимущества: высокая наглядность. Недостатки*:* неоднозначность декодирования и зависимость от внешних словарей.
5. Контрастивные и объясняющие задачи. Преимущества: высокая согласуемость с человеческими интерпретациями. Ограничения: необходимость дополнительной разметки и усложнение архитектуры.

Интерпретируемость в текстовых моделях: особенности, сложности, перспективы

Интерпретируемость моделей в области NLP сталкивается с уникальными трудностями, обусловленными спецификой текстовых данных и архитектурой современных моделей. В отличие от изображений или табличных данных, текст обладает высокой степенью абстракции, неоднозначностью и вариативностью выражения одного и того же смысла, что усложняет извлечение прозрачных интерпретаций.

Особенности работы с текстовой информацией: линейность и дискретность текста; семантическая неоднозначность; привязка к языковому и культурному фону; неустойчивость к переформулировкам.

Основные сложности интерпретации: высокая размерность входных данных; неявность признаков; контекстуальность значений; глубокая архитектурная сложность.

Сегодня применяются следующие основные подходы к интерпретации: методы визуализации внимания (Attention-based); локальные методы интерпретации (LIME, SHAP); поиск рациональных обоснований (Rationale extraction**)**; проекция эмбеддингов; построение логико-семантических цепочек**.**

Перспективы и направления развития моделей

Перспективы в области интерпретируемости нейросемантических моделей связаны с развитием новых методов и инструментов, способных эффективно объяснять работу сложных архитектур. Использование методов визуализации эмбеддингов и анализа внимания в контексте моделей Transformer открывает возможности для более глубокого понимания их поведения. Например, в [31] предложен метод выравнивания эмбеддингов с семантическими признаками, что делает представления более интерпретируемыми и способствует интеграции нейросетей в семантически насыщенные системы.

Разработка таких подходов не только повысит доверие пользователей к искусственному интеллекту, но и расширит его применение в критически важных областях, таких как медицина, право и финансы. Исследователи подчеркивают, что «изучение возможностей нейросетей в решении математических задач позволяет определить пределы применимости искусственного интеллекта, а также диапазон задач, которые способствуют развитию самого ИИ» [32]. Интерпретируемость моделей станет важным шагом в определении их возможностей и ограничений.

В последние годы в области интерпретируемости нейросемантических моделей сформировались несколько многообещающих направлений, способных существенно расширить возможности объяснения поведения сложных архитектур. Ниже приведены ключевые векторы развития с указанием соответствующих источников.

1. Развитие self-interpretable архитектур [33]. Создание моделей, изначально нацеленных на объяснимость, без необходимости внешних интерпретирующих модулей. К таким подходам относятся атрибутивные, функционально-базированные, концепт-базированные, прототипные и правило-ориентированные модели, которые раскрывают логику вывода внутри самой архитектуры.
2. Механистическая интерпретируемость для безопасности ИИ [32]. Исследования, направленные на восстановление и анализ скрытых алгоритмов сложных моделей (например, трансформеров), с целью выявления потенциально опасных или неверных паттернов поведения и обеспечения надёжности систем.
3. Нейросимволическое объяснение через концепт-базированные модели [34]**.** Интеграция символических онтологий и концептов в распределённые эмбеддинги позволяет строить *Deep Concept Reasoner*, который генерирует интерпретируемые логические правила на основе семантических признаков.

4. Структурные *Neural Additive Models (SNAMs)* [35]. Совмещение классических статистических методов с мощью DNN через *Structural Neural Additive Models,* где каждая компонентная функция остаётся простой и визуализируемой, обеспечивая баланс между точностью и прозрачностью.
5. Унификация и стандартизация метрик интерпретируемости [36]. Разработка общепринятых критериев оценки объяснимости (*fidelity, stability, comprehensibility*), создание бенчмарков для NLP-задач с пост-хок методами (LIME и SHAP).
6. Композиционность и причинная интерпретация в NLP [37]**.** Исследования в области *compositional behavior* и *causal probe-based* анализов, направленные на понимание того, как модели формируют сложные семантические конструкции и причинно-следственные связи при работе с текстом.
7. Генерация естественноязычных объяснений [38]**.** Разработка подходов, при которых модели самостоятельно формируют пояснения на естественном языке, что повышает доступность интерпретации для неквалифицированных пользователей и облегчает верификацию решений.

Выводы

В настоящей статье проведен комплексный анализ подходов к обеспечению интерпретируемости нейросемантических моделей, используемых в задачах обработки естественного языка. Рассмотрены как встроенные, так и пост-хок методы объяснения, включая attention-механизмы, проекцию эмбеддингов и локальные методы интерпретации. Особое внимание уделено современным проблемам интерпретируемости: непрозрачность векторных представлений, нестабильность контекста, чувствительность к искажениям и отсутствие универсальных метрик. Выявлен компромисс между точностью и интерпретируемостью, особенно актуальный для глубинных моделей. Обоснована значимость интерпретируемости в критически важных приложениях, включая медицину, право и автономные системы. Представлены актуальные направления исследований: self-interpretable архитектуры, нейросимво-лические модели, генерация естественноязычных объяснений и развитие метрик.

Таким образом, интерпретируемость следует рассматривать не как вспомогательную характеристику, а как ключевое условие надёжности и этической приемлемости ИИ-систем. Будущие исследования должны быть направлены на формализацию объяснений, унификацию критериев интерпретируемости и разработку гибридных моделей, сочетающих точность и прозрачность.

Список литературы

1. Модели нейросетей: что такое и чем отличаются друг от друга // unisender:сайт – 2025 (24.05.2025).
2. Amodei D. Concrete Problems in AI Safety // arXiv. 2016. URL: https://arxiv.org/pdf/1606.06565 (accessed May 2, 2025).
3. Жарова, М. Модели BERT для машинного обучения // habr. 2024. URL: https://habr.com/ru/companies/skillfactory/articles/862130 (28.05.2025).
4. ChatGPT 2025 // habr – 2025, URL: https://habr.com/ru/companies/bothub/articles/887988/ (accessed May 28, 2025).
5. BioBERT – модель обработки биомедицинских текстов // neurohive – 2019, URL: https://neurohive.io/ru/novosti/biobert/ (28.05.2025).
6. ClinicBERT // HuggingFace – 2023, URL: https://huggingface.co/tdobrxl/ClinicBERT (accessed May 2, 2025).
7. RuT5, RuRoBERTa, RuBERT: как мы обучили серию моделей для русского языка // habr – 2021, URL: https://habr.com/ru/companies/sberdevices/articles/567776/ (28.05.2025).
8. DistilBERT: The compact NLP Powerhouse // opengenus – 2023, URL: https://iq.opengenus.org/ distilbert/ (accessed May 28, 2025).
9. XLM-RoBERTa // huggingface, URL: https://huggingface.co/docs/transformers/model_doc/xlm-roberta (accessed May 28, 2025).

10. Beltagy L., Lo K., Cohan A., SciBERT: a pretrained languge model for scientific text // arXiv. 2019, URL: https://huggingface.co/docs/transformers/model_doc/xlm-roberta (accessed May 28, 2025).
11. T. Pires, E. Schlinger, D. Garrette, How multilingual is mBERT? // arxiv. 2020, URL: https://arxiv.org/abs/2005.09093 (accessed May 28, 2025).
12. Reformer – эффективный трансформер // habr. 2020, URL: https://habr.com/ru/articles/522622/ (29.05.2025).
13. L. Beltagy, M. Peters, A. Cohan, LongFormer: the long-document Transformer // arxiv. 2020, URL: https://arxiv.org/abs/2004.05150 (accessed May 29, 2025).
14. F. Landola, A. Shaw, R. Krishna, K. Keutzer, SqueezeBERT: what can computer vision teach NLP about efficient neural networks? // arxiv. 2020, URL: https://arxiv.org/abs/2006.11316 (accessed May 29, 2025).
15. W. Wang, et al. MiniLM: Self Attention distillation for task-agnostic compression of pre-trained Transformers // arxiv. 2020, URL: https://arxiv.org/abs/2002.10957 (accessed May 29, 2025).
16. Google T5 1.1 model // metatext – 2024, URL: https://metatext.io/models/google-t5-v1_1-xxl (accessed May 29, 2025).
17. Новая языковая модель Google PalM // habr – 2022, URL: https://habr.com/ru/news/659603/ (28.05.2025).
18. Яндекс выложил YaLM – крупнейшая GPT-подобная нейросеть // habr – 2022, URL: https://habr.com/ru/companies/yandex/articles/672396/ (28.05.2025).
19. Lepikhin D., et al. GShard: Scaling giant models with conditional computation and automatic Sharding // arxiv. 2020, URL: https://arxiv.org/abs/2006.16668 (accessed May 29, 2025).
20. GODEL: может ли чатбот работать в коротких беседах? Майкрософт говорит да! // analyticsindiamag 2022, URL: https://analyticsindiamag.com/global-tech/can-a-chatbot-indulge-in-small-talks-microsoft-says-yes/ (29.05.2025).
21. DialoGPT – a state of the art large-scale pretrained response generation model // huggingface, URL: https://huggingface.co/microsoft/DialoGPT-medium (accessed May 29, 2025).
22. Zhao H., et al. Explainability for Large Language Models: A Survey // *arXiv preprint*. 2024. arXiv: 2309.01029. URL: https://arxiv.org/pdf/2309.01029 (accessed May 29, 2025).
23. Räuker T., Ho A., Casper S., Hadfield-Menell D. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks // arxiv. 2022, URL: https://arxiv.org/pdf/2207.13243 (accessed May 28, 2025).
24. Xia B., Wang X., Yamasaki T., Semantic Explanation for Deep Neural Networks Using Feature Interactions // ACM Transactions on Multimedia Computing, Communications, and Applications. V.17, no2, 2021: 1–20. DOI: 10.1145/3474557.
25. Ермоленко Т.В., Классификация ошибок в тексте на основе глубокого обучения // Проблемы искусственного интеллекта. Т.3, №14, 2019. – С. 47–57.
26. Смирнов И.В., Интеллектуальный анализ текстов на основе методов разноуровневой обработки естественного языка. – М.: ФИЦ ИУ РАН, 2023. – 354 с.
27. Sun X., et al. Interpreting Deep Learning Models in Natural Language Processing: A Review // *arXiv*, 2021. arXiv:2110.10470. URL: https://arxiv.org/pdf/2110.10470 (accessed May 29, 2025).
28. Ji Y., et al. A Comprehensive Survey on Self-Interpretable Neural Networks // *arXiv*, 2025. arXiv: 2501.15638. URL: https://arxiv.org/pdf/2501.15638 (accessed May 2, 2025).
29. Анцыферов С.С. Методология развития интеллектуальных систем/ С.С. Анцыферов, А.С. Сигов, К.Н. Фазилова. Проблемы искусственного интеллекта, №2(25), 2022. – С.42–47.
30. Bereska L., Gavves S. Mechanistic Interpretability for AI Safety // OpenReview:сайт, URL: https://arxiv.org/pdf/2501.15638 (accessed May 29, 2025).
31. Старченко С.Н., Рудаков А.В. Построение интерпретируемых моделей для анализа пользовательских отзывов // Информационные технологии и вычислительные системы. 2020. №3. – С. 73–83.
32. Luber M., Thielmann A., Safken B. Structural Neural Aditive Models: enhanced Interpretable Machine Learning // arxiv:сайт, URL: https://arxiv.org/pdf/2302.09275 (accessed May 2, 2025).
33. Madsen A., Reddy S., Chandar S. Post-Hoc Interpretability for Neural NLP: A Survey // ACM Computing Surveys, no55(155), 2022: 1–42.
34. Киселёв С.А. Архитектуры современных языковых моделей: от BERT до GPT-3 // Системы и средства информатики, Т.33, №2, 2023. – С. 95–114.
35. Мельникова Н.Н., Шкляр В.Л. Методы интерпретации решений нейросетевых моделей в задачах анализа текста // Проблемы программирования, №5, 2021. – С. 102–110.
36. Жураев И.И., Попов С.В. Интерпретируемость алгоритмов машинного обучения в задачах медицины // Искусственный интеллект и принятие решений, №1, 2022. – С. 54–64.
37. Чупров И.Ю., Фёдоров В.А. Интерпретируемость нейросетевых моделей: проблемы и подходы // Вестник Московского университета. Серия 1: Математика и механика, №6, 2023. – С. 41–56.
38. Смирнов А.В., Кузнецова Е.П. Анализ композиционности в нейросетевых языковых моделях // Вестник компьютерных и информационных технологий, №2, 2024. – С. 58–67.

References

1. Model neuroseteum: Chto takoe I Chem otlichayutsya drug OT druga // unisender: site – 2025 (accessed May 2, 2025).
2. Amodei D. Concrete Problems in AI Safety. // archive. 2016. URL: https://arxiv.org/pdf/1606.06565 (accessed May 2, 2025).
3. Jarowa M. Model BERT dlya mashinnogo obucheniya // habr. 2024. URL: https://habr.com/ru/companies/skillfactory/articles/862130 (accessed May 28, 2025).
4. ChatGPT 2025 // habr-2025, URL: https://habr.com/ru/companies/bothub/articles/887988/ (accessed May 28, 2025).
5. BioBERT-model obrabotki biomedisinskix tekstov // neurohive. 2019, URL: https://neurohive.io/ru/novosti/biobert/ (accessed May 28, 2025).
6. ClinicBERT // HuggingFace-2023, URL: https://huggingface.co/tdobrxl/ClinicBERT (accessed May 2, 2025).
7. RuT5, RuRoBERTa, RuBERT: Kak Mi obuchili seriyu modeley dlya russkogo yazika // habr – 2021, URL: https://habr.com/ru/companies/sberdevices/articles/567776/ (accessed May 28, 2025).
8. Distillbert: the compact NLP Powerhouse // opengenus-2023, URL: https://iq.opengenus.org/distilbert/ (accessed May 28, 2025).
9. XLM-RoBERTa // huggingface, URL: < BR > https://huggingface.co/docs/transformers/model_doc/xlm-roberta (accessed May 28, 2025).
10. Beltagy L., Lo K., Cohan A. SciBERT: a pretrained languge model for scientific text // archive – 2019, URL: < br > https://huggingface.co/docs/transformers/model_doc/xlm-roberta (accessed May 28, 2025).
11. T. Pires, E. Schlinger, D. Garrett, How multilingual is mBERT? // archive-2020, URL: https://arxiv.org/abs/2005.09093 (accessed May 28, 2025).
12. Reformer-effektivniy transformer // habr. 2020, URL: https://habr.com/ru/articles/522622/ (accessed May 29, 2025).
13. L. Beltagy, M. Peters, A. Cohan, LongFormer: the long-document Transformer // archive – 2020, URL: < BR > https://arxiv.org/abs/2004.05150 (accessed May 29, 2025).
14. F. Landola, A. Shaw, R. Krishna, K. Keutzer, SqueezeBERT: what can computer vision teach NLP about efficient neural networks? // archive-2020, URL: https://arxiv.org/abs/2006.11316 (accessed May 29, 2025).
15. W. Wang, et al. MiniLM: Self Attention distillation for task-agnostic compression of pre-trained Transformers // archives. 2020, URL: < BR > https://arxiv.org/abs/2002.10957 (accessed May 29, 2025).
16. Google T5 1.1 model // metatext-2024, URL: https://metatext.io/models/google-t5-v1_1-xxl (accessed May 29, 2025).
17. Novaya yazikovaya model Google PalM // habr-2022, URL: https://habr.com/ru/news/659603/ (accessed May 28, 2025).
18. Yandex vilojil YaLM-krupneyshaya GPT-podobnaya neuroset // habr. 2022, URL: https://habr.com/ru/companies/yandex/articles/672396/ (accessed May 28, 2025).
19. Lepikhin D., et al. GShard: Scaling giant models with conditional computing and automatic Sharding // archive. 2020, URL: < BR > https://arxiv.org/abs/2006.16668 (accessed May 29, 2025).
20. GODEL: mojet Lee chatbot rabotat v korotkix besedax? Mycrosoft Govorit da! // analyticsindiamag-2022, URL: https://analyticsindiamag.com/global-tech/can-a-chatbot-indulge-in-small-talks-microsoft-says-yes/ (accessed May 29, 2025).
21. DialoGPT – a state of the art large-scale pretrained response generation model // huggingface, URL: < br > https://huggingface.co/microsoft/DialoGPT-medium (accessed May 29, 2025).
22. Zhao H., et al. Explainability for Large Language Models: a Survey // archive preprint. 2024. archive: 2309.01029. URL: https://arxiv.org/pdf/2309.01029 (accessed May 29, 2025).
23. Räuker T., Ho A., Casper S., Hadfield-Menell D. Toward Transparent AI:a Survey on Interpreting the Inner Structures of Deep Neural Networks // archive: site. – 2022, URL: https://arxiv.org/pdf/2207.13243 (accessed May 28, 2025).
24. Xia B., Wang X., Yamasaki T., Semantic exploration for Deep Neural Networks Using Feature Interactions // ACM Transactions on Multimedia Computing, Communications, and Applications. V.17, no 2, 2021: 1–20. DOI: 10.1145/3474557.
25. Ermolenko T.V., Classification oshibok V tekste na osnove glubokogo obucheniya // problemi iskusstvennogo intellekta. V.3, no 14, 2019: 47–57.
26. Smirnov I.V., Intellectualny analiz tekstov na osnove metodov raznourovnevoy obrabotki estestvennogo yazika. Moscow, FITS IO RAN, 2023, 354 p.

27. Sun X., et al. Interpreting Deep Learning Models in Natural Language Processing: a Review // archive, 2021. archive: 2110.10470. URL: https://arxiv.org/pdf/2110.10470 (accessed May 29, 2025).
28. Ji Y., et al. A Comprehensive Survey on Self-Interpretable Neural Networks // archive, 2025. archive: 2501.15638. URL: https://arxiv.org/pdf/2501.15638 (accessed May 2, 2025).
29. Antsiferov S.S. Methodology razvitia intellectualnix system/ s.S. Antsiferov, A.S. Sigov, K.N. Fazilova. Problemi iskusstvennogo intellekta, no 2(25), 2022: 42–47.
30. Beresca L., Gavves S.. Mechanistic Interpretability for AI Safety // OpenReview: site, URL: https://arxiv.org/pdf/2501.15638 (accessed May 29, 2025).
31. Starchenko S.N., Rudakov A.V. Postroenie interpretiruemix modeley dlya analiza polzovatelskix otzivov // Informasionnie technologii I vichislitelnie system, no 3, 2020: 73–83.
32. Luber M., Thielmann A., Safken B. Structural Neural Aditive Models: enhanced Interpretable Machine Learning / / archive: site, URL: https://arxiv.org/pdf/2302.09275 (accessed May 2, 2025).
33. Madsen A., Reddy S., Chandar S. Post-Hoc Interpretability for Neural NLP: a Survey // ACM Computing Surveys, no 5(155), 2022: 1–42.
34. Kiselyov S.A. Architekturi sovremennix yazikovix modeley: OT BERT do GPT-3 // Sistemi I sredstva informatiki, V.33, no 2, 2023: 95–114.
35. Melnikova N.N., Shklyar V.L. Method interpretasii resheniy neurosetevix modelei V zadachax analiza texta / / problemi programmirovaniya, no 5, 2021: 102–110.
36. Zhuraev I.I., Popov S.V. Interpretiruemost algorithmov mashinnogo obucheniya V zadachax mediisini / / iskusstvenniy intellect I prinyatie resheniy, no 1, 2022: 54–64.
37. Chuprov I.Yu., Fyodorov V.A. Interpretiruemost neurosetevix modeley: problemi I podhodi // Vestnik Moskovskogo universiteta. Series 1: Mathematics I mechanics, no 6, 2023: 41–56.
38. Smirnov A.V., Kuznesova E.P. Analysis kompozisionnosti v neurosetevix yazikovix modelyax // Vestnik komiternix I informasionnix technologiy, no 2, 2024: 58–67.

RESUME

K. A. Nikitenko, A. V. Zviagintseva

Interpretability of Neurosemantic Models in Applied Domains

Neurosemantic models based on deep learning and transformer architectures (e.g., BERT, GPT) are widely used in applied domains such as medicine, law, and computational linguistics. However, their complexity and opacity raise critical concerns about transparency, safety, and user trust. Interpretability is becoming a vital property for AI systems deployed in high-risk environments.

The article provides a structured review of interpretability types (global vs. local), methods (intrinsic and post-hoc), and evaluation approaches. A comparative table of modern neurosemantic models is included, presenting key features such as parameter count, domain-specific training data, and area of application. The study also discusses common interpretability issues: embedding opacity, contextual instability, lack of universal metrics, and sensitivity to input perturbations. Methodologies like attention visualization, PCA, LIME, SHAP, and symbolic grounding are reviewed and critically analyzed.

The analysis reveals a trade-off between interpretability and accuracy. While models with high precision often function as “black boxes,” interpretable models may sacrifice performance. A taxonomy of methods to overcome this trade-off is presented. Applied examples from clinical NLP (BioBERT, ClinicalBERT) and legal AI systems are discussed. The article also highlights emerging self-interpretable and neuro-symbolic models as promising solutions.

Interpretability should be considered a core requirement in neurosemantic modeling, not an optional feature. Future research should focus on developing hybrid models, integrating human-readable explanations, and standardizing interpretability benchmarks. These advances are crucial for building trustworthy, reliable AI systems in sensitive applications.

РЕЗЮМЕ

К. А. Никитенко, А. В. Звягинцева
Интерпретируемость нейросемантических моделей при их применении в прикладных областях

В статье исследуется проблема интерпретируемости нейросемантических моделей, получивших широкое распространение в различных прикладных сферах – от медицины и юриспруденции до информационной безопасности, цифровой лингвистики и технической диагностики. Акцент сделан на необходимость понимания внутренней логики нейросетей, особенно трансформерных архитектур (BERT, GPT, T5 и др.), в условиях их высокой точности и одновременно ограниченной прозрачности.

Представлена классификация подходов к интерпретируемости: по типу (глобальная и локальная), по способу реализации (встроенные и пост-хок методы), а также по формату выходной информации (визуализация, текстовые пояснения, атрибутивные признаки). Рассматриваются такие методы, как LIME, SHAP, attention-механизмы, PCA, embedding projection и пр.

Особое внимание уделено анализу компромисса между интерпретируемостью и точностью: приведены примеры архитектурных решений, позволяющих минимизировать потери в производительности при сохранении объяснимости. В статье представлена сводная таблица с характеристиками популярных нейросемантических моделей, данными их обучения и сферами применения. Анализируются такие модели как BioBERT, ClinicalBERT, RuBERT, T5, DistilBERT, Galactica, GatorTron, PaLM и другие, с фокусом на их способности обеспечивать интерпретируемые результаты в различных предметных областях.

Выделены ключевые проблемы интерпретации: непрозрачность векторных представлений, нестабильность интерпретаций в контексте, уязвимость к атакующим примерам, слабая согласуемость с человеческими объяснениями, отсутствие универсальных метрик. Для каждой проблемы предложены возможные пути устранения или снижения влияния.

Также рассмотрены перспективные направления развития: self-interpretable архитектуры, нейросимволические модели, использование онтологий, концепт-базированные и модульные архитектуры, генерация естественноязыковых пояснений, стандартизация метрик интерпретируемости. Подчёркнута роль интерпретируемости как необходимого условия внедрения ИИ в критически значимые сферы.

Никитенко Кирилл Андреевич – аспирант кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», 283001, Донецк, ул. Университетская, 24, n1kitenkok@yandex.ru. *Область научных интересов:* компьютерное зрение, машинное обучение, нейронные сети. Число научных публикаций – 5.

Звягинцева Анна Викторовна – д.т.н., доцент, профессор кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», 283001, Донецк, ул. Университетская, 24, zviagintsevaav@gmail.com. *Область научных интересов*: системный анализ, событийная и комплексная оценка; безопасность и управление социально-экономическими и техногенными системами; информационно-аналитические системы; обработка и анализ данных. Число научных публикаций – более 150.

Статья поступила в редакцию 01.06.2025