Моделирование и анализ данных
2025, *15*(4), 156—164.
https://doi.org/10.17759/mda.2025150410
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2025, *15*(4), 156—164.
https://doi.org/10.17759/mda.2025150410
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

КРАТКИЕ СООБЩЕНИЯ | BRIEF MESSAGES

Научная статья | Original paper

УДК 004.885

Семантический анализ результатов выполнения тестовых заданий

Б.Ю. Поляков
Московский государственный психолого-педагогический университет
Москва, Российская Федерация
✉ deslion@yandex.ru

Резюме

Контекст и актуальность. Автоматизация проверки открытых ответов в образовательном и профессиональном тестировании остается сложной задачей из-за необходимости привлечения экспертов для разметки данных. Современные большие языковые модели открывают новые возможности генерации синтетических данных для обучения классификаторов. **Цель.** Оценить возможность применения синтетических данных, сгенерированных большими языковыми моделями, для обучения автоматических классификаторов текстовых ответов в тестировании. **Методы.** Эксперимент включал генерацию 100 примеров ответов через LLM, их предобработку (токенизация, стемминг, TF-IDF) и обучение двух моделей — логистической регрессии и RBF-сети с последующей оценкой на тестовой выборке. **Результаты.** Точность моделей составила 80% и 65—90% соответственно. Обнаружены системные ограничения: зависимость от ключевых слов, нечувствительность к семантическим инверсиям, контекстуальная слепота. **Выводы.** Подход перспективен для создания вспомогательных инструментов оценки, но текущие ограничения не позволяют полностью заменять эксперта-оценщика.

Ключевые слова: большие языковые модели, генеративный ИИ, автоматизация тестирования, обработка текста

Для цитирования: Поляков, Б.Ю. (2025). Семантический анализ выполнения тестовых заданий. *Моделирование и анализ данных, 15*(4), 156—164. https://doi.org/10.17759/mda.2025150410

Semantic analysis of test responses using synthetic data generation

B.Y. Polyakov
Moscow State University of Psychology and Education, Moscow, Russian Federation
✉ deslion@yandex.ru

Abstract

Context and Relevance. Automating the assessment of open-ended responses in educational and professional testing remains challenging due to the need for expert-labeled data. Modern large language models offer new opportunities for generating synthetic training data for text classifiers. **Objective.** To evaluate the feasibility of using synthetic data generated by large language models for training automated classifiers of text responses in educational and professional testing. **Methods.** The experiment involved generating 100 response examples using LLMs, followed by text preprocessing (tokenization, stemming, TF-IDF) and training two classification models — logistic regression and RBF network, with subsequent evaluation on a test dataset. **Results.** The models achieved accuracy of 80% and 65—90% respectively. Systematic limitations were identified: overdependence on keywords, insensitivity to semantic inversions, and contextual blindness in classification. **Conclusions.** The approach shows promise for developing auxiliary assessment tools, though current limitations prevent complete replacement of human evaluators. Further refinement is needed for practical implementation.

Keywords: LLM, large language models, generative AI, test automatization, text processing

For citation: Polyakov, B.Y. (2025). Semantic analysis of test responses using synthetic data generation. *Modelling and Data Analysis, 15*(4)*,* 156—164. (In Russ.). https://doi.org/10.17759/mda.2025150410

Введение

Тестирование с использованием опросников открытого типа является широко распространенным инструментом оценки знаний в образовательной и корпоративной практике. В отличие от тестов со множественным выбором, данный формат позволяет не только проверить фактологические знания, но и оценить способность респондента к формулированию мыслей, построению аргументации и демонстрации системного понимания предметной области. Эта методика успешно применяется как при проведении аттестаций в учебных заведениях, так и в процессе профессионального отбора кандидатов при трудоустройстве, где требуется оценить практическое владение специализированными компетенциями.

Однако ключевым ограничением метода выступает ресурсоемкость процесса проверки ответов. Каждый текст или устный ответ требует индивидуального анализа и оценки квалифицированным экспертом, что приводит к временным затратам и субъективности в оценивании. В качестве технологического решения данной проблемы предлагается разработка автоматизированной системы бинарной классификации текстовых ответов («правильно» / «неправильно») методами машинного обучения. (Нежников, Марьенков, 2024).

Помимо ресурсоемкости, значительной проблемой остается субъективность экспертной оценки. Разные проверяющие могут по-разному интерпретировать критерии оценки, особенно когда ответы содержат элементы неполного знания или нестандартные формулировки. Эта вариабельность снижает надежность и валидность результатов тестирования, что критически важно как в образовательных экзаменах, так и в профессиональной сертификации. Автоматизация процесса проверки позволила бы не только снизить нагрузку на экспертов, но и стандартизировать оценочные процедуры, обеспечивая единообразие применяемых критериев.

Внедрение подобных систем сопряжено с фундаментальной проблемой — необходимостью предварительного формирования репрезентативной выборки размеченных текстовых данных. Традиционный подход к созданию обучающих наборов требует привлечения предметных экспертов для написания и аннотирования сотен или тысяч примеров, что сопряжено со значительными временными и финансовыми затратами. Проблема недостатка размеченных данных особенно остро стоит в новых или быстро развивающихся дисциплинах, где экспертные сообщества еще не сформировали обширных корпусов текстов для обучения моделей. В таких условиях традиционные подходы к сбору данных становятся узким местом, затягивая разработку и внедрение автоматизированных систем на месяцы и годы. Современные большие языковые модели (Large Language Models, LLM), прошедшие обучение на колоссальных массивах текстовой информации из различных областей знания, предлагают принципиально новое решение (Ребенок, 2024). Они способны не только к осмысленному перефразированию, но и к генерации оригинальных текстовых единиц, релевантных заданной тематике и соответствующих указанным стилистическим и содержательным ограничениям. Это открывает возможность создания «синтетических экспертов» — виртуальных агентов, способных генерировать объемные и семантически разнообразные обучающие выборки «по требованию».

В этой связи представляется обоснованным объединение технологических возможностей: использование интерактивных чат-ботов на основе продвинутых LLM для генерации синтетической обучающей выборки с последующим обучением на этих данных традиционных моделей бинарной классификации текстов. Такой гибридный подход позволяет преодолеть проблему нехватки размеченных данных и создает предпосылки для разработки эффективных систем автоматической оценки открытых ответов в различных предметных областях.

Жизнеспособность данного подхода проверяется на практическом примере, с детальным описанием процесса генерации синтетических данных, подготовки

признаков и обучения нескольких типов классификационных моделей, а также анализом их производительности и ограничений.

Описание метода

Для проверки жизнеспособности предложенного подхода было проведено экспериментальное исследование, состоящее из трех ключевых этапов:

1. Генерация синтетического набора данных с использованием языковых моделей;
2. Предобработка текстовых данных и формирование признаков-предикторов;
3. Обучение и оценка различных классификационных моделей.

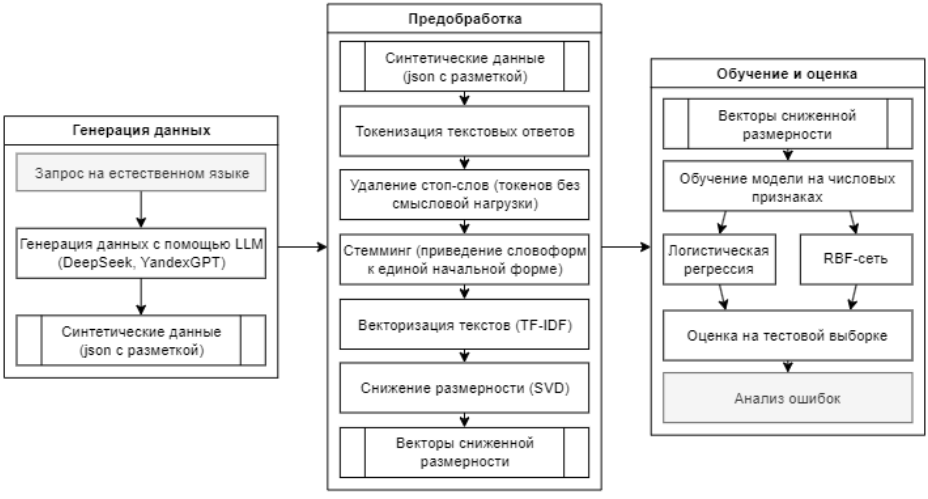

Рис 1. Процесс генерации данных и обучения классификаторов

Fig. 1. Data generation and classifier training process

В качестве примера был выбран следующий вопрос: «С чем связано увеличение количества детей с диагнозом «аутизм» или «расстройство аутистического спектра»?». Для создания размеченных данных использовались современные языковые модели DeepSeek и YandexGPT через веб-интерфейс (при увеличении необходимых объёмов генерации к моделям можно обращаться и через API).

Процесс генерации синтетических данных итеративен. Исходный запрос к языковым моделям неоднократно уточнялся для обеспечения максимального разнообразия генерируемых формулировок и исключения шаблонности. Для класса правильных ответов в запросе на генерацию явно перечислялись ключевые факторы (улучшение диагностики, расширение критериев, рост осведомленности), но при этом требовалось избегать дословного копирования этих формулировок в каждом ответе.

Для класса неправильных ответов использовался список распространенных псевдонаучных мифов, почерпнутых из анализа дискурса в социальных сетях и ненаучных медиа. Это позволило создать выборку, репрезентативно отражающую типичные ошибки и заблуждения, с которыми может столкнуться реальный эксперт.

Всего было сгенерировано 100 текстовых примеров, равномерно распределенных между двумя классами:

- Класс «правильные ответы»: научно обоснованные объяснения, включающие улучшение диагностических методик, расширение диагностических критериев, повышение осведомленности среди медицинских работников и населения
- Класс «неправильные ответы»: распространенные мифы и псевдонаучные концепции, такие как связь с вакцинацией, влияние ГМО, электромагнитного излучения и других факторов, не имеющих научного подтверждения

Для нормализации текстовой информации был реализован общепринятый процесс предобработки данных Python-скриптом (Воронин и др., 2017):

- Токенизация с использованием библиотеки NLTK;
- Удаление стоп-слов русского языка (не имеющие «семантического веса» токены);
- Стемминг (приведение различных словоформ к начальной) с применением SnowballStemmer;
- Векторизация текстов методом TF-IDF (Term Frequency-Inverse Document Frequency);
- Снижение размерности до 10—25 компонент с помощью сингулярного разложения (SVD).

Были исследованы две принципиально различные архитектуры классификаторов (Нежников, Марьенков, 2024):

1. Логистическая регрессия с L2-регуляризацией — классический линейный метод, обеспечивающий интерпретируемость результатов;
2. Нейронная сеть на радиально-базисных функциях (RBF) — нелинейная модель, способная выявлять сложные зависимости в данных.

Выбор моделей классификации был обусловлен необходимостью сравнения интерпретируемости и способности к обобщению. Логистическая регрессия, будучи линейной моделью, позволяет проанализировать веса признаков (ключевых слов) и понять, какие из них наиболее значимы для принятия решения, что ценно для анализа предметной области. В свою очередь, RBF-сеть, как нелинейный классификатор, способна выявлять более сложные, неочевидные зависимости и паттерны в данных, которые могут быть недоступны линейным методам, потенциально повышая точность, но жертвуя при этом интерпретируемостью.

Обучающая выборка была разделена в соотношении 50/50 для обучения и тестирования моделей. Для RBF-сети дополнительно использовалось разделение 80/20 с выделением валидационной подвыборки для контроля переобучения.

Результаты

Экспериментальное исследование продемонстрировало принципиальную возможность использования синтетически сгенерированных данных для обучения моделей бинарной классификации. Обе тестируемые архитектуры показали удовлетворительные результаты на тестовой выборке:

Модель логистической регрессии продемонстрировала точность 100%, что говорит об отсутствии ложноположительных срабатываний для данного класса. Однако полнота для этого же класса составил 0.67, что свидетельствует о пропуске части корректных ответов.

RBF-сеть показала вариативную точность в диапазоне 65—90% в зависимости от установленного порога классификации и параметров обучения, подтвердив повышенную чувствительность нелинейных моделей к гиперпараметрам.

Наиболее показательным этапом стал анализ работы моделей на новых, не входивших в обучающую выборку текстах, включая короткие и намеренно искаженные формулировки. Были выявлены систематические ошибки, указывающие на фундаментальные ограничения подхода, основанного на поверхностных текстовых признаках.

1. Лексическая хрупкость: Классификаторы демонстрировали чрезмерную зависимость от отдельных ключевых слов, игнорируя общий смысл высказывания.
   - Пример: Фраза «улучшили методы диагностики» была отнесена к неправильным ответам, в то время как «улучшение диагностики и осведомленности о РАС» классифицировалась верно. Модель, обученная на данных, где доминировала лемма «улучш» (от «улучшение»), не распознала однокоренной глагол.
2. Смысловая нечувствительность: Модели оказались неспособны распознавать семантические инверсии и логические противоречия.
   - Пример: Ответ «меньше осведомлённости о РАС» был ошибочно отнесен к категории правильных, тогда как «больше осведомлённости о ФАС» (Фетальный Алкогольный Синдром) — к неправильным. Это указывает на опору на изолированные лексемы «осведомлен» и «рас» без анализа их семантической связки.
3. Контекстуальная слепота: Модели некорректно интерпретировали утверждения, содержащие как правильные, так и неправильные ключевые слова.
   - Пример RBF-сети: научно обоснованное утверждение «Увеличение связано с улучшением диагностики и осведомленности о РАС» было классифицировано как неправильное, в то время как псевдонаучное утверждение «Рост числа диагнозов вызван вакцинацией детей» — как правильное. Это подчеркивает, что нелинейная модель может улавливать неинтерпретируемые паттерны, характерные для синтетических данных, но не для реальной логики.

Проведенный эксперимент выявил проблему переобучения на артефактах генерации. Модели запоминали не смысловые паттерны, а специфические формулировки и сочетания лемм, характерные для синтетических данных, что снижало их обобщающую способность на реальные, вариативные формулировки. Текущий подход демонстрирует потенциал для использования в качестве инструмента предварительной

сортировки и фильтрации явно некорректных ответов в условиях массового тестирования, однако для достижения экспертного уровня точности требуются более совершенные методы, такие как тонкая настройка (fine-tuning) предобученных языковых моделей (BERT и аналогичных) на репрезентативных наборах реальных данных (Нежников, Марьенков, 2024).

Несмотря на выявленные ограничения, исследование подтвердило потенциальную применимость гибридного подхода для задач предварительной классификации ответов и автоматического выявления явно некорректных формулировок. В условиях массового тестирования даже текущая точность моделей может обеспечить экономию временных ресурсов экспертов.

Перспективными направлениями дальнейших исследований видятся:

- Дообучение предварительно тренированных языковых моделей (tuning);
- Разработка ансамблей, комбинирующих несколько классификаторов;
- Расширение подходов к аугментации данных и генерации более разнообразных текстовых вариаций (включая тексты различной длины — от нескольких слов до нескольких предложений).

Эксперимент подтвердил принципиальную возможность использования синтетических данных, сгенерированных большими языковыми моделями, для обучения классификаторов текстовых ответов. Однако достигнутый уровень точности и выявленные ограничения указывают на то, что в настоящее время подобные системы могут эффективно использоваться лишь как вспомогательный инструмент в процессе экспертной оценки, а не как полноценная замена человека-эксперта.

Заключение

Несмотря на выявленные систематические ошибки, практическая ценность подхода не должна недооцениваться. В сценариях массового предварительного отбора, где требуется отфильтровать заведомо неверные или нерелевантные ответы, даже модель с точностью 70—80% может обеспечить значительную экономию временных ресурсов. Эксперт в таком случае проверяет не 100% ответов, а лишь ту часть, которую модель пометила как «правильные» или вызвавшие сомнение, что может сократить общий объем работы на 30—50%. Таким образом, система функционирует не как автономный судья, а как «интеллектуальный фильтр» или ассистент эксперта, повышающий эффективность его труда.

Дальнейшее совершенствование методов генерации данных и архитектур моделей открывает перспективы для создания более надежных систем автоматизированной оценки, способных работать в условиях ограниченной доступности размеченных данных.

Ограничения. Проведенное исследование имеет ряд ограничений, которые необходимо учитывать при интерпретации его результатов. Основным ограничением является относительно небольшой объем синтетической обучающей выборки (100 примеров), что могло повлиять на способность моделей к обобщению и повысить риск переобучения на артефактах генерации. Кроме того, использованный

подход к генерации данных, хотя и направленный на разнообразие, по своей природе ограничен теми паттернами и формулировками, которые способны воспроизвести выбранные языковые модели, и может не в полной мере отражать все многообразие реальных ответов респондентов. Наконец, валидация моделей проводилась на ограниченном наборе новых примеров; для более надежной оценки их производительности и устойчивости необходимы обширные тесты на репрезентативных выборках реальных данных из различных предметных областей.

Limitations. The conducted study has several limitations that should be considered when interpreting its results. The primary limitation is the relatively small size of the synthetic training sample (100 examples), which may have affected the models' ability to generalize and increased the risk of overfitting to generation artifacts. Furthermore, the data generation approach used, although aimed at diversity, is inherently limited by the patterns and formulations that the chosen language models can reproduce and may not fully capture the entire variety of real respondents' answers. Finally, model validation was performed on a limited set of new examples; extensive testing on representative samples of real-world data from various subject domains is required for a more robust assessment of their performance and resilience.

Список источников / References

1. Воронин, В.М., Курицин, С.В., Наседкина, З.А., Ицкович, М.М. (2017). Использование латентного семантического анализа как альтернативы пропозиционального анализа в исследованиях понимания текста. *Гуманизация образования. 2017. № 2*. (с. 11—19) https://www.elibrary.ru/item.asp?id=29369554
   Voronin, V.M., Kuritsin, S.V., Nasedkina, Z.A., Itstovich, M.M. (2017). Using a latent semantic analysis as alternatives of sentential analysis in studies of text understanding. *Humanization of education, 2017(2),* 11—19 (In Russ.) https://www.elibrary.ru/item.asp?id=29369554
2. Нежников, Р.И., Марьенков, А.Н. (2024). Сравнительный анализ моделей трансформера для классификации неструктурированной текстовой информации. *Прикаспийский журнал: управление и высокие технологии. 2024. № 2 (66)* (с. 32—38). https://www.elibrary.ru/item.asp?id=71199707
   Nezhnikov, R.I., Marenkov, A.N. (2024). Comparative Analysis of Transformer Models for Classification of Unstructured Text Information. *Caspian Journal: Control and High Technologies, 2024, 2 (66)*, 32—38. (In Russ.). https://www.elibrary.ru/item.asp?id=71199707
3. Ребенок, К.В. (2024) Эффективность нейросетевых алгоритмов в автоматическом реферировании и суммаризации текста. *Вестник НГУ. Серия: Информационные технологии. 2024. № 1*. (с. 49—61) https://doi.org/10.25205/1818-7900-2024-22-4-49-61
   Rebenok, K.V. (2024). Efficiency of Neural Network Algorithms in Automatic Abstracting and Summarization Text. *Vestnik NSU. Series: Information Technologies*. 2024;22(4):49—61. (In Russ.) https://doi.org/10.25205/1818-7900-2024-22-4-49-61

Информация об авторах
Борислав Юрьевич Поляков, младший научный сотрудник, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-6457-9520, e-mail: deslion@yandex.ru

Information about the authors
Borislav Yu. Polyakov, Junior Research Assistant, Moscow State University of Psychology and Education, Moscow, Russian Federation, ORCID: https://orcid.org/0000-0002-6457-9520, e-mail: deslion@yandex.ru

Вклад авторов
Поляков Б.Ю. — аннотирование, написание и оформление рукописи; планирование исследования; контроль за проведением исследования; проведение эксперимента.

Contribution of the authors
Borislav Yu. Polyakov — annotations, writing and design of the manuscript; planning of the research; control over the research; conducting the experiment.

Конфликт интересов
Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest
The authors declare no conflict of interest.

Декларация об этике
Исследование выполнено с использованием синтетических данных, сгенерированных языковыми моделями. Поскольку в исследовании не участвовали люди, получение одобрения этического комитета не требовалось.

Ethics statement
The study was conducted using synthetic data generated by language models. As the research did not involve human participants, ethics committee approval was not required.

Поступила в редакцию 24.10.2025 — Received 2025.10.24
Поступила после рецензирования 11.11.2025 — Revised 2025.11.11
Принята к публикации 13.11.2025 — Accepted 2025.11.13
Опубликована 28.12.2025 — Published 2025.12.28