Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

УДК 004.021:303.1

ЮЧКОВ
Андрей Константинович

АЛГОРИТМ ОЦЕНКИ РЕЗУЛЬТАТОВ СОЦИОЛОГИЧЕСКИХ ОПРОСОВ ПО НЕПОЛНЫМ ДАННЫМ

АВТОРЕФЕРАТ
диссертации на соискание степени магистра технических наук
по специальности 7-06-0612-03 Системы управления информацией

Научный руководитель
Севернёв Александр Михайлович
кандидат технических наук, доцент

Минск 2025

ВВЕДЕНИЕ

Социологические опросы и анкетирования применяются в самых разных областях – от изучения общественного мнения и электоральных предпочтений до анализа поведения потребителей и оценки эффективности социальных программ. Их результаты формируют основу для принятия решений во многих сферах: в здравоохранении – например, для выявления факторов низкого темпа вакцинации или диспансеризации; в образовании – для определения причин низкой успеваемости и выделения эффективных методик обучения; в городском строительстве – для оценки качества городской среды и её доступности; в маркетинге – для понимания потребительских предпочтений и адаптации коммерческих стратегий. Ценность таких исследований заключается в способности получать информацию непосредственно от участников, выявлять скрытые тенденции и строить прогнозы социально значимых явлений.

Однако наличие пропущенных данных в результатах опросов может существенно снижать точность оценок, искажать выводы о социальных трендах и, как следствие, вести к принятию неверных управленческих решений. Проблема отсутствующих значений впервые привлекла внимание статистиков ещё в начале *XX* века, когда Р.А. Фишер и Дж. Нейман заложили базовые принципы работы с неполными наблюдениями, предложив простейшие стратегии – удаление неполных записей или подстановку средних значений. Такие методы эффективны лишь при крайне редких пропусках (не более 5 %) и однородной выборке.

В 1970-1980-х годах были разработаны более гибкие подходы: алгоритм *EM* и метод *MICE*, позволяющие моделировать распределение пропущенных значений и учитывать неопределённость через создание нескольких полных версий данных. С развитием вычислительных мощностей широкое распространение получили методы машинного обучения, такие как алгоритм *k*-ближайших соседей, случайные леса и метод *SVM*. Эти алгоритмы способны выявлять сложные нелинейные взаимосвязи между признаками и часто превосходят традиционные регрессионные модели по качеству восстановления данных.

С 2010-х годов в задачах импутации стали активно использоваться глубокие нейронные сети – автоэнкодеры и сети *GAN*, которые демонстрируют высокую эффективность при работе с высокоразмерными и сильно зашумлёнными данными, хотя и требуют значительных вычислительных ресурсов.

Параллельно с этим в области обработки естественного языка появились большие языковые модели (*LLM*) – архитектуры, обученные на терабайтах текстовых данных и обладающие способностью генерировать связные тексты и

рассуждения. Таким образом, при формулировании задачи импутации как «текстовой» задачи – например, при представлении строк таблицы как предложений – *LLM* могут предлагать достаточно точные оценки для пропущенных полей. Это открывает перспективу применения универсального, «текстового» подхода к импутации, не требующего специализированной настройки под каждый конкретный набор данных.

Целью настоящей работы является разработка концептуальной основы применения *LLM*-агентов для восстановления пропущенных значений в социологических данных и определение их преимуществ относительно классических и *ML*-методов.

Для достижения поставленной цели в данной работе решаются следующие задачи:

1 Определить специфику работы с данными в сфере социологических исследований.

2 Оценить сильные и слабые стороны классических подходов в контексте анализа данных социологических опросов.

3 Изучить распространенные модели *LLM*, оценить качество их работы в рассматриваемой сфере и определить наиболее оптимальную модель.

4 Сформулировать базовые принципы применения *LLM* для импутации пропущенных значений, а также изучить и протестировать возможные стратегии взаимодействия с *LLM*.

5 Определить критерии и метрики для оценки качества восстановления данных и влияния на последующие аналитические модели.

Актуальность исследования определяется необходимостью повышения надёжности и воспроизводимости научных исследований в области социальных наук. Проблема неполных данных остаётся одной из ключевых преград для качественного анализа, а появление *LLM*-подходов открывает принципиально новые пути её решения. Данная работа позволит оценить перспективность данного направления, а в случае его перспективности также может стать основой дальнейших теоретических и прикладных исследований в данной области.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью диссертационной работы является разработка алгоритма оценки результатов социологических опросов по неполным данным.

Для достижения заданной цели необходимо решить следующие задачи:

– проанализировать современное состояние и тенденции развития алгоритмов импутации данных в различных сферах деятельности, исследовать ограничения и преимущества существующих методов;

– произвести теоретическое обоснование необходимости разработки нового алгоритма непосредственно для заполнения пропущенных значений в социологических данных на основе текущего состояния данной сферы;

– разработать алгоритм, учитывающий ограничения существующих методов, и провести оценку его применимости и возможностей;

– разработать программное средство, демонстрирующее работу алгоритма.

Область исследования

Содержание диссертации соответствует образовательному стандарту углубленного высшего образования по специальности 7-06-0612-03 Системы управления информацией.

В данной диссертационной работе объектом исследования является процесс импутации пропущенных данных. Предметом исследования является алгоритм заполнения пропусков при помощи больших языковых моделей

Теоретическая и методологическая основа исследования. В основу диссертации легли работы зарубежных исследователей в следующих областях:

– моделирование и классификация типов пропусков в данных;

– разработка и применение статистических методов импутации;

– исследование алгоритмов машинного обучения, ансамблевых моделей, а также нейросетевых подходов в контексте заполнения пропусков в данных;

– применение больших языковых моделей и гибридных алгоритмов в задачах обработки табличных и социологических данных.

Информационная база исследования сформирована на основе технической литературы, открытой информации, сведений из электронных ресурсов, а также материалов научных конференций и семинаров.

Научная новизна работы

На текущий момент тема использования больших языковых моделей в контексте восстановления данных является широко исследуемой в связи с

постоянно увеличивающимися способностями и качеством работы данных моделей. Особенную важность такой подход получает при использовании небольших наборов данных, в которых может не хватать записей для полноценного обучения нейросетевых моделей и моделей машинного обучения.

Актуальность работы определяется высокой значимостью полноты и качества социологических данных при их дальнейшем использовании в прикладном анализе, моделировании общественных процессов и выработке решений на основе эмпирических данных.

Теоретическая значимость работы состоит в разработке алгоритма импутации пропущенных значений на основе больших языковых моделей, способных учитывать содержательные связи между признаками и адаптироваться к различным типам данных.

Практическая значимость работы заключается в создании прикладного программного средства, предназначенного для восстановления пропущенных значений в табличных данных социологических исследований с использованием большой языковой модели, позволяющего исследователям и аналитикам данных повысить качество их работы и выводов, производимых на основе восстановленных данных.

Личный вклад соискателя

Все результаты и положения, выносимые на защиту, получены автором лично. Научный руководитель принимал участие в постановке задач, определении возможных путей их решения, в предварительном анализе, обсуждении исследований и их результатов, проведенных автором лично.

Апробация результатов диссертации

Результаты, полученные в ходе выполнения диссертационной работы докладывались и обсуждались на: 60-й юбилейной научной конференции аспирантов, магистрантов и студентов (Минск: БГУИР, 2024) [1–А., 2–А.], международной конференции Информационные технологии и системы (Минск: БГУИР, 2024) [3–А.], 61-й научной конференции аспирантов, магистрантов и студентов (Минск: БГУИР, 2025) [4–А.] и опубликованы в виде тезисов в материалах к перечисленным выше конференциям. По результатам была получена благодарность на конференции 2024 года.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении определена область исследования, обоснована актуальность диссертационной работы, показана ее научная новизна и практическая значимость, кратко проанализировано общее состояние проблемы и основные пути ее решения.

Первая глава «Теоретические и методологические основы восстановления пропусков в социологических данных» посвящена анализу природы пустых значений и классических подходов к их заполнению. В ней даётся детальная классификация механизмов возникновения пропусков, описываются простейшие статистические методы, а также современные приёмы машинного обучения и нейросетевые архитектуры. Отдельный раздел посвящён возможностям больших языковых моделей для контекстно-осмысленной импутации, схемам дообучения и гибридным стратегиям. Также глава содержит обзор литературных источников по современным разработкам и сравнительный анализ рассмотренных методов.

Вторая глава «Разработка алгоритма восстановления пропусков с использованием *LLM*» описывает постановку задачи и требования к системе, формализует её через ввод матрицы данных и булевой маски пропусков, а также целевую функцию с учётом квадратичных и категориальных потерь. Далее приводится обоснование выбора модели *LLM*, стратегия формирования запросов и отбор контекстных примеров, подробная схема предобработки и маскирования данных, логика генерации запросов и последовательность взаимодействия с *API*, а также механизм пост-валидации и логика запасного механизма.

Третья глава «Реализация алгоритма и разработка консольного приложения» содержит описание технической структуры ПО: использование *Python* для работы с исходными данными, компоненты маскирования, генерации и парсинга ответов, а также модуль резервной импутации. Представлены сведения о тестовых наборах данных и методиках оценки, а также сравнительный анализ полученных результатов с базовыми методами.

В заключении сформулированы основные результаты диссертационной работы. Цитирования обозначены ссылками на публикации, указанные в «Библиографическом списке».

ЗАКЛЮЧЕНИЕ

В ходе работы над магистерской работой был разработан алгоритм оценки результатов социологических опросов по неполным данным, отличающийся возможностью работы на небольших выборках с приемлемой точностью.

В ходе выполнения магистерской диссертационной работы были решены следующие задачи:

– проанализировано современное состояние и тенденции развития алгоритмов импутации данных в различных сферах деятельности, исследованы ограничения и преимущества существующих методов;

– проведено сравнительное тестирование различных *LLM* непосредственно для заполнения пропущенных значений в социологических данных;

– разработан алгоритм, учитывающий ограничения существующих методов, и проведена комплексная оценка его применимости и возможностей;

– разработано программное средство, демонстрирующее работу алгоритма.

В ходе сравнения полученного алгоритма с существующими аналогами было показано его преимущество на небольших выборках (100, 1000 записей) при небольшом проценте записей с пропущенными значениями.

Возможной областью применения данного алгоритма может стать обработка социологических данных в различных профильных компаниях и государственных органах, занимающихся сбором статистических данных.

Результаты, полученные в ходе выполнения диссертационной работы, докладывались и обсуждались на:

– 60-й и 61-й научной конференции аспирантов, магистрантов и студентов БГУИР (на русском и английском языках);

– XIV международной научной конференции «Информационные технологии и системы ИТС-2024».

СПИСОК ПУБЛИКАЦИЙ АВТОРА

[1–А.] Ючков, А. К. Заполнение пропусков в социологических данных с нелинейными зависимостями / А. К. Ючков // 60-я научная конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» : материалы конференции по направлению 2 : Информационные технологии и управление (Минск, 22–26 апреля 2024 года) / редкол. : Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2024. – С. 42.

[2–А.] Uchkov, A. K. Modern methods of analysing sociological datasets with missing values / A. K. Uchkov // Актуальные вопросы экономики и информационных технологий: материалы 60-й юбилейной научной конференции аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» (Минск, 22–26 апреля 2024 года) – Минск: БГУИР, 2024. – С. 762-764.

[3–А.] Ючков, А. К. Анализ качества оценки результатов социологических исследований по неполным данным / А. К. Ючков, К. А. Хаджинова, А. А. Навроцкий / Информационные технологии и системы 2024 (ИТС 2024): материалы междунар. науч. конф. (Минск, 20 ноября 2025 года). / редкол. : Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2024. – С. 189-190.

[4–А.] Uchkov, A. K. Potential biases and fairness in LLM-based data imputation for sociological research / A. K. Uchkov // Актуальные вопросы экономики и информационных технологий: материалы 61-й научной конференции аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» (Минск, 20–25 апреля 2025 года). – Минск : БГУИР, 2025. – С. 740-741.