



УТВЕРЖДАЮ
Директор ИПУ РАН
академик РАН
Д. А. Новиков
«23» октября 2025г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного учреждения науки
Институт проблем управления им. В. А. Трапезникова Российской академии наук
на диссертационную работу Ватолина Алексея Сергеевича
«Обучение и оценивание мультязычных нейросетевых моделей семантического
векторного представления научных текстов», представленную на соискание ученой
степени кандидата технических наук по специальности 2.3.8 Информатика и
информационные процессы.

Актуальность темы исследования обусловлена многократным увеличением количества публикаций за последние десятилетия в ведущих наукометрических базах данных (Scopus, Web of Science, eLibrary.ru). Это существенно усложняет для исследователей задачи поиска и анализа релевантных источников, вызывая необходимость в автоматических инструментах обработки научных текстов.

Современные нейросетевые модели показывают превосходство над традиционными статистическими методами в задачах анализа текстов, однако большая их часть обучена исключительно на англоязычных данных. Существующие мультязычные решения, поддерживающие русский язык, зачастую характеризуются большим количеством параметров и требуют значительных и дорогостоящих вычислительных ресурсов. Такие особенности ограничивают возможности практического внедрения решений для их широкого использования.

Таким образом, актуальной задачей является разработка компактных и эффективных моделей, востребованных в российской науке и ориентированных преимущественно на русский и английский языки. Работа А.С. Ватолина предлагает комплексный подход к решению указанных проблем, включая создание новых двуязычных моделей семантического представления научных текстов (SciRus) и открытых инструментов для их оценки (RuSciBench, RuSciFact). Реализация предложенного подхода позволяет снять обозначенные ограничения и значительно повысить эффективность научной деятельности, что подтверждает безусловную актуальность и своевременность диссертационной работы.

Структура и содержание диссертации. Диссертация состоит из введения, четырех глав, заключения и списка литературы.

Во введении обоснована актуальность избранной темы исследования, сформулированы цель и задачи работы, определены научная новизна и практическая значимость полученных результатов, а также представлены основные положения, которые автор выносит на защиту.

В первой главе проведён анализ теоретических основ и современных методов построения семантических векторных представлений текстовой информации. Рассмотрена эволюция методов от традиционных статистических подходов к современным нейросетевым моделям, основанным на архитектуре трансформер. Формализована постановка задачи векторизации текстов, описаны её ключевые этапы, включая токенизацию, предобучение языковых моделей и их дальнейшее дообучение для решения задач оценки семантической близости текстов.

Вторая глава посвящена вопросам разработки, обучения и оценки качества семейства двуязычных моделей SciRus, предназначенных для векторного представления научных текстов. Описаны наборы данных, используемые для обучения моделей, раскрыта архитектура разработанных решений, а также представлена двухэтапная методология их обучения, включающая маскированное языковое моделирование и контрастивное дообучение на парах «заголовок–аннотация» и «цитирующая–цитируемая статьи». Приведены результаты сравнительной оценки разработанных моделей с существующими аналогами по критериям качества и вычислительной эффективности, а также описан опыт практического внедрения данных моделей в информационно-поисковую систему eLibrary.ru.

Третья глава посвящена разработке инструментария для объективной оценки качества моделей векторизации русскоязычных научных текстов. Представлен разработанный автором мультизадачный двуязычный бенчмарк RuSciBench, основанный на корпусах данных российской научной электронной библиотеки и включающий различные задачи: классификацию, регрессию и информационный поиск. Проведено комплексное тестирование и анализ широкого спектра моделей, позволившее выявить их преимущества, ограничения и степень языковой специализации.

Четвёртая глава рассматривает актуальную проблему верификации научных фактов на русском языке. В главе изложена методика создания первого специализированного русскоязычного бенчмарка RuSciFact, основанная на полуавтоматической генерации утверждений с помощью больших языковых моделей с последующей экспертной проверкой. Задача декомпозируется на этапы информационного поиска релевантного контекста и последующей классификации, для каждого из которых представлены результаты экспериментальной оценки существующих моделей.

В заключении диссертации сформулированы основные результаты и выводы, сделанные по итогам проведённого исследования.

Основные результаты работы и их новизна заключаются в следующем.

1. Разработаны двуязычные модели SciRus-tiny и SciRus-small для векторного представления научных текстов, которые, обладая существенно меньшим количеством параметров, демонстрируют качество, сопоставимое с многоязычными аналогами, а на задачах русскоязычного анализа превосходят их.
2. Создан и апробирован RuSciBench — первый мультизадачный двуязычный (русско-английский) бенчмарк для оценки качества моделей векторного представления научных текстов, обеспечивающий стандартизированную процедуру тестирования и интегрированный в международный бенчмарк MTEB.
3. Предложена оригинальная полуавтоматическая методология формирования наборов данных для верификации научных фактов, включающая генерацию утверждений с помощью больших языковых моделей и их последующую экспертную проверку. На её основе сформирован RuSciFact — первый русскоязычный набор данных для оценки качества решений по векторизации текста в задачах фактологической проверки.

Достоверность полученных результатов подтверждается использованием апробированных методов машинного обучения, проведением серии численных экспериментов на обширных массивах реальных научных данных и сравнительным анализом разработанных моделей с ведущими аналогами на стандартных бенчмарках.

Объективность и воспроизводимость исследований обеспечиваются интеграцией предложенного бенчмарка RuSciBench в международный фреймворк MTEB, а также публикацией в открытом доступе исходного кода и весов разработанных моделей для бенчмарков RuSciBench и RuSciFact.

Основные положения диссертации получили апробацию на международных конференциях и отражены в пяти публикациях в рецензируемых научных изданиях, индексируемых в базах Scopus и Web of Science.

Значимость полученных результатов

Разработанные в диссертации компактные двуязычные модели SciRus являются значимыми для решения практических задач. Они показывают высокую эффективность при существенно меньших вычислительных затратах по сравнению с крупными многоязычными аналогами, что упрощает их внедрение в научно-информационные системы. Следует отметить успешное внедрение модели SciRus-tiny в поисковую систему электронной научной библиотеки eLibrary.ru. Это свидетельствует о высокой степени готовности предложенных автором решений к использованию в производственных условиях. Также особую ценность представляют созданные автором бенчмарки RuSciBench и RuSciFact, которые являются первыми стандартизированными инструментами для объективного сравнения и оценки качества моделей в исследуемой области.

Замечания к диссертации носят преимущественно рекомендательный характер:

1. В разделе 2.4, посвящённом обучению моделей SciRus, приводятся графики сходимости функции потерь и промежуточные результаты на бенчмарках SciDocs и RuSciBench. Целесообразно было бы дополнить материал более глубоким анализом динамики отдельных метрик качества в ходе обучения, что позволило бы убедительно обосновать выбор оптимального момента прекращения обучения.
2. В разделе 3.2.2 при описании задачи прогнозирования количества цитирований, недостаточно раскрыт вопрос о влиянии фактора времени на целевую переменную. Поскольку количество цитирований естественным образом увеличивается с возрастом публикаций, важно явно указать, как этот фактор учитывался при формировании выборок и при оценке качества предложенных моделей.
3. В разделе 3.3 в рамках сравнительного анализа моделей на бенчмарке RuSciBench было бы желательно подробнее раскрыть причины, по которым компактные модели SciRus демонстрируют показатели, сопоставимые с существенно более крупными моделями общего назначения. Более глубокая интерпретация ключевых факторов успеха (таких как особенности архитектуры моделей, специфика обучающих наборов данных и выбранная методология дообучения) значительно повысила бы научную ценность и практическую значимость представленных результатов.

Перечисленные замечания не снижают общей высокой оценки диссертационного исследования.

Заключительная оценка

Диссертация Ватолина Алексея Сергеевича на тему «Обучение и оценивание мультязычных нейросетевых моделей семантического векторного представления научных текстов» является завершённой научно-квалификационной работой, в которой решается актуальная научная задача, имеющая важное практическое значение.

Представленные в работе результаты и выводы корректно обоснованы, обладают научной новизной и достоверностью. Результаты работы своевременно и полно опубликованы в 5 печатных трудах, индексируемых в базах Scopus и Web of Science, а также докладывались автором на международных конференциях. Автореферат верно и полно отражает содержание диссертационной работы.

Диссертация соответствует всем критериям, установленным «Положением о присуждении ученых степеней», предъявляемым к диссертациям на соискание ученой степени кандидата наук, а ее автор — Ватолин Алексей Сергеевич — заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.8 — «Информатика и информационные процессы». Диссертационная работа, автореферат и отзыв на работу рассмотрены на заседании научного семинара Института проблем

управления им. В.А. Трапезникова Российской академии наук (протокол №1 от 12 сентября 2025 года).

Заведующий лабораторией №25,
доктор технических наук,
ведущий научный сотрудник

Д. А. Губанов

Сведения о составителе отзыва

Фамилия, имя, отчество: Губанов Дмитрий Алексеевич

Ученая степень: доктор технических наук

Наименование отрасли наук, научных специальностей, по которым защищена диссертация: 05.13.10 – Управление в социальных и экономических системах (технические науки)

Должность: Заведующий лабораторией № 25, ведущий научный сотрудник

Контактный телефон: +7 495 198-17-20, доб. 1402

Адрес электронной почты: gubanov@ipu.ru

Сведения о ведущей организации

Федеральное государственное бюджетное учреждение науки «Институт проблем управления им. В.А. Трапезникова Российской академии наук» (ИПУ РАН)

Сайт: www.ipu.ru

Почтовый адрес: 117997, ГСП-7, В-342, г. Москва, ул. Профсоюзная, д. 65

Телефон: +7 (495) 334-89-10

E-mail: dan@ipu.ru

