Компьютерные инструменты в образовании, 2025
№ 1: 61–75
УДК: 004.89
http://cte.eltech.ru
doi:10.32603/2071-2340-2025-1-61-75

ПРИМЕНЕНИЕ ТРАНСФОРМЕРОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В СИСТЕМЕ РЕКОМЕНДАЦИЙ НАУЧНЫХ РУКОВОДИТЕЛЕЙ

Будило Е. А.[1], студентка, ✉ elizaveta.budilo@gmail.com

[1]Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), ул. Профессора Попова, 5, корп. 3, 197022, Санкт-Петербург, Россия

Аннотация

В настоящей работе предложена система рекомендаций для выбора научного руководителя, основанная на архитектуре трансформеров и современных методах машинного обучения. Система анализирует академические данные студента, включая изученные дисциплины и оценки по ним, а также профессиональные характеристики преподавателей. Экспериментальные результаты демонстрируют значительное превосходство предложенного подхода над традиционными методами: при тестировании достигнута точность рекомендаций $0,3230$ против $0,1106$ (метод на основе частоты положительных оценок) и $0,1637$ (подход с использованием классификации через машинное обучение). Полученные результаты подтверждают эффективность системы в оптимизации процесса подбора научного руководителя, что способствует повышению качества научно-исследовательской деятельности студентов за счёт персонализированного сопоставления компетенций.

Ключевые слова: *система рекомендаций, машинное обучение, трансформеры, обработка естественного языка, научное руководство, анализ образовательных данных, текстовые эмбеддинги.*

Цитирование: Будило Е. А. Применение трансформеров и методов машинного обучения в системе рекомендаций научных руководителей // Компьютерные инструменты в образовании. 2025. № 1. С. 61–75. doi:10.32603/2071-2340-2025-1-61-75

1. ВВЕДЕНИЕ

Современные рекомендательные системы в образовательной сфере стремительно эволюционируют от традиционных алгоритмов к сложным интеллектуальным системам, способным учитывать данные пользователей. Классические подходы, такие как коллаборативная фильтрация [1] и контентно-ориентированные методы [2], несмотря на определённые успехи, сталкиваются с существенными ограничениями при решении академических задач — прежде всего, с проблемой «холодного старта» [3] и необходимостью обработки гетерогенных образовательных данных.

С развитием методов машинного обучения и появлением глубоких нейронных сетей открылись новые перспективы для создания более адаптивных и персонализированных

рекомендательных систем. Проведённые исследования в этой области продемонстрировали преимущества нейросетевых архитектур в задачах рекомендаций, обеспечивая значительное повышение точности по сравнению с классическими методами [4, 5].

В настоящей работе рассматривается задача разработки рекомендательной системы для выбора научных руководителей студентами на основе комплексного анализа различных данных о студентах и преподавателях. Цель работы — разработать подход, позволяющий рекомендовать наиболее подходящего научного руководителя на основе информации о научной деятельности преподавателей и с учётом предпочтений студента, выраженных через его «любимые» дисциплины.

Предлагается использовать методы машинного обучения и трансформерные модели [6] для формирования эмбеддингов как студентов, так и научных руководителей, что позволяет учитывать текстовые описания учебных программ, тем выпускных квалификационных работ (далее — ВКР), научных публикаций и других связанных данных. Задача сводится к классификации студентов по научным руководителям с последующим сравнением предложенных подходов с классическими методами рекомендаций.

2. ОБЗОР ЛИТЕРАТУРЫ

В последние годы в научной литературе по образовательным рекомендательным системам особое внимание уделяется интеграции современных методов машинного обучения и больших языковых моделей с традиционными алгоритмами. Гибридные подходы позволяют повысить точность и персонализацию рекомендаций, а также преодолеть такие ограничения классических систем, как проблема «холодного старта» и недостаточная интеграция разнородных образовательных данных. Это особенно актуально для задач подбора научных руководителей, где требуется учёт индивидуальных интересов студентов и профессиональных компетенций преподавателей.

2.1. Внедрение больших языковых моделей в рекомендательные системы

Значительный прогресс в области рекомендательных систем связан с внедрением больших языковых моделей (LLM). В существующем исследовании [7] проведён систематический анализ современных решений, демонстрирующий, что интеграция LLM позволяет существенно расширить функциональные возможности рекомендательных систем за счёт глубокого понимания естественного языка, генерации контекста и способности к рассуждению. В отличие от традиционных подходов, основанных на раздельной генерации эмбеддингов и последующем ранжировании, LLM реализуют унифицированный подход, формируя рекомендации непосредственно на основе текстовых и поведенческих данных пользователя. В рамках обзора выделяются ключевые стратегии интеграции LLM: предварительное обучение на обширных корпусах, дообучение (fine-tuning) на специализированных задачах и применение prompting для динамического управления генерацией рекомендаций [8].

Особое внимание уделяется адаптации моделей к специфике предметной области, а также возможностям LLM по интерпретации сложных текстовых описаний объектов и профилей пользователей. Авторы отмечают, что использование LLM способствует не только повышению точности рекомендаций, но и улучшает объяснимость системы за счет генерации естественно-языковых обоснований, что может быть полезно для образовательных задач. Вместе с тем, отмечаются вызовы, связанные с вычислительной эффек-

тивностью и необходимостью адаптации моделей к изменяющимся пользовательским предпочтениям и контексту. В качестве решений предлагаются оптимизация архитектур и гибридные схемы интеграции LLM с классическими компонентами рекомендательных систем.

2.2. Практические решения для рекомендаций научных руководителей

В рамках задачи подбора научных руководителей особый интерес представляют решения, сочетающие современные методы обработки естественного языка и машинного обучения. В исследовании [9] предложен комплексный подход к построению рекомендательной системы научных руководителей, объединяющий методы веб-скрейпинга, продвинутой обработки естественного языка и машинного обучения, включая использование Sentence-BERT для формирования семантических эмбеддингов и тематическую кластеризацию посредством K-means. Такой подход позволяет проводить глубокий анализ профилей преподавателей и эффективно сопоставлять их с интересами студентов.

Гибридный сбор данных, объединяющий структурированные источники (университетские сайты) и неструктурированные (Google Scholar), расширяет информационную базу и способствует повышению качества рекомендаций. Персонализация достигается за счёт сопоставления векторных представлений интересов студентов и профилей преподавателей с использованием косинусного сходства. Практическая реализация решения и его тестирование на реальных пользователях подтверждают применимость подхода. Вместе с тем были выявлены проблемы с качеством исходных данных и точностью кластеризации, что указывает на необходимость дальнейших исследований и доработки системы.

2.3. Применение трансформеров в рекомендательных системах

Наиболее перспективным направлением в последние годы стало применение архитектур трансформеров в задачах рекомендаций. Систематический обзор литературы, проведённый в [10], демонстрирует значительный прирост качества рекомендаций при использовании моделей на основе механизма внимания (self-attention).

Трансформеры, такие как BERT, GPT и их модификации, позволяют эффективно обрабатывать текстовые описания интересов и достижений, учитывать контекстные зависимости и интегрировать временные и тематические аспекты в единую модель.

Проведённые исследования [11] и [12] продемонстрировали применение трансформеров в задачах последовательных рекомендаций, где модели SASRec и BERT4Rec показали увеличение метрик точности (Recall, NDCG) на 10–20 % по сравнению с классическими методами. Эти исследования подтверждают потенциал трансформеров для моделирования сложных взаимосвязей между пользователями и объектами рекомендаций.

2.4. Вывод

Анализ существующих решений свидетельствует о том, что классические методы и даже современные практические решения, основанные на статических эмбеддингах и кластеризации, не в полной мере решают задачи персонализации и учёта сложных взаимосвязей в образовательных данных. Большие языковые модели и трансформеры обладают уникальной способностью интегрировать разнородные источники информации, учитывать семантические, временные и поведенческие аспекты, а также обеспечивать высокий уровень точности и объяснимости рекомендаций.

В связи с этим применение трансформеров и современных методов машинного обучения для формирования эмбеддингов студентов и научных руководителей представляется наиболее обоснованным и перспективным направлением для построения рекомендательных систем в образовательной среде. Такой подход позволяет глубоко анализировать образовательные траектории и научные интересы, интегрировать различные типы данных и обеспечивать масштабируемость, адаптивность и прозрачность рекомендаций, что подтверждается как теоретическими исследованиями, так и экспериментальными результатами.

3. МЕТОДОЛОГИЯ

3.1. Описание данных

В настоящей работе используются данные, предоставленные университетом СПбГЭТУ «ЛЭТИ». Среди наборов данных, применяемых в исследовании, можно отметить:

— список утверждённых тем ВКР;
— протоколы защит ВКР;
— темы аспирантских диссертаций;
— данные о студентах с параметрами обучения;
— рабочие программы дисциплин;
— оценки студентов по каждой дисциплине;
— научные публикации преподавателей;
— патенты и свидетельства о регистрации программ для ЭВМ;
— а также другие связанные наборы данных.

Основным набором данных является список утверждённых тем ВКР, содержащий информацию только за 2024 год. Данный набор состоит из 2337 строк. После удаления дубликатов и объединения с наборами данных о студентах и их оценках количество строк сократилось до 2258. В выборке представлены студенты различных образовательных программ: бакалавриата, специалитета и магистратуры. Всего 445 научных руководителей курировали ВКР.

Поскольку основной задачей настоящего исследования является классификация по научным руководителям, каждый из 445 преподавателей рассматривается как отдельный класс. 97 научных руководителей курируют по одной теме ВКР, а один руководитель — одновременно 44 темы. Среднее количество тем на преподавателя составляет 5,074. Нижний квартиль равен 2, медиана — 3, а верхний квартиль — 6.

Дополнительным набором данных служат протоколы защит ВКР. В них содержатся выставленные оценки, оригинальность, объём работы (количество страниц и слайдов), а также вопросы, заданные на защите. Аннотации к ВКР в данном наборе отсутствуют.

Среди наборов данных, описывающих студентов, имеются параметры обучения и оценки по дисциплинам. Параметры обучения включают номер группы, курс, учебный год, срок обучения, статус (бюджет, целевой приём, контракт), статус обучения (учащийся, отчислен, закончил, академический отпуск), гражданство, специальность, шифр специальности, уровень образования (бакалавриат, магистратура, специалитет, аспирантура), форму образования (очная, заочная, очно-заочная), кафедру и факультет.

Набор данных оценок студентов по каждой дисциплине включает следующие столбцы: первая попытка сдачи, итоговая оценка, семестр оценки, название дисциплины, форма образования (очная, заочная, очно-заочная), кафедра и тип ведомости (практика,

выпускная работа, контрольная работа, дифференцированный зачет, экзамен, курсовой проект, реферат, государственный экзамен), а также учебный год, в котором была выставлена оценка.

Кроме того, имеется набор данных с рабочими программами дисциплин. Рабочая программа — документ, описывающий дисциплину, включая список тем (разделов), которые студенты изучают, и вопросы к экзамену. Это описание может быть использовано для создания эмбеддингов (векторов-признаков) с помощью трансформера.

Среди наборов данных, описывающих научных руководителей, можно отметить темы аспирантских диссертаций, научные публикации, а также патенты и свидетельства о регистрации программ для ЭВМ. В наборе данных аспирантских диссертаций имеются только тема диссертации и связывающие ID с аспирантом и преподавателем.

Набор данных научных публикаций содержит следующие столбцы: название публикации, DOI (если имеется), ISBN (если имеется), наличие индексации WoS, RSCI, Scopus, РИНЦ и ВАК, ключевые слова (если имеются), аннотация (если имеется), research area (WoS), category (WoS), topic (Scopus), topic cluster (Scopus). Набор данных состоит из 10100 строк, из которых 6911 имеют аннотацию.

Набор данных патентов и свидетельств о регистрации программ для ЭВМ включает название, описание и область техники. Набор данных содержит 687 строк, из которых 163 имеют описание. Область техники включает информационные системы и технологии, сенсорику и измерительные технологии, электронику и приборостроение, биомедицинскую инженерию, электротехнику и электротехнологии, робототехнику, автоматику и управление, радиотехнику и телекоммуникации.

3.2. Подходы к решению

В рамках настоящего исследования первоначально рассматривалась возможность использования оценки за ВКР в качестве критерия для формирования рекомендаций научных руководителей. Согласно предложенной методике, студентам, получившим оценку «отлично», рекомендовался бы соответствующий преподаватель, в то время как при оценке ниже «отлично» рекомендация данного руководителя не предоставлялась бы.

Однако данный подход обладает рядом существенных ограничений:

- Общее количество доступных записей сокращается до 2114, что может негативно сказаться на статистической значимости и обобщаемости результатов.
- Между средней оценкой студента по всем изученным дисциплинам и оценкой за ВКР есть умеренная корреляция ($r = 0,5$), что указывает на косвенное влияние оценок по дисциплинам на итоговую оценку ВКР и, соответственно, на рекомендации.
- Анализ распределения оценок за ВКР по 34 кафедрам выявил значительную неоднородность: две кафедры выставляют исключительно оценку «отлично», четыре кафедры — в 90 % случаев, а 14 кафедр — в 75 % случаев. Такая вариативность снижает объективность и универсальность использования оценки ВКР в качестве единственного критерия для рекомендаций.

При попытках реализовать этот подход модели адаптировались к оценкам студентов и кафедр, что приводило к рекомендации большого количества преподавателей, не связанных со специальностью студента, и в результате снижалась релевантность и качество рекомендаций. В связи с выявленными недостатками использование оценки за защиту ВКР в качестве основного критерия для рекомендаций было признано неэффективным, и от него пришлось отказаться. Вместо этого в настоящем исследовании предлагается иной подход: если студент выбирает научного руководителя, то данный руководитель

будет рекомендоваться студентам с похожими академическими профилями. Таким образом, задача формирования рекомендаций сводится к решению задачи классификации.

Из доступных данных о студентах имеются сведения об изучаемых ими дисциплинах и полученных по ним оценках. В связи с этим в работе вводится концепция «любимых» дисциплин, под которыми понимаются учебные курсы, вызывающие наибольший интерес у студента и являющиеся потенциальной основой для выбора темы ВКР или начала научно-исследовательской деятельности.

Система рекомендаций, основанная на анализе «любимых» дисциплин и полного перечня изученных студентом дисциплин, направлена на подбор наиболее релевантных научных руководителей для каждого конкретного студента. В рамках исследования рассматриваются два основных подхода для реализации данной задачи:

— формирование эмбеддингов (векторных представлений) студентов с последующей классификацией с использованием методов машинного обучения;
— совместное построение эмбеддингов студентов и преподавателей с решением задачи мультирегрессии, на основе которой осуществляется выбор наиболее близкого по признакам научного руководителя.

Далее проводится сравнительное тестирование предложенных методов с классическими подходами, включающими:

- Детерминированный алгоритм рекомендаций, основанный на выборе преподавателя с кафедры обучения студента, учитывающий академические оценки учащегося у данного преподавателя.
- Классификацию по оценкам студентов методами деревьев решений [13] и градиентного бустинга CatBoost [14].

Предложенные в работе подходы предусматривают применение архитектур трансформеров для генерации векторных представлений студентов и преподавателей. Для валидации эффективности методов выполнена сравнительная оценка качества рекомендаций, на основании которой осуществляется выбор оптимальной архитектуры трансформера для анализируемого датасета.

3.3. Оценка трансформера на датасете

Для оценки качества эмбеддингов, получаемых с помощью трансформеров, были сформулированы следующие предположения:

- Темы, курируемые одним научным руководителем, должны быть семантически схожими, что отражается в высокой косинусной близости соответствующих векторных представлений.
- Темы, относящиеся к разным научным руководителям, должны быть семантически различными, то есть иметь низкую косинусную близость.

На основе этих предположений был сформирован датасет, включающий все возможные пары тем, принадлежащих одному научному руководителю (положительные примеры), и равное количество пар, состоящих из темы данного руководителя и случайно выбранной темы другого руководителя (отрицательные примеры).

Каждой паре была присвоена бинарная метка: 1 — если темы считаются схожими, 0 — если темы различны. В итоге был сформирован набор данных, состоящий из 21932 строк.

Далее для каждой темы были вычислены эмбеддинги с использованием выбранного трансформера, после чего для каждой пары вычислялась косинусная близость между соответствующими векторами. На основании полученных значений и бинарных меток бы-

ла проведена оценка качества модели с использованием площади под ROC-кривой (AUC-ROC) [15].

В сравнении исследовались следующие трансформеры: rubert-tiny2 [16], paraphrase-multilingual-mpnet-base-v2 [17], sbert-large-nru-ru [18] и multilingual-e5-large [19]. Все исследуемые трансформеры основаны на архитектуре BERT [20].

Аналогичные действия были проделаны с предположением, что темы одной кафедры являются похожими (набор данных имеет 264962 строк), а также с предположением, что темы одного факультета являются похожими (набор данных имеет 1006004 строк). Результат сравнения приведён в таблице 1.

Таблица 1. Сравнение качества моделей по ROC AUC

Модель (архитектура)	**Научный руководитель** ($N = 21932$)	**Кафедра** ($N = 264962$)	**Факультет** ($N = 1006004$)
RuBERT-Tiny (RU)	0,7902	0,7570	0,6678
MPNet-Base (Multilingual)	0,8224	0,7948	0,7161
SBERT-Large (RU)	0,7544	0,7280	0,6509
E5-Large (Multilingual)	0,8442	0,8053	0,7092

Из данных таблицы видно, что мультиязычные модели демонстрируют более высокие показатели по сравнению с русскоязычными аналогами. Модель multilingual-e5-large показала наилучшие результаты при оценке семантической близости тем на уровнях научных руководителей и кафедр, однако в задаче с факультетами она немного уступила paraphrase-multilingual-mpnet-base-v2. В связи с этим в рамках настоящего исследования была выбрана модель multilingual-e5-large как наиболее эффективная для дальнейшего использования.

3.4. Формирование векторного представления студента

В рамках первого подхода, предложенного в настоящем исследовании, формируется векторное эмбеддинг-представление студента с использованием трансформерной модели, после чего выполняется обучение модели машинного обучения на основе данных о выборе научных руководителей для ВКР студентов 2024 года.

Для формирования обучающего и тестового наборов данных было принято предположение, что «любимыми» дисциплинами студента являются те, по которым студент получил оценку выше медианы оценок по данной дисциплине в текущем учебном году на соответствующей кафедре. Иными словами, для каждой дисциплины рассчитывается медиана оценок студентов, обучающихся в одной или нескольких группах на одной кафедре. Если оценка конкретного студента превышает эту медиану, дисциплина считается «любимой» для данного студента.

Данный подход ориентирован на анализ одного учебного года и одной кафедры, что позволяет учитывать вариативность преподавательского состава и обеспечивает равные условия для студентов. Использование медианы вместо среднего арифметического предотвращает искажения, возникающие в случаях, когда оценки распределены неравномерно. Например, если большинство студентов получили «отлично», а одному студенту поставлена оценка «хорошо», средняя оценка группы будет занижена, что приведёт к неверной классификации дисциплин как «любимых» для большинства студентов.

Для генерации эмбеддингов дисциплин используется их рабочая программа. На основе текстового описания рабочей программы с помощью трансформера multilingual-e5-large вычисляются соответствующие эмбеддинги. Далее для каждого студента формируется взвешенный средний вектор эмбеддингов, включающий как «любимые» дисциплины, так и все изучаемые дисциплины, в соотношении 3 : 1. Иными словами, эмбеддинги «любимых» дисциплин учитываются с весом 4, а эмбеддинги остальных изучаемых дисциплин — с весом 1 (при этом «любимые» дисциплины также входят в общий набор изученных). Таким образом, итоговое векторное представление студента отражает как предпочтительные, так и изучаемые дисциплины, с акцентом на первые.

В рамках данного подхода не учитываются оценки, полученные студентами на первых двух курсах бакалавриата и специалитета, поскольку дисциплины в этот период преимущественно являются общеобразовательными и не отражают профильные интересы студентов.

Все эмбеддинги дисциплин вычисляются заранее, позволяя исключить необходимость вызова трансформерной модели во время процесса рекомендаций, что существенно снижает вычислительные затраты и повышает эффективность системы.

На основе сформированного вектора эмбеддингов студента строится классификационная модель, использующая алгоритмы дерева решений и градиентного бустинга из библиотеки CatBoost [14] для предсказания научного руководителя. Обучение модели проводится на парных данных, сформированных из набора данных по темам ВКР, где каждая пара состоит из представления студента и соответствующего научного руководителя, под руководством которого студент выполнял ВКР. Для оценки качества модель обучается и тестируется на выборках, разделённых в соотношении 4 : 1 (обучающая и тестовая выборки соответственно).

3.5. Формирование векторного представления научного руководителя

Во втором подходе, предложенном в настоящем исследовании, наряду с вектором эмбеддингов студента формируется также вектор эмбеддингов научного руководителя с использованием трансформерной модели.

Для построения векторного представления научного руководителя используются разнообразные источники информации: дисциплины, которые он ведёт; темы ВКР, которые он курировал; научные публикации; патенты и свидетельства о регистрации программ для ЭВМ, а также диссертации, по которым осуществляется научное руководство аспирантами. Для каждого из этих наборов данных формируется отдельный вектор эмбеддингов, после чего вычисляется взвешенное среднее, представляющее итоговый вектор научного руководителя.

При формировании вектора дисциплин, аналогично процессу создания вектора эмбеддингов студента, используются текстовые описания рабочих программ дисциплин. Поскольку дисциплины играют ключевую роль в выборе научного руководителя студентами и содержат достаточно объемные описания, в итоговый вектор научного руководителя вектор дисциплин включается с весом 3.

Темы ВКР также имеют важное значение для выбора научного руководителя. Однако в предоставленных данных отсутствуют аннотации к темам, что снижает точность получаемых эмбеддингов. В связи с этим вектор тем ВКР включается в итоговый вектор с весом 2.

Научные публикации, как правило, сопровождаются аннотациями. При их наличии для создания эмбеддингов используются как заголовки публикаций, так и аннотации,

что повышает качество представления. Вектор научных публикаций включается в итоговый вектор с весом 2.

Патенты, свидетельства о регистрации программ для ЭВМ и диссертации, по которым осуществляется научное руководство аспирантами, в указанном наборе данных представлены только темами и названиями без подробных описаний. В связи с этим соответствующие векторы включаются в итоговый вектор с весом 1.

Итоговое векторное представление научного руководителя создаётся посредством объединения эмбеддингов из различных информационных источников с учётом их значимости посредством весовых коэффициентов. Такой подход позволяет сформировать более точное отражение научных интересов и профессионального опыта руководителя, что является ключевым фактором для повышения качества и релевантности рекомендаций в системе подбора научных руководителей.

3.6. Метод решения

В настоящей работе предлагается метод решения задачи рекомендации научного руководителя с использованием нейронной сети, принимающей на вход вектор эмбеддингов студента и генерирующей на выходе вектор эмбеддингов рекомендованного научного руководителя. Для разработки и обучения модели применён фреймворк машинного обучения PyTorch [21].

Архитектура нейронной сети включает два скрытых полносвязных слоя, каждый из которых сопровождается нормализацией [24], функцией активации GELU [22] и слоем dropout [23] с вероятностью исключения 0.3. Обучение модели осуществляется на парных данных, сформированных из набора информации по темам ВКР, где каждая пара состоит из векторов студента и научного руководителя, под руководством которого студент выполнял ВКР. Данные разделены на обучающую и тестовую выборки в пропорции 4 : 1.

В качестве функции потерь используется CosineEmbeddingLoss [25], которая применяется для оценки сходства (или различия) между двумя векторами эмбеддингов в процессе обучения нейронной сети. Все метки при обучении принимаются равными 1, что означает, что все пары считаются схожими.

Диаграмма потока данных предложенного решения представлена на рисунке 1.

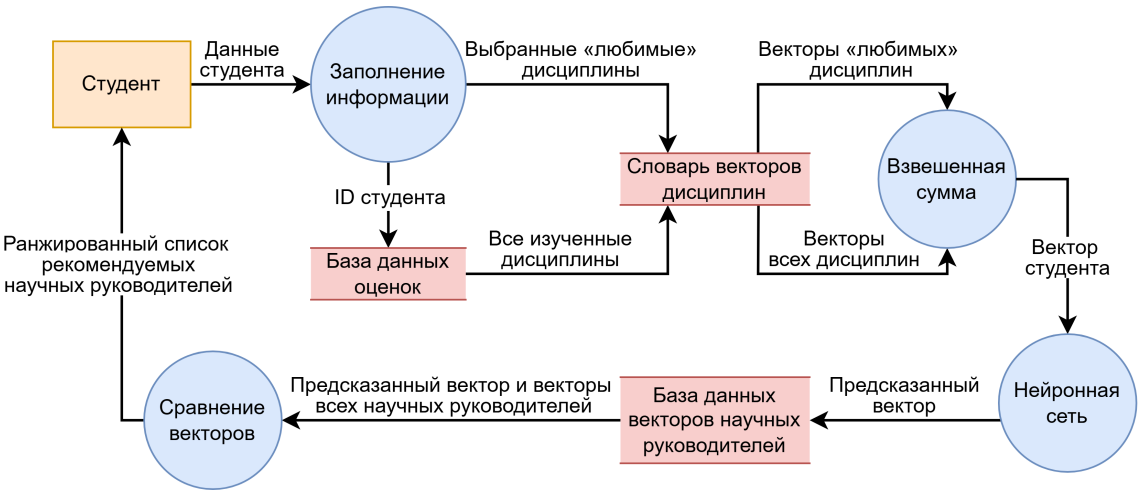

Рис. 1. Диаграмма потока данных в рекомендательной системе

Внедрение предлагаемой рекомендательной системы предполагает, что студент сможет самостоятельно определять наиболее интересные для себя дисциплины, которые будут использоваться для формирования его индивидуального векторного представления. В соответствии с этим, на этапе предсказания вектор эмбеддингов студента формируется на основе «любимых» дисциплин, выделенных либо непосредственно самим студентом, либо определённых на основе его оценок за период обучения. Из заранее вычисленных эмбеддингов дисциплин рассчитывается взвешенное среднее, которое служит входным вектором для нейронной сети. На выходе сеть генерирует вектор эмбеддингов предполагаемого научного руководителя. Для выбора конкретного руководителя выполняется поиск по косинусному расстоянию между предсказанным вектором и эмбеддингами всех доступных научных руководителей. В результате формируется ранжированный список научных руководителей по возрастанию косинусного расстояния.

Предложенный метод на основе нейронной сети и использования эмбеддингов позволяет эффективно моделировать сложные взаимосвязи между академическими интересами студентов и характеристиками научных руководителей.

3.7. Базовые подходы для сравнительного анализа

Для оценки эффективности предложенного метода использованы три альтернативных подхода: наивный метод, метод на основе дерева решений и метод на основе градиентного бустинга.

В контексте рассматриваемой задачи наивный подход предполагает, что студент выбирает научного руководителя, у которого он получил наибольшее количество положительных оценок. Для реализации данного метода, аналогично процессу формирования эмбеддингов, выделяются «любимые» дисциплины студента — дисциплины, по которым оценки превышают медиану оценок по соответствующей дисциплине за текущий учебный год. При этом оценки, полученные студентом на первом и втором курсах, не учитываются, что обусловлено преимущественно общеобразовательным характером дисциплин в этот период.

Выбор научного руководителя осуществляется среди преподавателей кафедры, на которой обучается студент. Каждому преподавателю присваиваются баллы в зависимости от его участия в «любимых» дисциплинах студента: за ведение лекций — 4 балла, за проведение практических занятий, лабораторных работ и других видов учебной деятельности — 2 балла. В случае, если студент получил оценку «отлично» по дисциплине, независимо от того, входит ли она в число «любимых», преподавателю дополнительно начисляется 1 балл за данную дисциплину.

На основании суммированных баллов формируется ранжированный список потенциальных научных руководителей по убыванию их общего рейтинга.

Для методов на основе дерева решений и градиентного бустинга был сформирован набор данных, представляющий оценки студентов по различным дисциплинам. Столбцы данного набора содержат оценки по всем изучаемым дисциплинам с указанием типа ведомости (практика, курсовая работа, дифференцированный зачет, экзамен), а строки соответствуют отдельным студентам. Значения в ячейках отражают оценки студентов. Данные были разделены на обучающую и тестовую выборки в пропорции $4:1$.

На обучающей выборке были обучены классификаторы, построенные на основе методов дерева решений и градиентного бустинга (с использованием библиотеки CatBoost [14]), которые предсказывали наиболее подходящего научного руководителя для каждого студента.

4. РЕЗУЛЬТАТЫ

В качестве основной метрики для оценки и сравнения результатов использовался классический показатель точности (accuracy), который определяется как доля правильно рекомендованных научных руководителей среди всех пар студент–преподаватель. В подходах, формирующих ранжированный список преподавателей, корректной рекомендацией считается ситуация, когда первый в списке преподаватель совпадает с фактическим научным руководителем студента.

Во всех экспериментах обучающие и тестовые выборки оставались одинаковыми. В случае наивного подхода обучающая выборка не использовалась. Для реализации метода дерева решений применялась библиотека sklearn [26], а для градиентного бустинга — библиотека CatBoost [14]. Обе модели обучались с использованием стандартных значений гиперпараметров.

Для нейронной сети использовалась функция потерь CosineEmbeddingLoss [25], планировщик скорости обучения ReduceLROnPlateau [27] и оптимизатор AdamW [28]. Обучение проводилось в течение 800 эпох с размером пакета 16. На рисунке 2 представлены графики динамики потерь и точности в процессе обучения.

Результаты всех экспериментов представлены в таблице 2.

Анализ данных показывает, что подходы, основанные на использовании трансформеров, демонстрируют значительно лучшие показатели точности по сравнению с классическими методами (0,3230 против 0,1637). Кроме того, данные подходы обеспечивают большую вариативность рекомендаций, поскольку выбор «любимых» дисциплин позволяет учитывать индивидуальные предпочтения студентов.

Таблица 2. Сравнение рассмотренных подходов по точности

Подход	Раздел	Точность
Наивный подход	3.7	0,1106
Дерево решений по оценкам	3.7	0,0774
Градиентный бустинг по оценкам	3.7	0,1637
Дерево решений по эмбеддингам студента	3.4	0,25
Градиентный бустинг по эмбеддингам студента	3.4	0,3163
Нейронная сеть по эмбеддингам студента и преподавателя	3.6	**0,3230**

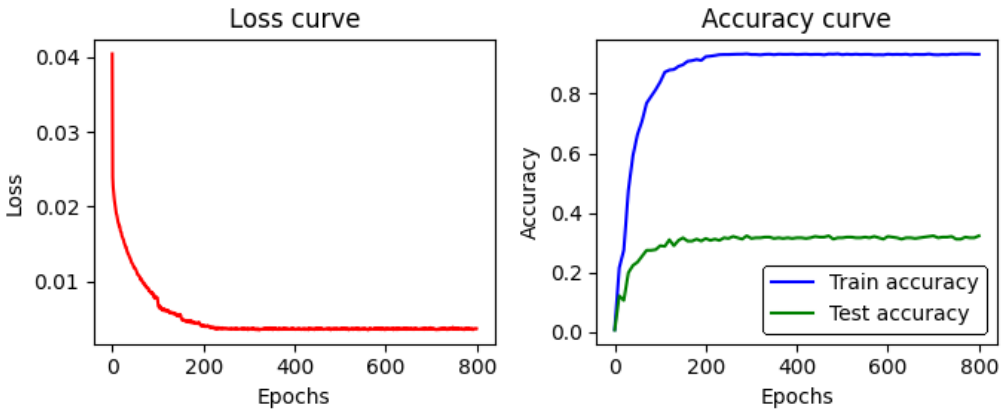

Рис. 2. Графики потерь и точности во время обучения

Подход, учитывающий эмбеддинги научных руководителей, показывает точность выше, чем метод, основанный только на эмбеддингах студентов с применением алгоритма CatBoost (0,3230 против 0,3163). Различия в точности невелики, что связано с более высокой сложностью модели CatBoost по сравнению с предложенной нейронной сетью (размер модели 212 МБ против 32 МБ). Тем не менее, даже при использовании относительно простой архитектуры нейронной сети, подход с учётом эмбеддингов преподавателей демонстрирует наилучшие результаты.

5. ЗАКЛЮЧЕНИЕ

В настоящей работе представлена методология разработки системы рекомендаций научных руководителей с использованием трансформерных моделей. При построении системы учитывается информация о дисциплинах, изученных студентами, а также косвенные данные об их оценках. Помимо этого, используется информация о потенциальных научных руководителях университета, их научных интересах и темах ВКР, которые они курировали.

На основе предоставленных данных предложенный подход продемонстрировал превосходство над наивным методом, основанным исключительно на оценках студентов, с показателями точности 0,3230 против 0,1106 и 0,1637 соответственно. Следует отметить, что значение точности 0,3230 нельзя считать высоким, что обусловлено ограниченностью объёма исходных данных: обучающая выборка содержит всего 1806 записей, что является недостаточным для задачи классификации с 445 классами.

Представленный трансформерный подход обладает преимуществом, позволяющим не использовать трансформер непосредственно на этапе предсказания, что существенно снижает вычислительные затраты. Кроме того, использование эмбеддингов научных руководителей предоставляет возможность включения новых преподавателей без опыта научного руководства, что повышает гибкость и масштабируемость системы.

Перспективы дальнейших исследований:

- Проведение анализа на более объемном и разнообразном наборе данных, например с использованием данных за предыдущие годы или добавлением синтетических данных.
- Поиск и использование аннотаций к темам ВКР для повышения качества эмбеддингов.
- Применение более точных и современных трансформерных моделей для формирования эмбеддингов.
- Дообучение трансформеров на специализированных научных текстах и данных.
- Разработка архитектуры нейронной сети, оптимизированной для задачи рекомендации научных руководителей.

Список литературы

1. *Herlocker J. L. et al.* Evaluating collaborative filtering recommender systems // ACM Transactions on Information Systems. 2004. Vol. 22, № 1. P. 5–53. doi: 10.1145/963770.963772
2. *Pazzani M. J., Billsus D.* Content-Based Recommendation System // The Adaptive Web. Berlin: Springer Berlin Heidelberg, 2007. P. 325–341. doi:10.1007/978-3-540-72079-9_10
3. *Lam X. N. et al.* Addressing cold-start problem in recommendation systems // Proc. of the 2nd Int. Conf. on Ubiquitous Information Management and Communication. 2008. P. 208–211. doi:10.1145/1352793.1352837

4. *He X. et al.* Neural Collaborative Filtering // Proc. of the 26th International Conference on World Wide Web. 2017. P. 173–182. doi:10.1145/3038912.3052569
5. *Kang W. C., McAuley J.* Self-Attentive Sequential Recommendation // Proc. of the 2018 IEEE Int. Conf. on Data Mining (ICDM). 2018. P. 197–206. doi:10.1109/icdm.2018.00035
6. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 31. P. 1–15.
7. *Zhao Z. et al.* Recommender Systems in the Era of Large Language Models (LLMs) // IEEE Transactions on Knowledge and Data Engineering. 2024. Vol. 36, № 11. P. 6889–6907. doi:10.1109/tkde.2024.3392335
8. *Zhao W. X. et al.* A survey of large language models // aarXiv:2303.18223v16 [cs.CL]. 2023.
9. *Veerannagari V. R.* Developing A Professor Recommendation System. Ames, Io, USA: Iowa State University, 2024.
10. *Pohan H. I. et al.* Recommender System Using Transformer Model: A Systematic Literature Review // Proc. of 2022 1st International Conference on Information System & Information Technology (ICISIT). 2022. P. 376–381. doi: 10.1109/icisit54091.2022.9873070
11. *Sun F. et al.* BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer // Proc. of the 28th ACM international conference on information and knowledge management. 2019. P. 1441–1450. doi:10.1145/3357384.3357895
12. *Kang W. C., McAuley J.* Self-Attentive Sequential Recommendation // Proc. of the 2018 IEEE International Conference on Data Mining (ICDM). 2018. P. 197–206. doi:10.1109/icdm.2018.00035
13. *Breiman L., Friedman J., Olshen R., Stone C.* Classification and Regression Trees. Belmont, CA, USA: Wadsworth Int. 1984.
14. *Prokhorenkova L. et al.* CatBoost: unbiased boosting with categorical features // Advances in neural information processing systems. 2018. Vol. 31. P. 1–11.
15. *Bradley A. P.* The use of the area under the ROC curve in the evaluation of machine learning algorithms // Pattern recognition. 1997. Vol. 30, № 7. P. 1145–1159. doi:10.1016/s0031-3203(96)00142-2
16. *Solovyev V. et al.* A BERT-Based Classification Model: The Case of Russian Fairy Tales // Journal of Language and Education. 2024. Vol. 10, № 4. P 98–111. doi:10.17323/jle.2024.24030
17. *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 3982–3992. doi:10.18653/v1/d19-1410
18. *Snegirev A. et al.* The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design // Proc. of the 2025 Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025. Vol. 1. P. 236–254. doi:10.18653/v1/2025.naacl-long.12
19. *Wang L. et al.* Multilingual e5 text embeddings: A technical report // arXiv:2402.05672v1 [cs.CL]. 2024.
20. *Devlin J. et al.* Bert: Pre-training of deep bidirectional transformers for language understanding // Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186. doi:10.18653/v1/N19-1423
21. *Paszke A.* Pytorch: An imperative style, high-performance deep learning library // arXiv:1912.01703v1 [cs.LG]. 2019.
22. *Hendrycks D., Gimpel K.* Gaussian error linear units (gelus) // arXiv:1606.08415v5 [cs.LG]. 2016.
23. *Hinton G. E. et al.* Improving neural networks by preventing co-adaptation of feature detectors // arXiv:1207.0580v1 [cs.NE]. 2012.
24. *Ba J. L., Kiros J. R., Hinton G. E.* Layer normalization // arXiv:1607.06450v1 [stat.ML]. 2016.
25. *Barz B., Denzler J.* Deep Learning on Small Datasets without Pre-Training using Cosine Loss // Proc. of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020. doi:10.1109/wacv45572.2020.9093286
26. *Pedregosa F. et al.* Scikit-learn: Machine learning in Python // Journal of machine Learning research. 2011. Vol. 12. P. 2825–2830.
27. *Zeiler M. D.* Adadelta: an adaptive learning rate method // arXiv:1212.5701v1 [cs.LG]. 2012.
28. *Loshchilov I., Hutter F.* Decoupled weight decay regularization // arXiv:1711.05101v3 [cs.LG]. 2017.

Поступила в редакцию 15.03.2025, окончательный вариант — 10.04.2025.

Будило Елизавета Александровна, студентка 4-го курса бакалавриата, кафедра информационных систем СПбГЭТУ «ЛЭТИ», ✉ **elizaveta.budilo@gmail.com**

Computer tools in education, 2025
№ 1: 61–75
http://cte.eltech.ru
doi:10.32603/2071-2340-2025-1-61-75

Application of Transformers and Machine Learning Methods in the System of Recommendations of Academic Supervisor

Budilo E. A.[1], Student, ✉ elizaveta.budilo@gmail.com

[1]Saint Petersburg Electrotechnical University, 5, building 3, Professora Popova st., 197022, Saint Petersburg, Russia

Abstract

This paper proposes a recommendation system for academic supervisor selection based on transformer architecture and modern machine learning methods. The system analyzes a student's academic data, including courses studied and grades for them, as well as the professional characteristics of teachers. The results of the experiment demonstrate a significant superiority of the proposed approach over traditional methods: during testing, the achieved accuracy of recommendations was 0.3230 versus 0.1106 (method based on the frequency of positive grades) and 0.1637 (approach using classification through machine learning). The obtained results confirm the effectiveness of the system in optimizing the process of selecting a supervisor, which helps to improve the quality of students' research activities through a personalized comparison of competencies.

Keywords: *recommendation system, machine learning, transformers, natural language processing, academic supervision, educational data analysis, text embeddings.*

Citation: E. A. Budilo, "Application of Transformers and Machine Learning Methods in the System of Recommendations of Academic Supervisor," *Computer tools in education*, no. 1, pp. 61–75, 2025 (in Russian); doi:10.32603/2071-2340-2025-1-61-75

References

1. J. L. Herlocker et al., "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, 2004, doi: 10.1145/963770.963772
2. M. J. Pazzani and D. Billsus, "Content-Based Recommendation Systems," in *The Adaptive Web*, Berlin: Springer Berlin Heidelberg, pp. 325–341, 2007; doi:10.1007/978-3-540-72079-9_10
3. X. N. Lam et al., "Addressing cold-start problem in recommendation systems," in *Proc. of the 2nd Int. Conf. on Ubiquitous Information Management and Communication*, pp. 208–211, 2008, doi: 10.1145/1352793.1352837
4. X. He et al., "Neural Collaborative Filtering," in *Proc. of the 26th International Conference on World Wide Web*, pp. 173–182, 2017; doi:10.1145/3038912.3052569
5. W.-C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," in *Proc. of the 2018 IEEE Int. Conf. on Data Mining (ICDM)*, pp. 197–206, 2018, doi:10.1109/icdm.2018.00035
6. A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 31, pp. 1–15, 2017.
7. Z. Zhao et al., "Recommender Systems in the Era of Large Language Models (LLMs)," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 6889–6907, 2024; doi:10.1109/tkde.2024.3392335
8. W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223v16 [cs.CL]*.

9. V. R. Veerannagari, *Developing A Professor Recommendation System*, Ames, Io, USA: Iowa State University, 2024.
10. H. I. Pohan et al., "Recommender System Using Transformer Model: A Systematic Literature Review," in *Proc. of 2022 1st International Conference on Information System & Information Technology (ICISIT)*, pp. 376–381, 2022, doi: 10.1109/icisit54091.2022.9873070
11. Sun F. et al, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proc. of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019; doi:10.1145/3357384.3357895.
12. W.-C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," in *Proc. of the 2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018; doi:10.1109/icdm.2018.00035
13. L. Breiman et al., *Classification and Regression Trees*, Belmont, CA, USA: Wadsworth Int. 1984.
14. L. Prokhorenkova et al., "CatBoost: unbiased boosting with categorical features," in *Advances in neural information processing systems*, vol. 31, pp. 1–11, 2018.
15. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997; doi:10.1016/s0031-3203(96)00142-2
16. V. Solovyev et al., "A BERT-Based Classification Model: The Case of Russian Fairy Tales," *Journal of Language and Education*, vol. 10, no. 4, pp. 98–111, 2024; doi:10.17323/jle.2024.24030
17. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019; doi:10.18653/v1/d19-1410
18. A. Snegirev et al., "The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design," in *Proc. of the 2025 Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 236–254, 2025; doi:10.18653/v1/2025.naacl-long.12
19. Wang L. et al., "Multilingual e5 text embeddings: A technical report," 2024, *arXiv:2402.05672v1 [cs.CL]*.
20. J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019; doi:10.18653/v1/N19-1423
21. A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703v1 [cs.LG]*.
22. D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus),"2016, *arXiv:1606.08415v5 [cs.LG]*.
23. Hinton G. E. et al., "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580v1 [cs.NE]*.
24. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450v1 [stat.ML]*.
25. B. Barz and J. Denzler, "Deep Learning on Small Datasets without Pre-Training using Cosine Loss," in *Proc. of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020; doi:10.1109/wacv45572.2020.9093286
26. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
27. M. D. Zeiler, "Adadelta: an adaptive learning rate method," 2012, *arXiv:1212.5701v1 [cs.LG]*.
28. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101v3 [cs.LG]*.

Received 15-03-2025, the final version — 10-04-2025.

Elizaveta Budilo, 4rd year Student of the Bachelor's Degree Program, Information Systems Department, SPbETU «LETI», ✉ elizaveta.budilo@gmail.com