Использование технологий искусственного интеллекта для повышения качества данных и аналитики научной деятельности университета

В. П. Семенов

Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова

4245110@mail.ru

Р. В. Соколов[1], И. Л. Андреевский[2], А. Е. Бобчинский[3], И. В. Подгорных[4]

Санкт-Петербургский государственный экономический университет

[1] rsok7@rambler.ru, [2] ail@unecon.ru, [3] artembobchinskiy@yandex.ru, [4] ira.podgornykh@mail.ru

Аннотация. **В статье формулируются проблемы обеспечения качества исходных данных по публикационной активности как основы аналитики научной деятельности университета. Выявляются направления прикладного использования технологий искусственного интеллекта (ИИ) для повышения качества результатов тематического анализа публикационной активности, предлагается концептуальный алгоритм предобработки исходных данных, решаются задачи применения предобученных мультиязычных трансформерных моделей семантического кодирования текста и повторного ранжирования в целях сопоставления публикаций с наиболее распространенными системами индексирования и рубрикации научных публикаций.**

Ключевые слова: публикационная активность, классификация научных публикаций, предобученные мультиязычные трансформерные модели, обработка естественного языка, технологии искусственного интеллекта

I. Введение

В век цифровых технологий стремительно меняются подходы в обучении и проведении научных исследований. Изменяется роль крупных университетов.

Современный университет является не только образовательной площадкой, но и выступает как мощный центр генерации новых знаний, центр научных исследований, в котором качественная подготовка кадров неразрывно связана с фундаментальными исследованиями и их верификацией в авторитетных публикациях.

Важно выстраивать систему объективного мониторинга качества научной продукции и результативности научной деятельности через оценку публикационной активности. Такой подход позволяет трансформировать количественные значения в достоверные показатели академической результативности и интеллектуального авторитета. Также это является способом продвижения организации и публикаций отдельных сотрудников, делая бренд университета узнаваемым, а выпускников востребованными на рынке труда. Дополнительным результатом является возможность подготовки регламентных отчетов в штатном режиме, где заполнение отчетных форм выполняется в штатном режиме и не отнимает значительного количества времени.

Растет и ширится применение ИИ в разных сферах. Не обошел данный тренд и исследуемую область. Использование ИИ делает возможным публикационный анализ делать более интеллектуально. В настоящий момент публикаций по данной тематике недостаточно много, а данная деятельность ведется не системно. В этом контексте тематика данной статьи актуальна, а реализуемый исследовательский проект носит прикладной и практический характер.

Конечной целью проводимого исследования является повышение качества исходных данных для анализа публикационной активности, результатов анализа в целях упрощения формирования необходимой статистической отчетности и принятия более обоснованных управленческих решений.

II. Текущее состояние и описание проблем

Имеющиеся на рынке прикладные информационные системы (модуль «Наука» в программном продукте «1С:Университет», модуль «Управление НИОКР» в системе «Галактика Управление вузом» на базе Галактика ERP, модуль «НИР и инновации» системы ТАНДЕМ.Университет, решение Парус-Предприятие (для вузов) и аналогичные решения) занимаются больше решением учетных задач, сколько публикаций было сделано, кем, за какой период, сколько научно-исследовательских работ и грантов проведено организацией за отчетный период и получено финансирования в разрезе их источников и т.п. Однако в большинстве случаев дело не доходит до качественного анализа с дальнейшей визуализацией результатов на дашбордах.

Для проводимого коллективом авторов анализа публикационной активности университета берутся данные из нескольких источников: выгрузка информации по всем публикациям организации в формате xml, выгрузка списка сотрудников и подразделение профиля организации, выгрузки по зарубежным публикациям (WoS, Scopus), выгрузка данных по действующим сотрудникам организации, справочники библиографических кодов ББК, УДК, кодов ГРНТИ и пр., данные библиографических кодов их личных профилей сотрудников и др.

Имеющийся массив первичных данных для целей анализа публикационной активности требуется специализированным образом подготовить, устранить

неточности, дополнить недостающей информацией, исключить противоречия. То есть с технической стороны можно говорить про задачи повышения качества данных.

Технологии искусственного интеллекта (ИИ) показали себя с наилучшей стороны в задачах рутинной обработки данных.

В выгрузке данных по публикациям присутствуют все аффилированные с организацией авторы (не только действующие сотрудники). Для анализа необходимо выбрать только публикации текущих сотрудников. Так как выгрузка выполняется на регулярной основе, то необходимо учитывать срезы на определенные даты. Не у всех публикаций сопоставлены библиографические коды (ББК, УДК, РФФИ, РКП, др.) тематике публикации, могут быть проставлены чаще всего 1–2 кода.

Привязка публикации к библиотечным кодам носит зачастую формальных характер, ограничиваясь указанием укрупненной категории или сферы деятельности или проставляется примерно, по аналогии с прошлыми публикациями автора. Это характерно для кода ГРНТИ. В 10–20% публикаций от общего объема выгрузки данные кода ГРНТИ пропущены. Данные других кодов могут быть занесены все вместе в одно поле, могут содержать различные знаки разделителей, опечатки, лишние символы и пр. Ключевые слова публикации зачастую дублируют слова из заголовка публикации. В отдельных случаях ключевые слова не попадают в выгрузку. Не у всех есть авторов аффилиации с организацией в профилях библиографических сервисов.

Проведение качественного тематического анализа публикационной активности таким образом затруднительно. Повышение качества имеющихся данных становится трудоемкой рутинной операцией, требующей значительных временных усилий.

В то же время проведение детального количественного и качественного тематического анализа публикаций необходимо для решения ряда прикладных задач:

- проверка соответствия тематик имеющихся публикаций профилю организации;
- подготовка регламентной отчетности по результативности научной деятельности организации за отчетный период с корректной разноской по тематике и категориям (точные, естественные, гуманитарные, общественные, технические (прикладные) науки);
- классификация и привязка публикации к коду научной специальности ВАК и формирование отчетов по диссертационным советам и его участникам;
- поиск и подбор сотрудников, специализирующихся на определенной тематике для рецензирования научных публикаций, участия в качестве официальных оппонентов, т.п.;
- оценка качества публикаций (мусорные и хищнические издания), работа на корзину, публикации с нулевой цитируемостью, с привязкой к отраслевой направленности и др.

Для решения перечисленных задач необходимо повышение качества исходных данных. Данные задачи предлагается решать с использованием технологий ИИ, хорошо зарекомендовавших себя в задачах рутинной обработки данных, организации датапайплайнов и т.п.

III. Использование технологий ИИ в задачах повышения качества исходных данных и их классификации

К числу наиболее актуальных задач, которые могут быть решены с использованием технологий ИИ, относятся:

- интеллектуальное дополнение недостающих данных кодам УДК, ББК, ГРНТИ или др. для повышения точности тематического анализа;
- детализация тематики публикации в рамках укрупненной классификации публикаций по направлениям (например, в отношении кодов ГРНТИ или др.);
- классификация и привязка публикации к коду научной специальности для членов диссертационных советов (когда известен код УБК, ББК, ГРНТИ, есть данные соответствия кодов паспортов специальностей ВАК и т.п.);
- оценка персональных достижений для соблюдения формальных требований к занимаемой должности и др.

Для достижения поставленных задач с технической точки зрения необходимо выполнить ряд практических шагов. Первичная выгрузка о публикационной активности университета в формате XML была заказана из личного кабинета, а потом извлечена из архива с сайта elibrary.ru [5], а с помощью специально разработанного авторами парсера в среде MS Visual Studio 2022 на языке C# с использованием технологии LINQ for XML разобрана и последовательно импортирована в базу данных mySQL 8.0. В эту же базу данных были добавлены данные о действующих сотрудниках, их библиографических идентификаторы и т.д. С помощью дополнительно разработанных скриптов на Python в среде Jupyter Notebook [2] проведена в базе данных нормализация данных о названиях публикаций, аннотациях, названиях источников публикаций для разных языков, ключевых библиографических кодах (например, первоначально названия публикации хранились в виде нескольких строк в таблице базы с указанием используемого языка, а после стали храниться в виде отдельных полей таблицы базы данных). С использованием запросов на SQL построена предварительная выгрузка данных в формат CSV для дальнейшего тематического анализа в среде облачной платформы Visiology.

IV. Постановка задачи и предлагаемый путь решения

Проведенный анализ полученной выгрузки данных о публикационной активности организации показал, что в таком виде многих данных не хватает или они требуют очистки. Первоочередной задачей стало дополнение существующей выборки метаинформацией (коды УДК, ББК, ГРНТИ). Так как задачи однотипны для каждого кода, то, научившись использовать технологии ИИ по автоматическому дополнению какого-то одного кода с использованием нейросетевых моделей на основе совокупности таких признаков, как название публикации, ключевые слова и аннотация, другие задачи

можно решить по аналогии. Для исследования были взяты описательные данные только для двух языков - русского и английского.

Так как коды ББК имеют свои особенные правила формирования [1], а проставленные в выборке коды ГРНТИ [4] не обладают большой вариативностью, то задача по работе с кодами УДК [3] показалась авторам более интересной с научной точки зрения. Кодов УДК в справочнике более 118 тысяч, а ГРНТИ немногим меньше 8 тысяч [4], при этом многие публикации в имеющейся выборке характеризуются кодом укрупненной, а не детализированной категории. В этом плане УДК более предпочтителен. Был предложен и реализован алгоритм на Python в среде Jupyter Notebook. Обучение нейросетевой модели велось в Google Colab.

V. Разработка нейросетевой модели

На предварительном этапе исследования был сформирован набор данных с публикациями для автоматизированного присвоения рубрик УДК. Исходный массив включал более 80 тыс. записей, однако для дальнейшей работы были сохранены только публикации, в которых был указан хотя бы один код УДК. После этого набор был дополнительно очищен и из большого числа служебных и нерелевантных полей были оставлены текстовые признаки публикации и целевой признак (наименование УДК). В итоговой базе, использованной в экспериментальной части, содержалось около 28 тыс. публикаций, для которых удалось одновременно получить и номер УДК, и его текстовое описание. Именно этот массив стал основой для всех последующих экспериментов.

Первоначально задача заключалась в необходимости предсказания подходящих кодов УДК по характеристикам публикации (название, ключевые слова и аннотация). Однако такой подход оказался методологически несостоятельным. В результате анализа было выявлено около 3430 уникальных кодов УДК. Анализ их частотного распределения показал высокую несбалансированность. В таких условиях модель, предсказывающая код как независимый класс, неизбежно сталкивается с проблемой недостаточной обученности на редких рубриках и не способна устойчиво обобщать для множества редких классов. Поэтому прямое предсказание номера УДК было признано недостаточно перспективным.

В связи с этим задача была переформулирована как семантически осведомленный поиск по текстам меток. Основанием для такого перехода стало различие между кодом и названием УДК: номер УДК является формальным индексом и сам по себе не несет языкового смысла, тогда как текстовое описание рубрики обладает содержательной семантикой и может быть сопоставлено с текстом публикации. Для реализации этой постановки потребовалось присоединить к публикациям названия УДК. Поскольку в исходной базе публикаций таких расшифровок не было, был создан отдельный справочник. Из внешнего источника [3] была создана таблица, содержащая около 118 тыс. кодов УДК и соответствующих им описаний. Проведена нормализация номеров УДК в публикациях, заключающаяся в удалении шумовых символов, унификации записи, а также разбиении множественных УДК внутри одной ячейки с использованием разделителя. Эта процедура не решила абсолютно все проблемы сопоставления, однако позволила корректно наполнить столбец с текстовыми описаниями УДК для 28 тыс. публикаций организации. Если публикации соответствовало несколько УДК, их описания также записывались последовательно через разделитель в другом поле таблицы.

После формирования набора данных с входными и выходными признаками был выполнен анализ структуры данных. Было показано, что все 3430 кодов УДК в используемой выборке соответствуют ровно одному названию, тогда как только 25 названий соответствуют двум или более кодам, то есть доля неоднозначных названий составила 0.73%. Это позволило использовать название УДК как основной целевой признак. Задача должна рассматриваться как многометочное ранжирование и присвоение (multi-label ranking and assignment), а не как обычная одноклассовая классификация, поскольку на одну публикацию может приходиться несколько УДК.

Далее были исследованы варианты формирования характеристик (входных признаков) публикации. Поскольку каждое содержательное поле (название, ключевые слова, аннотация) могло быть представлено на русском и английском языке, были проверены две языковые стратегии. Первая, когда приоритетно используется русский текст, а при его отсутствии – английский. Вторая, когда русская и английская версии объединяются в одно представление. Для этих стратегий были сформированы семь комбинаций входных признаков: только название, только ключевые слова, только аннотация, а также их основные сочетания. Таким образом, возникло 14 сценариев входного состояния характеристик публикации. Такой подход был необходим, поскольку заранее нельзя было определить, что эффективнее для поиска рубрики УДК: максимально полный текст или, напротив, более короткое, но тематически концентрированное описание.

На следующем этапе исследования были использованы базовые модели сравнения (baseline). Их применение имело две цели: оценить, насколько задача решается без тяжелых нейросетевых архитектур, а также отсечь заведомо проигрышные сценарии до запуска более ресурсоемких моделей. Набор данных был разбит на обучающую, валидационную и тестовую выборки в пропорции 70:15:15. При разбиении была использована групповая стратификация, чтобы сохранить покрытие классов и одновременно избежать утечки через текстовые дубликаты. По итогам этого этапа лучшим baseline подходом оказался «TF-IDF-представление на основе словарных и символьных n-грамм», то есть модель, использующая одновременно словарные и символьные признаки. Наиболее перспективными сценариями для дальнейшего этапа стали: выбор наилучшего доступного языка только для заголовка, ключевых слов и аннотации; объединение доступных языков для заголовка, ключевых слов и аннотации; выбор наилучшего доступного языка только для заголовка и ключевых слов; объединение доступных языков для заголовка и ключевых слов.

На этапе выбора наилучшего архитектурного решения была реализована схема, в которой отдельно кодируются названия, ключевые слова и аннотации публикаций, а также названия рубрик УДК, после чего их векторные представления сравниваются в общем семантическом пространстве. Были протестированы три

мультиязычные модели семантического кодирования текста: multilingual-e5-base, paraphrase-multilingual-mpnet-base-v2 и bge-m3. На валидации лучшей оказалась модель paraphrase-multilingual-mpnet-base-v2 в сценарии «выбор наилучшего доступного языка только для заголовка и ключевых слов». На валидационной выборке данная конфигурация обеспечила полноту (метрика Recall) среди первых 10 позиций ранжированного списка 0.198, среди первых 20 позиций – 0.287, среди первых 50 позиций – 0.414, средний обратный ранг (метрика MRR) в пределах первых 10 позиций – 0.097 и нормированную дисконтированную кумулятивную полезность (метрика nDCG) в пределах первых 10 позиций – 0.117. На тестовой выборке были получены близкие значения: полнота среди первых 10 позиций составила 0.198, среди первых 20 позиций – 0.275, среди первых 50 позиций – 0.409, средний обратный ранг в пределах первых 10 позиций – 0.095, а нормированная дисконтированная кумулятивная полезность в пределах первых 10 позиций – 0.116. Лучшим оказалось использование сочетания названия и ключевых слов публикации. Следовательно, для данной задачи решающим оказался не объем текста, а концентрация тематического сигнала. Одновременно гипотеза о выигрыше от мультиязычного слияния не получила убедительного подтверждения, что связано с ограниченным присутствием английского языка в наборе данных.

После выбора лучшей нейросетевой модели семантического поиска был реализован второй этап, повторное ранжирование кандидатов. На предыдущем шаге модель уже формировала для каждой публикации полный ранжированный список всех 3430 рубрик УДК по степени их семантической близости к тексту публикации. Далее из этого полного списка отбирались два укороченных набора кандидатов – первые 20 и первые 50 позиций, которые затем повторно оценивались и переупорядочивались более точной нейросетевой моделью bge-m3. Сопоставление двух режимов показало, что использование списка из первых 20 кандидатов обеспечивает более высокое качество верхней части ранжирования, чем режим с 50 кандидатами. Финальная система с повторным ранжированием на тестовой выборке обеспечила полноту среди первых 10 позиций 0.215, полноту среди первых 20 позиций 0.295, средний обратный ранг в пределах первых 10 позиций 0.106 и нормированную дисконтированную кумулятивную полезность в пределах первых 10 позиций 0.127. Тем самым она превзошла как лучший лексический baseline, так и лучшую модель семантического поиска без этапа повторного ранжирования.

После построения для каждой публикации ранжированного списка рубрик УДК потребовалось определить, какие из них следует считать не просто кандидатами, а рекомендованными нейросетевой моделью. Для этого на валидационной выборке экспериментально был подобран порог релевантности: если оценка рубрики моделью оказывалась выше этого порога, рубрика включалась в итоговый набор рекомендаций, а если ниже – не включалась. Данный порог использовался как граница между режимом «показать рубрику как возможный вариант» и режимом «считать рубрику присвоенной». Лучшим по результатам валидации оказался порог, равный 0.95.

В режиме ранжирования система работает сравнительно успешно: она умеет поднимать релевантные рубрики УДК в верхнюю часть списка кандидатов. Однако в более строгом режиме, когда требуется автоматически преобразовать этот список в окончательный набор присвоенных рубрик, качество заметно снижается. На тестовой выборке в режиме автоматического порогового отбора были получены следующие значения: микро-F1 = 0.058, F1 по образцам = 0.057 и точность подмножества = 0.049. Первая метрика характеризует общее качество присвоения рубрик по всем решениям системы, вторая – качество на уровне отдельных публикаций, а третья показывает долю публикаций, для которых система полностью и безошибочно восстановила весь набор УДК.

На практике это означает, что модель интерпретируется как инструмент поддержки эксперта, который формирует упорядоченный список наиболее релевантных рубрик УДК, а окончательное решение остается за человеком. В условиях большого числа редких и семантически близких рубрик УДК наиболее перспективен подход двухэтапной архитектуры, описанной выше.

Достоверность результатов подбора УДК была проверена в ручном режиме на независимых данных, отсутствующих в первоначальном наборе данных.

VI. Заключение

Таким образом, в результате исследования формулируются проблемы обеспечения качества исходных данных по публикационной активности как основы аналитики научной деятельности университета; выявляются направления прикладного использования технологий ИИ для повышения качества результатов тематического анализа публикационной активности; предлагается концептуальный алгоритм предобработки исходных данных на основе данных о публикационной активности; разработана нейросетевая модель предсказания кодов УДК для научных публикаций.

Полученные результаты могут способствовать повышению качества тематического анализа публикационной активности. Дальнейшее развитие исследования предполагает необходимость учета иерархической структуры УДК, а также дополнительную обработку неоднозначных и коротких названий рубрик типа «Теория» или «Общие вопросы» и т.п. с целью повышения точности предсказания за счет более сложной калибровки.

Список литературы

[1] Библиотечно-библиографическая классификация (ББК). Сокращенные таблицы [Электронный ресурс]. Режим доступа: http://roslavl.library67.ru/files/382/bbk.pdf

[2] Веб-платформа для интерактивных вычислений Jupyter Notebook [Электронный ресурс]. Режим https://jupyter.org

[3] ГОСТ Р 7.0.90-2016 Универсальная десятичная классификация. Структура, правила ведения и индексирования [Электронный ресурс]. Режим доступа: https://docs.cntd.ru/document/1200142869

[4] Государственный рубрикатор научно-технической информации (ГРНТИ) [Электронный ресурс]. Режим доступа: https://грнти.рф/

[5] Российский информационно-аналитический портал в области науки, технологии, медицины и образования, научная электронная библиотека [Электронный ресурс]. Режим доступа: https://elibrary.ru