MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 32, NO. 1, 2025
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Hierarchical Classification of Scientific Articles Using Deep Learning (Using the UDC Hierarchy as an Example)

V. Y. Mamedov[1], D. A. Kovalevsky[1], D. A. Morozov[1], S. S. Stolyarov[1], S. S. Ospichev[1]

DOI: 10.18255/1818-1015-2025-1-80-94

[1]Novosibirsk National Research State University, Novosibirsk, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received February 14, 2025
Revised February 24, 2025
Accepted February 26, 2025

The exponential growth in scientific publications has heightened the need for robust tools to organize and retrieve research effectively. The Universal Decimal Classification (UDC) serves as a valuable framework for categorizing articles by subject area. However, manual assignment of UDC codes is often prone to inaccuracies or oversimplification, limiting its utility. In this study, we present a novel approach for the automated assignment of UDC codes to scientific articles using BERT-based models. Our methodology was trained and evaluated on a dataset comprising over 19,000 articles in mathematics and related disciplines. To address the hierarchical structure of UDC, we developed two specialized evaluation metrics: hierarchical classification accuracy and hierarchical recommendation accuracy. We also explored multiple strategies for flattening hierarchical labels. Our results demonstrated a hierarchical recommendation accuracy of 0.8220. Furthermore, blind expert evaluation revealed that discrepancies between reference and predicted labels often stem from errors in the original UDC code assignments by article authors. Our approach demonstrates strong potential for automating the classification of scientific articles and can be extended to other hierarchical classification systems.

Keywords: text classification; hierarchical text classification; universal decimal classifier; deep learning

INFORMATION ABOUT THE AUTHORS

Mamedov, Valentin Y.	ORCID iD: 0009-0004-4154-5522. E-mail: v.mamedov@g.nsu.ru PhD student
Kovalevsky, Danil A.	ORCID iD: 0009-0002-8484-7366. E-mail: d.kovalevskii@g.nsu.ru Master student
Morozov, Dmitry A. (corresponding author)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com Junior Researcher
Stolyarov, Stepan S.	ORCID iD: 0009-0005-7651-6948. E-mail: s.stolyarov@g.nsu.ru Junior Researcher
Ospichev, Sergey S.	ORCID iD: 0000-0001-9912-6364. E-mail: s.ospichev@nsu.ru Deputy Director of the Mathematical Center in Akademgorodok, PhD

For citation: V. Y. Mamedov, D. A. Kovalevsky, D. A. Morozov, S. S. Stolyarov, and S. S. Ospichev, "Hierarchical classification of scientific articles using deep learning (using the UDC hierarchy as an example)", *Modeling and Analysis of Information Systems*, vol. 32, no. 1, pp. 80–94, 2025. DOI: 10.18255/1818-1015-2025-1-80-94.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 32, № 1, 2025
сайт журнала: www.mais-journal.ru

Иерархическая классификация научных статей при помощи глубокого обучения (на примере иерархии УДК)

В. Ю. Мамедов[1], Д. А. Ковалевский[1], Д. А. Морозов[1], С. С. Столяров[1], С. С. Оспичев[1]

DOI: 10.18255/1818-1015-2025-1-80-94

[1]Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 14 февраля 2025 г.
После доработки 24 февраля 2025 г.
Принята к публикации 26 февраля 2025 г.

В условиях стремительного роста числа научных публикаций актуальной задачей становится разработка эффективных инструментов для их систематизации и поиска. Одним из таких инструментов является универсальная десятичная классификация (УДК), которая позволяет структурировать статьи по тематическим областям. Однако ручное присвоение кодов УДК зачастую оказывается неточным или недостаточно детализированным, что снижает эффективность использования этого подхода. В данной статье предлагается подход к автоматическому присвоению кодов УДК научным статьям с использованием моделей на основе архитектуры BERT. Для обучения и оценки модели был использован набор данных, содержащий более 19 тысяч статей по математике и смежным наукам. Мы разработали две специализированные метрики качества, учитывающие иерархическую природу УДК: иерархическую классификационную точность и иерархическую рекомендательную точность. Кроме того, мы предложили несколько стратегий преобразования иерархических меток в плоские. В ходе экспериментов нам удалось достичь значения иерархической рекомендательной точности 0,8220. Дополнительно проведено слепое тестирование с участием экспертов, которое выявило, что часть расхождений между эталонными и сгенерированными метками обусловлена некорректным выбором кода УДК авторами статей. Предложенный подход демонстрирует высокий потенциал для автоматической классификации научных статей и может быть адаптирован для других иерархических систем классификации.

Ключевые слова: классификация текстов; иерархическая классификация текстов; универсальный десятичный классификатор; глубокое обучение

ИНФОРМАЦИЯ ОБ АВТОРАХ

Мамедов, Валентин Юрьевич	ORCID iD: 0009-0004-4154-5522. E-mail: v.mamedov@g.nsu.ru Аспирант
Ковалевский, Данил Анатольевич	ORCID iD: 0009-0002-8484-7366. E-mail: d.kovalevskii@g.nsu.ru Магистрант
Морозов, Дмитрий Алексеевич (автор для корреспонденции)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com Младший научный сотрудник
Столяров, Степан Сергеевич	ORCID iD: 0009-0005-7651-6948. E-mail: s.stolyarov@g.nsu.ru Младший научный сотрудник
Оспичев, Сергей Сергеевич	ORCID iD: 0000-0001-9912-6364. E-mail: s.ospichev@nsu.ru Заместитель директора Международного Математического Центра, канд. физ.-мат. наук

Для цитирования: V. Y. Mamedov, D. A. Kovalevsky, D. A. Morozov, S. S. Stolyarov, and S. S. Ospichev, “Hierarchical classification of scientific articles using deep learning (using the UDC hierarchy as an example)”, *Modeling and Analysis of Information Systems*, vol. 32, no. 1, pp. 80–94, 2025. DOI: 10.18255/1818-1015-2025-1-80-94.

Введение

Количество ежегодно публикуемых научных статей стабильно растёт. Так, в 2019 году количество документов, индексируемых Google Scholar превысило 389 млн [1], а по состоянию на февраль 2023 года в Scopus было проиндексировано более 90 млн публикаций[1]. Важной особенностью является ускорение этого роста: если за 1980 год было опубликовано в совокупности около одного миллиона статей, то к концу 2010-х число ежегодно публикуемых статей превысило семь миллионов [2].

Вместе со стремительным ростом числа работ всё более актуальными становятся инструменты для поиска и фильтрации публикаций. Большой популярностью пользуются сервисы семантического поиска статей (например, Semantic Scholar[2]). Подобные сервисы опираются на формирование семантических векторов статей и сопоставление их с запросом пользователей. Среди прочих инструментов следует упомянуть поиск по ключевым словам. Этот подход давно зарекомендовал себя в научной среде, однако по тем или иным причинам выбранные авторами ключевые слова могут отсутствовать в библиографической базе знаний. Преодолеть эту трудность помогают алгоритмы автоматической генерации ключевых слов. В последнее время лучших результатов в этой области удаётся добиться при помощи генеративных нейронных сетей, в том числе больших языковых моделей (LLM) [3, 4].

При этом невозможно утверждать, что какой-либо из этих инструментов является идеальным. Автоматические методы реферирования и формирования семантических векторов могут сталкиваться с проблемами из-за использования схожей лексики в различных не связанных областях науки. Ключевые слова обладают большой гибкостью в описании тематики, однако это же и усложняет поиск в случае, например, выбора синонимичных слов или фраз. Так, авторы могут указать ключевую фразу «нейронные сети», и такая статья не будет найдена при поиске по фразе «глубокое обучение». Унифицировать синонимичные понятия позволяет использование иерархических систем классификации тематики. Одной из таких систем является универсальная десятичная классификация.

Универсальная десятичная классификация (УДК) — это библиографическая система классификации, широко используемая по всему миру[3]. Она используется для систематизации документов, в том числе научных статей, из всех областей человеческого знания. Как и многие другие системы классификации, УДК имеет иерархическую структуру. Код УДК представляет собой последовательность десятичных цифр, разделённую для удобства чтения точками. Уточнение классификации происходит с помощью добавления дополнительных символов к последовательности справа. Таким образом, чем больше символов в УДК, тем точнее задана тематика статьи. Средняя глубина дерева классификации УДК равняется 6–9 уровням. Пример дерева представлен на рис. 1.

Общей проблемой ключевых слов и систем классификации, подобных УДК, на практике оказывается недостаточная точность определения авторами тематики своей статьи. Так, авторы могут указать слишком общие ключевые слова [5] или выбрать недостаточно глубокий код УДК. В иных случаях ключевые слова или код УДК могут быть не указаны вовсе, если этого не предполагает формат журнала или если издание использует другую систему классификации, например, Mathematics Subject Classification (MSC)[4] (причём однозначно транслировать коды между различными системами не всегда возможно). Решением этой проблемы может стать автоматическая генерация соответствующих значений. При этом, в последние годы было опубликовано множество алгоритмов автоматической генерации ключевых слов, в том числе, на материале русского языка [5, 6], тогда

[1] https://blog.scopus.com/posts/scopus-now-includes-90-million-content-records
[2] https://www.semanticscholar.org
[3] https://udcc.org/index.php
[4] https://zbmath.org/classification/

```
0 Общий отдел. Наука изнание~. Организация. Информационные технологии. Информация. Документация.
    Библиотечное дело. Учреждения. Публикации вцелом~
...
5 Математика иестественные~ науки
  50 Общие вопросы математических иестественных~ наук
  51 Математика
    510 Фундаментальные иобщие~ проблемы математики
    ...
    517 Анализ
      517.1 Введение ванализ~
      ...
      517.9 Дифференциальные, интегральные идругие~ функциональные уравнения. Конечные разности.
    Вариационное исчисление. Функциональный анализ
        517.91 Общая теория обыкновенных дифференциальных уравнений
        ...
        517.95 Дифференциальные уравнения вчастных~ производных
          517.951 Общая теория дифференциальных уравнений исистем~ счастными~ производными
          ...
          517.956 Линейные иквазилинейные~ уравнения исистемы~ уравнений
            517.956.1 Уравнения спостоянными~ коэффициентами
            517.956.2 Эллиптические уравнения исистемы~
              ...
              517.956.23 Эллиптические уравнения высокого порядка
                517.956.232 Общие свойства эллиптических уравнений высокого порядка
                517.956.233 Краевые задачи дляэллиптических~ уравнений высокого порядка
          ...
        ...
      ...
    ...
  ...
```

Fig. 1. Fragment of the UDC code tree

Рис. 1. Фрагмент дерева кодов УДК

как исследований, направленных на классификацию статей согласно УДК, нам удалось обнаружить гораздо меньше.

Цель настоящей работы состоит в создании автоматической системы присвоения научным статьям кодов согласно УДК. Для достижения этой цели нами был использован набор данных, содержащий информацию о более чем 19 тысячах статей по математике и смежным наукам. Мы рассмотрели ряд подходов, опирающихся на архитектуру BERT. Для того, чтобы учесть при классификации иерархическую природу классов УДК, мы предложили специальные метрики качества и опробовали несколько стратегий преобразования иерархических меток в плоские. Нам удалось достичь качества модели 0, 8220 в смысле иерархической рекомендательной точности (эта метрика подробно описана в разделе 3.5). Для большей репрезентативности исследования, помимо подсчёта автоматических метрик, мы привлекли к оценке качества модели экспертов. Слепое тестирование показало, что по крайней мере часть расхождений между эталонными и сгенерированными метками может быть объяснена ошибками в наборе данных.

Настоящая работа устроена следующим образом. В разделе 1 приводится краткий обзор предметной области. В разделе 2 описан собранный для нужд эксперимента набор данных. Стратегии преобразования меток и методология эксперимента описываются в разделе 3. Результаты экспериментов и их обсуждение приведены в разделе 4. Наконец, в разделе 5 подводятся итоги работы.

1. Обзор предметной области

Классификацию текстов следует считать одной из наиболее изученных областей обработки естественного языка. Долгое время в этой области лучшие результаты показывали методы классическо-

го машинного обучения, такие как метод опорных векторов, наивный байесовский классификатор и логистическая регрессия, а для векторизации текстов использовались такие простые алгоритмы как TF-IDF и n-граммы. Эти методы показывали хорошие результаты на небольших наборах данных, но их эффективность ограничивалась слабым учётом контекста и семантики текста в целом [7]. С развитием глубокого обучения удалось достичь принципиального повышения качества автоматической разметки [8]. В настоящее время прогресс в классификации текстов сосредоточен в основном вокруг подходов, опирающихся на архитектуру Transformer, в частности, BERT-подобных [9] и GPT-подобных [10] моделей.

Задача выбора класса УДК является частным случаем задачи классификации текстов: иерархической классификацией. В этом случае различные возможные классы не независимы, а упорядочены в виде подвешенного дерева, причём, чем дальше вершина находится от корня, тем более узкий специализированный класс она представляет. В таком дереве считается, что классы-потомки некоторой вершины являются подклассами класса, связанного с вершиной-предком. Автоматическая иерархическая классификация текстов естественным образом наиболее часто востребована при упорядочивании больших коллекций документов с неполной или отсутствующей метаинформацией. К исследованным в ходе предыдущих исследований коллекциям относятся, например, библиотеки научных документов [11] или выписки из медицинских карт [12]. Используемые разными авторами методы при этом значительно различаются. Так, в работе [13] в качестве алгоритмов векторизации рассматриваются GloVe, word2vec и FastText, а в качестве классификаторов — FastText, XGBoost, метод опорных векторов и свёрточная нейронная сеть. Лучшие результаты при этом продемонстрировал алгоритм FastText. В качестве корпуса для исследования использовалась коллекция RCV1, содержащая статьи издательства Reuters [14]. В работе [12] для классификации медицинских записей согласно кодам заболеваний по МКБ предлагается подход, основанный на обучении сети Multilabel-XLNet с архитектурой Transformer. При этом сравнение десятков актуальных подходов на материале нескольких наборов данных [15, 16] показывает, что лучшие из этих методов различаются по качеству разметки слабо (в пределах нескольких процентов).

Подходы, направленные на классификацию текстов согласно УДК, встречаются в то же время достаточно редко. В работе [17] рассматривается задача присвоения меток УДК текстам научных статей и газет. Рассмотренные методы достаточно просты: для векторизации текстов использованы TF-IDF и FastText, а для классификации — метод опорных векторов, многослойный перцептрон, логистическая регрессия и наивный байесовский классификатор. Авторы рассматривают классификацию с точностью до двух символов в коде УДК (всего 73 класса). Полученные значения F-меры (с макроусреднением) для разных комбинаций методов векторизации и классификации находятся в интервале от 0,42 до 0,85. Применительно к русскому языку можно упомянуть работы [18, 19]. В работе [18] для классификации использован многослойный перцептрон, а в качестве набора данных использованы статьи, размещённые на портале КиберЛенинка[5]. Авторы ограничились определением первого символа в коде УДК (всего 10 классов). Предложенный подход не позволил эффективно разделить классы 5 «Математика и естественные науки» и 6 «Прикладные науки. Медицина. Технологии». В работе [19] проводится оценка значимости различных автоматически обнаруживаемых именованных сущностей при классификации текстов согласно УДК. Авторы исследуют поддерево 517 «Анализ» на примере именованных сущностей, относящихся к методам, постановке проблемы и вычислениям. К сожалению, авторы последних двух работ не приводят численных оценок качества автоматической классификации. Таким образом, область расстановки меток УДК остаётся достаточно слабо изученной, а применяемые методы значительно отстают от аналогичных, используемых в других задачах иерархической классификации текстов.

[5] https://cyberleninka.ru/

2. Данные

Нам не удалось найти в открытом доступе набора данных, который содержал бы всю необходимую для нашего исследования информацию, а именно: название статьи, ключевые слова, аннотацию и значение УДК. В связи с этим мы решили подготовить такой набор данных самостоятельно. В качестве домена мы выбрали научные статьи по математике и смежным наукам, а в качестве источника данных — базу данных Math-Net.Ru[6]. Для того, чтобы гарантировать достаточную полноту и корректность данных, совместно с экспертами мы определили список из 21 периодического издания, из числа индексируемых Math-Net.Ru, обладающих достаточно корректной разметкой УДК и не содержащих большого числа пропусков в описании статей. Для выбранных изданий мы собрали все имеющиеся в сервисе статьи, описание которых было полным. В итоговый набор данных вошло 19233 статей, что составляет около 6 % от всех имеющихся в сервисе статей.

Средняя длина заголовка статей составила около 83 символов. В совокупности в заголовках встретилось 33337 уникальных словоформ со средней длиной слова равной 7,9 символа. Средняя длина аннотации составила около 592 символов. В совокупности в аннотациях нашлось 139627 уникальных словоформ, причём их средняя длина оказалась чуть меньше, чем в случае заголовков: 7,3 символа. Наконец, среди ключевых фраз встретилось 47332 уникальных. Средняя длина ключевой фразы составила 23,8 символа.

Основные тематики статей, вошедших в выборку, относятся к областям математики и физики. Среди собранных данных мы обнаружили 4282 уникальных значения УДК. Важно отметить, что у заметной доли статей было указано два или более значения УДК. Всего таких статей оказалось 3123 (16,2 % от всех статей в наборе данных). Эти статьи были исключены из дальнейшего рассмотрения. Статей, содержащих единственное значение УДК, оказалось 16110 (83,8 % от всех статей в наборе данных). На эти статьи в совокупности пришлось 2138 уникальных значения УДК.

Собранные статьи распределены по УДК достаточно неравномерно, что проявляется как в количестве статей с заданным значением УДК, так и в различной степени детализации разметки. Так, самое популярное значение УДК — 539.3 (Механика деформируемых тел. Упругость. Деформации) — указано в 529 статьях, а 20 самых популярных классов содержат в совокупности 4776 статей, или в 29,7 % от всего набора данных, 100 самых популярных — 9244 статей (57,4 % данных). На каждое из 1600 наименее популярных значений УДК приходится менее пяти статей, в совокупности им соответствует всего 2633 статей (16,4 % данных).

Детализация значений УДК тоже значительно разнится от статьи к статье. Чем более детализированное значение УДК указано в статье, тем точнее оно задаёт предметную область статьи. При этом из табл. 1 видно, что наиболее популярным является шестой уровень вложенности. Помимо слишком общо заданной тематики статьи, значения УДК с малой глубиной вложенности осложняют процесс обучения и оценку качества модели. Это связано с тем фактом, что сгенерированная метка может оказаться подклассом истинной. В подобных случаях весьма затруднительна автоматическая интерпретация, является ли сгенерированное значение действительно ошибочным, или уточнение истинной метки допустимо.

3. Модели и методы

3.1. Стратегии формирования целевой метки

Все классы УДК могут быть представлены как дерево-подобная структура, где каждый узел дерева представляет класс, и рёбра между узлами представляют иерархические отношения между классами. В рамках нашей работы мы рассматривали поставленную задачу как задачу плоской классификации, то есть генерировали сразу итоговую метку класса, а не проводили последова-

[6] https://www.mathnet.ru

Table 1. Number of articles depending on the discreteness of UDC codes

Таблица 1. Количество статей в зависимости от детализации УДК

Уровень детализации УДК	Количество статей
2	9
3	554
4	3252
5	3971
6	5307
7	2050
8	691
9	217
10	17
11	7
12	9
13	3
14	1
15	2
16	1

тельность иерархичных генераций, в ходе которых модель итеративно уточняет сгенерированную метку. При этом анализ набора данных показал, что распределение объектов по классам и детализация УДК крайне неравномерны. В связи с этим нами были предложены несколько стратегий для формирования целевой метки на основе оригинального кода УДК. Пример использования этих стратегий приведён на рис. 2. Рассматриваются статьи из четырёх различных УДК (стопки белых прямоугольников, высота стопки соответствует числу статей с соответствующим УДК). Для каждой из четырёх стратегий в зелёных прямоугольниках указаны метки, присвоенные статьям из стопки согласно выбранной стратегии.

1. **Наивный подход (Naive classifier).** В этом подходе мы игнорируем иерархическую структуру УДК и рассматриваем каждый класс как независимый. Этот метод прост в реализации, но он не учитывает отношения между классами и может привести к потере информации.
2. **Класс «редкие» (With «rare» class).** Этот подход похож на предыдущий, но в этом случае все классы, содержащие менее N экземпляров в обучающей выборе объединяются в специальный класс «редкие». Это позволяет сбалансировать соотношение классов в обучающей выборке и улучшить обучение модели.
3. **Фиксированный уровень иерархии (Fixed level).** В этом подходе на предварительном этапе метки классов обрезаются до определённой глубины УДК L. В случае, если после этой операции некоторые классы получают одинаковые метки, такие классы объединяются. Это позволяет уменьшить число слабопредставленных классов, однако может привести к менее точным результатам из-за огрубления меток.
4. **Объединение классов по количеству примеров с учётом таксономии (Union by number of examples).** Этот подход заключается в итеративном объединении классов, которые имеют малое количество статей, в родительский. В ходе предварительного этапа рассматриваются поддеревья УДК на материале обучающей выборки. В случае, если поддерево содержит менее S экземпляров, такое поддерево заменяется единственной вершиной, а меткой вершины становится наибольший общий префикс меток вершин, составлявших исходное поддерево. Как и в предыдущих случаях, это позволяет сбалансировать обучающую выборку, при этом потеряв меньше информации, чем при стратегиях 2 и 3.

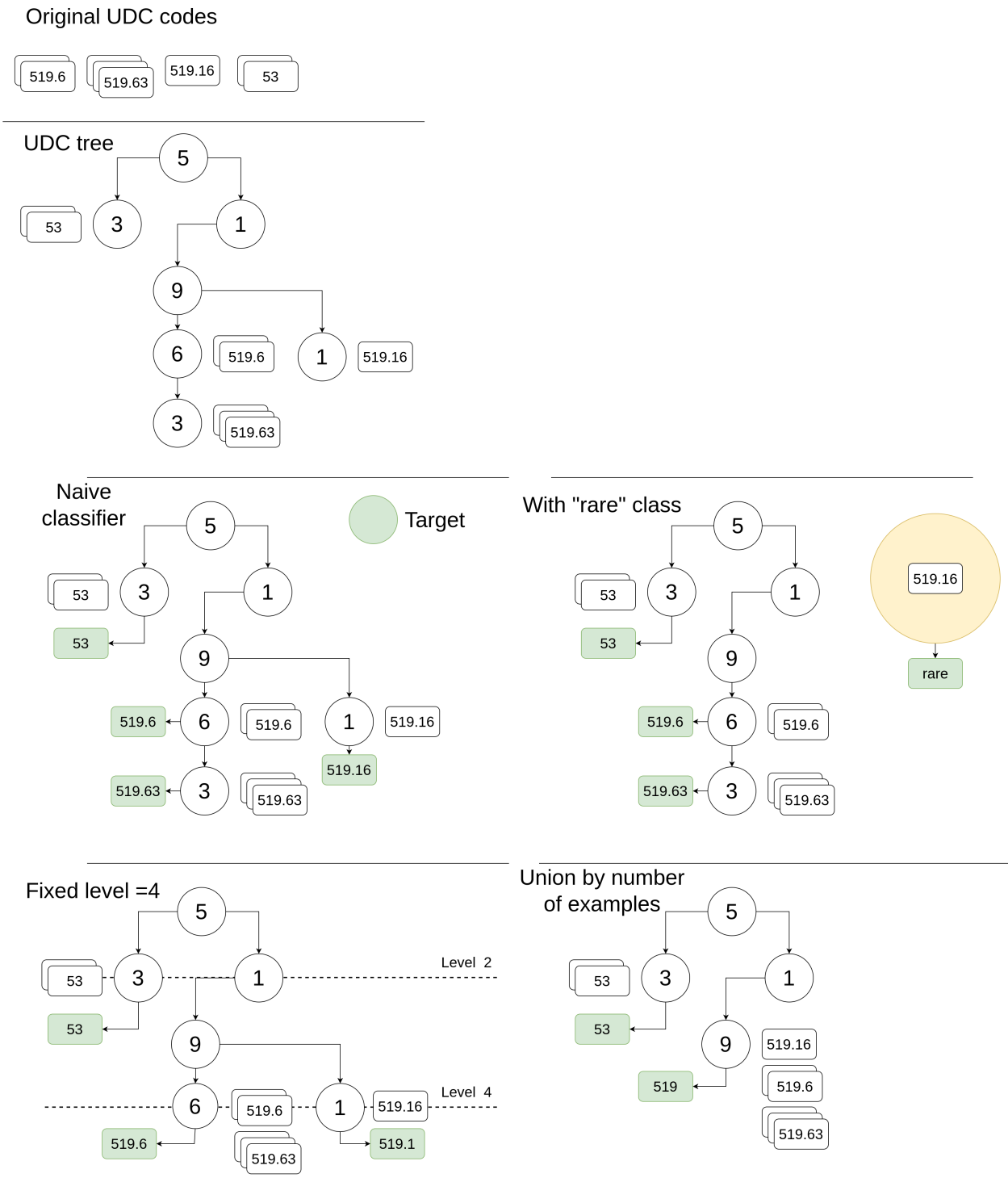

Fig. 2. Example of preparing target labels using different strategies

Рис. 2. Пример подготовки целевых меток при помощи различных стратегий

3.2. Стратегии формирования признакового описания

Собранный набор данных содержит значительное число полей, описывающих каждую публикацию. В качестве источников признакового описания мы выбрали три из них: заголовки статей, ключевые слова (указанные авторами публикаций) и аннотации. Мы изучили четыре способа формирования признакового описания на основе этих данных:

1. **Заголовки**. В этом случае в качестве признакового описания использовались только заголовки статей. Заголовки обычно содержат наиболее важную информацию о содержании статьи и могут быть полезны для её классификации. В то же время они могут быть недостаточно информативными для точной классификации, особенно если статья посвящена узкой тематике или в случае метафоричных заголовков[7].
2. **Заголовки + ключевые слова**. В этом случае в качестве признакового описания использовались заголовки и ключевые слова. Ключевые слова выбираются авторами так, чтобы отобразить тематику статьи, то есть решают схожую с метками УДК задачу. Следовательно, они могут предоставить дополнительную информацию о содержании статьи и помочь улучшить точность классификации. Тем не менее, предыдущие исследования [5] показали, что авторские ключевые слова не всегда идеально решают свою задачу. Кроме того, ключевые слова могут отсутствовать, в отличие от заголовка.
3. **Заголовки + аннотации**. В этом случае в качестве признакового описания использовались заголовки и аннотации. Аннотация статьи обычно содержит краткое описание проведённого исследования и может предоставить более полную информацию о теме статьи, чем заголовки и ключевые слова.
4. **Заголовки + ключевые слова + аннотации**. В этом случае в качестве признакового описания использовались все три выбранных поля. Подход предоставляет наиболее полную информацию о содержании статьи. Из общих соображений представляется, что это может быть наиболее эффективным способом формирования признакового описания с точки зрения точности итогового алгоритма.

3.3. Модели

Мы использовали две значительно различающиеся между собой предобученные языковые модели: многоязычную *intfloat/multilingual-e5-large*[8] [20] и ориентированную в первую очередь на русский язык *cointegrated/rubert-tiny2*[9].

Модель *multilingual-e5-large* является крупной (560 миллионов параметров) многоязычной моделью, поддерживающей 100 языков. Эта модель является дообученной на многоязычном наборе текстовых данных моделью *xlm-roberta-large*. Размер контекстного окна модели составляет 512 токенов. Эта модель ранее показала хорошие результаты на репрезентативном бенчмарке RuMTEB [21].

Крупные модели, подобные *multilingual-e5-large*, требуют для дообучения и использования значительных вычислительных мощностей. В качестве альтернативы мы рассмотрели модель *rubert-tiny2*, разработанную специально для обработки текстов на русском языке. Эта модель основана на архитектуре BERT и содержит 30 миллионов параметров. Размер её контекстного окна модели составляет 2048 токенов. За счёт малого размера возможно использование *rubert-tiny2* без графических ускорителей, что важно с прикладной точки зрения. При этом за счёт спецификации на русском языке модель может быть конкурентна на соответствующем домене.

[7] Особенно часто метафоричные заголовки используются в области искусственного интеллекта, как, например, в случае основополагающей для развития архитектуры Transformer статьи «Attention Is All You Need».

[8] https://huggingface.co/intfloat/multilingual-e5-large

[9] https://huggingface.co/cointegrated/rubert-tiny2

3.4. Постановка эксперимента

Для проведения численного эксперимента набор данных был случайным образом разделён на обучающую, валидационную и тестовую выборки в соотношении 70:15:15. Далее каждая из моделей дообучалась в течение 5 эпох с размером батча 12. В качестве оптимизатора использовался Adam [22]. Для модели *multilingual-e5-large* learning-rate зафиксирован в 10^{-5}, для модели *rubert-tiny2* — 10^{-4}. По результатам предварительных экспериментов были выбраны значения гиперпараметров для стратегий формирования целевой метки: в случае фиксированного уровня иерархии была выбрана длина метки 4, в случае объединения классов с учетом таксономии пороговый размер поддерева был выбран равным 100 статьям.

3.5. Метрики качества

Для оценки того, насколько хорошо модель обучилась, мы сравнивали её прогнозы с классами, сформированными при помощи соответствующей стратегии на базе авторской разметки. Например, для экспериментов с фиксированным уровнем иерархии, равным 4, мы сравнивали прогнозируемый класс не с исходным, а с полученным после огрубления разметки до четвёртого разряда в коде УДК. В качестве таких метрик мы использовали F-меру и Recall@k:

F-мера (F1). Одним из возможных сценариев использования нашей модели является присвоение статье наиболее вероятной метки УДК без рассмотрения остальных. Для оценки качества в таком случае хорошо подходит широко используемая в задачах классификации F-мера, позволяющая оценить баланс между точностью и полнотой модели. F-мера вычисляется как среднее гармоническое точности (precision) и полноты (recall). В свою очередь, точность вычисляется как

$$Precision = \frac{TP}{TP + FP},$$

а полнота — как

$$Recall = \frac{TP}{TP + FN}.$$

Вообще говоря, F-мера исходно разработана для бинарной классификации, и существует несколько способов обобщить её на многоклассовую постановку задачи. В настоящей работе мы использовали так называемое макроусреднение, при котором вычисляются глобальные TP, FP и FN без учёта класса.

Recall@k (R@k). В альтернативном случае потенциальный пользователь модели может рассмотреть несколько наиболее вероятных меток и выбрать из них самостоятельно. В таком случае подходящей метрикой является k-полнота (Recall@k) — доля статей, для которых истинный (авторский) УДК попадает в k наиболее вероятных предсказанных. Мы рассмотрели значения k равные 1, 2, 3 и 4.

Так как задача классификации научных статей имеет в первую очередь прикладное значение, мы сконструировали две метрики, отражающие различные сценарии использования итогового алгоритма:

Иерархическая классификационная точность (HA). С учётом того, что в процессе решения задачи классификации иерархическая природа УДК преобразуется в независимые (для модели) классы, необходимо уделять особое внимание тому, чтобы это не оказывало негативного влияния на пользовательскую оценку результата. Интуитивно понятно, что ошибка в старших разрядах УДК покажется пользователю значительно более грубой, чем ошибка в младших разрядах. Так, неумение модели отличить, например, юридические науки от физики (УДК 34 и 53, соответственно) намного менее приемлемо, чем ошибки в отделении консультационных экспертных систем (УДК

004.891.2) от диагностических (УДК 004.891.3). Для того, чтобы получить более справедливое представление о качестве создаваемой модели, мы разработали метрику, которую назвали *иерархической классификационной точностью*.

Для подсчёта иерархической классификационной точности целевое значение класса сравнивается с предсказанным поразрядно. Сравнение идёт слева направо до первого несовпадающего символа, таким образом, учитывается наибольший общий префикс двух классов p. Для подсчёта метрики суммируются веса w_i: совпадение i-го слева разряда добавляет к итоговому значению метрики $w_i = \frac{1}{i}$. Затем полученное значение делится на теоретический максимум величины, то есть на $\sum_{i=1}^{l} w_i$, где l — длина целевого значения УДК. Пусть тестовая выборка состоит из N статей, целевыми метками для которых являются $\overline{t} = (t^1, \ldots, t^N)$, а предсказанными — $\overline{p} = (p^1, \ldots, p^N)$. Тогда значение иерархической классификационной точности равно

$$\mathrm{HA}(\overline{t}, \overline{p}) = \frac{1}{N} \times \sum_{k=1}^{N} \frac{\sum_{i=1}^{lp_k} \frac{\mathbb{I}(t_i^k = p_i^k)}{i}}{\sum_{j=1}^{l_k} \frac{1}{j}},$$

где l_k — число разрядов в t^k, lp_k — длина наибольшего общего префикса t^k и p^k, t_i^k (p_i^k) — i-ый слева разряд t^k (p^k).

Иерархическая рекомендательная точность (HRA). Описывая взаимодействие с реальным пользователем системы, следует учитывать, что он способен в той или иной мере корректировать ошибки модели. В этом случае можно представить следующий сценарий использования: пользователь совершает последовательность выборов, углубляющих УДК, а модель на каждом шаге консультирует пользователя о наиболее вероятных с её точки зрения вариантах с учётом уже совершённых выборов. В таком случае, даже если модель перепутала, например, юридические науки и физику, пользователь сможет поправить её, и далее следует рассматривать предсказанные вероятности среди подклассов класса 53 «Физика». Заметим, что разрабатываемый нами классификатор позволяет такое использование: фактически, модель генерирует распределение вероятностей по классам. В таком случае, для любого префикса можно определить наиболее вероятный следующий разряд. Для оценки, учитывающей условные распределения, мы ввели метрику, которую назвали *иерархической рекомендательной точностью*. Пусть используются те же условные обозначения, что и в случае иерархической классификационной точности. Тогда значение иерархической рекомендательной точности равно:

$$\mathrm{HRA}(\overline{t}, \overline{p}) = \frac{1}{N} \times \sum_{k=1}^{N} \frac{\sum_{i=1}^{l_k} \frac{\mathbb{I}(t_i^k = p_i^k | (t_1^k, \ldots, t_{i-1}^k))}{i}}{\sum_{j=1}^{l_k} \frac{1}{j}},$$

где l_k — число разрядов в t^k, $p_i^k|(t_1^k, \ldots, t_{i-1}^k)$ — класс, который будет выбран в случае, если задан префикс $(t_1^k, \ldots, t_{i-1}^k)$, t_i^k (p_i^k) — i-й слева разряд t^k (p^k).

4. Результаты и их обсуждение

Полученные в ходе экспериментов результатов представлены в табл. 2. Для Recall@k приведены значения метрики при $k = 1 \ldots 4$. Для F-меры и Recall@k для каждого из типов классификаторов и каждой модели серым выделены лучшие полученные результаты в зависимости от способа формирования признакового описания. Для обеих иерархических точностей жирным на сером

фоне выделено лучшее полученное значение метрики среди всех изученных подходов. Серым выделены значения, отставшие от лучшего не более, чем на 1 п. п.

В большинстве случаев наибольших значений согласно F-мере и Recall@k удалось добиться при помощи наиболее подробного описания статей, включающего заголовок, ключевые слова и аннотацию. Это соотносится и с значениями иерархической точности обоих типов: за исключением классификатора, с использованием «редкого» класса лучшего результата с использованием полного признакового описания удалось добиться в четырёх из шести случаев согласно иерархической классификационной точности, в пяти из шести — согласно иерархической рекомендательной точности. При этом значения этих метрик существенно отличаются в зависимости от стратегии формирования целевых меток. Так, для фиксированного уровня иерархии удалось добиться значения Recall@1 более чем в два раза превышающего аналогичные значения для наивного классификатора (то есть для исходных меток из набора данных) и классификатора с использованием «редкого» класса.

Согласно обеим иерархическим метрикам, худшие результаты показал подход с «редким» классом. Остальные три подхода показали схожие результаты. Лучшего значения иерархической классификационной точности удалось добиться с использованием классификатора с фиксированным уровнем, полного признакового описания и модели multilingual-e5-large, а иерархической рекомендательной точности — с использованием наивного классификатора, полного признакового описания и модели multilingual-e5-large. Более того, результаты наивного классификатора для полного признакового описания и модели multilingual-e5-large уступили аналогичному классификатору с фиксированным уровнем менее 1 п. п.

Попарное сравнение иерархических точностей идентичных по архитектуре и способу формирования целевых меток классификаторов, использующих multilingual-e5-large и rubert-tiny2, показывает преимущество более крупной многоязычной модели. В большинстве пар разница составила более 1 п. п.

Дополнительно мы решили подробнее изучить случаи, при которых лучшая согласно иерархической рекомендательной сложности модель (наивный классификатор, полное признаковое описание, модель multilingual-e5-large) с высокой степенью уверенности выбрала метку УДК, отличающуюся от той, которая указана в наборе данных. Для этого мы упорядочили ошибки модели по степени уверенности и рассмотрели 100 первых среди них. В ходе слепого оценивания трое экспертов в области математики должны были ответить, какая из меток лучше подходит статье, если судить об этом по заголовку, ключевым словам и аннотации. Каждый из экспертов мог использовать одну из четырёх оценок: «первый УДК подходит лучше», «второй УДК подходит лучше», «оба УДК подходят хорошо», «оба УДК не подходят». Оказалось, что в 61 случае из 100 хотя бы два эксперта из трёх предпочли сгенерированную метку, и только в 19 случаев хотя бы два эксперта предпочли авторский код УДК. Это может подтверждает исходную гипотезу о том, что в части случаев выбираемый авторами УДК не лучшим образом соответствует их статье.

Для случаев, когда из двух предложенных большинство экспертов выбрали авторский вариант УДК, мы выделили три основных типа допущенных ошибок.

1. Недостаточно глубокая метка, например, 512.54 «Группы. Теория групп» при авторской 512.542 «Конечные группы».
2. Избыточно глубокая метка, например, 517.92 (в современной классификации отсутствует, заменён на 517.911 «Общие вопросы. Теоремы существования, теоремы единственности и теоремы о дифференциальных свойствах решений») при авторской 517.9 «Дифференциальные, интегральные и другие функциональные уравнения. Конечные разности. Вариационное исчисление.».
3. Авторская метка из слабо представленного класса, например, 004.8 «Искусственный интеллект».

Table 2. Experimental results **Таблица 2.** Результаты экспериментов

Способ учёта иерархии	Модель	Способ подачи данных	F1	R@1	R@2	R@3	R@4	HA	HRA
Наивный классификатор	**multilingual-e5-large**	Title	0,0461	0,2581	0,3410	0,4009	0,4387	0,7521	0,7935
		Title+KW	0,0619	0,3063	0,4072	0,4680	0,5108	0,7762	0,8214
		Title+Abstr	0,0611	0,3009	0,3919	0,4572	0,4968	0,7697	0,8173
		Title+KW+Abstr	0,0618	0,3032	0,4180	0,4806	0,5185	0,7756	**0,8220**
	rubert-tiny2	Title	0,0375	0,2437	0,3243	0,3730	0,4063	0,7337	0,7708
		Title+KW	0,0515	0,2815	0,3784	0,4275	0,4604	0,7541	0,7972
		Title+Abstr	0,0657	0,3113	0,3986	0,4599	0,4946	0,7584	0,8011
		Title+KW+Abstr	0,0724	0,3266	0,4086	0,4568	0,4928	0,7689	0,8042
С «редким» классом	**multilingual-e5-large**	Title	0,0470	0,2928	0,3973	0,4734	0,5203	0,5008	0,5236
		Title+KW	0,0650	0,3333	0,4586	0,5252	0,5788	0,5188	0,5481
		Title+Abstr	0,0696	0,3288	0,4500	0,5266	0,5811	0,5232	0,5527
		Title+KW+Abstr	0,0822	0,3257	0,4608	0,5392	0,5874	0,4830	0,5081
	rubert-tiny2	Title	0,0633	0,2671	0,3766	0,4347	0,4721	0,4259	0,5923
		Title+KW	0,0932	0,2941	0,4072	0,4824	0,5252	0,4791	0,6350
		Title+Abstr	0,1095	0,2995	0,4266	0,4946	0,5324	0,4660	0,6629
		Title+KW+Abstr	0,0930	0,2973	0,4311	0,4964	0,5392	0,4678	0,6325
Фиксированный (4-ый) уровень	**multilingual-e5-large**	Title	0,2092	0,6415	0,7710	0,8263	0,8548	0,7550	0,7733
		Title+KW	0,2408	0,7260	0,8316	0,8690	0,8907	0,7777	0,7909
		Title+Abstr	0,2335	0,6999	0,8189	0,8683	0,8922	0,7716	0,7885
		Title+KW+Abstr	0,2335	0,7433	0,8466	0,8855	0,9012	**0,7805**	0,7935
	rubert-tiny2	Title	0,1590	0,5966	0,7193	0,7747	0,8099	0,7430	0,7603
		Title+KW	0,1648	0,6699	0,7799	0,8234	0,8428	0,7629	0,7743
		Title+Abstr	0,1870	0,6654	0,7897	0,8353	0,8585	0,7619	0,7710
		Title+KW+Abstr	0,2368	0,7036	0,8099	0,8540	0,8795	0,7698	0,7787
С объединением классов (порог 100)	**multilingual-e5-large**	Title	0,3446	0,4126	0,5685	0,6536	0,7086	0,7444	0,7795
		Title+KW	0,4493	0,4932	0,6640	0,7410	0,7901	0,7687	0,8007
		Title+Abstr	0,4564	0,5009	0,6486	0,7302	0,7806	0,7652	0,7980
		Title+KW+Abstr	0,4707	0,5104	0,6671	0,7378	0,7955	0,7679	0,8037
	rubert-tiny2	Title	0,3580	0,4135	0,5383	0,6140	0,6694	0,7231	0,7587
		Title+KW	0,3580	0,4644	0,6149	0,7068	0,7563	0,7528	0,7877
		Title+Abstr	0,4334	0,4784	0,6288	0,6869	0,7351	0,7539	0,7835
		Title+KW+Abstr	0,4551	0,4901	0,6324	0,7009	0,7545	0,7536	0,7981

Среди случаев, когда эксперты не смогли определить лучшую метку, выделяется статья с кодом 544.452.2 «Горение. Распространение пламени» из области химии, для которой модель указала семантически схожий код 536.46 «Горение и другие реакции. Пламя» из области физики. Это показывает, что предметом отдельного исследования должен стать анализ семантической близости вершин дерева УДК, номинально далёких друг от друга с точки зрения расстояния по дереву.

Полученные нами результаты значительно расширяют ранее представленные в этой области. Обученные нами классификаторы работают с гораздо более детализированными кодами УДК. Этим, в частности, объясняется низкое качество наших моделей согласно F-мере в сравнении с [17]: если у авторов указанной работы рассматривалась классификация с фиксированным уровнем 2 (причём отсутствовала проблема вложенных классов), то в нашей работе наиболее приближенным следует считать эксперимент с фиксированным уровнем 4 (причём вложенные классы присутствуют, так как в наборе данных присутствовали статьи с уровнями 2 и 3). При этом значение HA для этой модели показывает, что модель с высокой точностью восстанавливает префиксы длины 4. Кроме того, в нашем подходе, в отличие от [17], не задействован полный текст статьи, что расширяет область потенциального применения алгоритма: во многих интернет-библиотеках научных статей полный текст работы в общем случае недоступен, в отличие от аннотации, заголовка и ключевых слов. Важным достижением в сравнении с предыдущими подходами мы считаем и внедрение двух иерархических метрик: в [17] используются лишь точность (accuracy), F-мера, точность (precision) и полнота (recall), не учитывающие иерархическую природу классификации, а в [18, 19] отсутствуют численные метрики качества классификации.

5. Заключение

В настоящей работе рассматривается задача автоматической иерархической классификации научных статей согласно универсальному десятичному классификатору. Актуальность исследования подобных подходов обусловлена быстрым ростом числа ежегодно публикуемых научных работ и необходимостью организации поиска по научным библиотекам. Мы рассмотрели подход, основанный на дообучении BERT-подобных моделей: многоязычной multilingual-e5-large и ориентированной на русский язык rubert-tiny2. Мы решали задачу при помощи плоских классификаторов, не учитывающих иерархичность классификации. При этом для учёта иерархичности УДК мы использовали несколько стратегий формирования целевых меток. Для оценки качества итоговых моделей нами были предложены две метрики, направленные на различные сценарии использования моделей: иерархическая классификационная точность (HA) и иерархическая рекомендательная точность (HRA). Лучших результатов удалось достигнуть при помощи модели multilingual-e5-large, использующей признаковое описание из заголовка, ключевых слов и аннотации к статьям. С точки зрения иерархической классификационной точности лучшей (HA=0,7805) оказалась модель, обученная с приведением целевых меток УДК к четвёртому уровню вложенности, а с точки зрения иерархической рекомендательной точности лучший результат (HRA=0,8220) показал классификатор, обученный на исходных метках УДК из набора данных. Эту же модель следует считать лучшей по совокупности метрик среди опробованных подходов. Выборочное тестирование случаев, когда предсказанный класс отличается от эталонного, показало, что по крайней мере некоторые расхождения объясняются некорректными кодами УДК, выбранными авторами статей. К ограничениям нашего исследования следует отнести малое число опробованных BERT-подобных моделей и отсутствие сравнения с иерархическими классификаторами. В ходе последующих экспериментов мы планируем устранить эти недостатки.

References

[1] M. Gusenbauer, “Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases”, *Scientometrics*, vol. 118, pp. 177–214, 2019. DOI: 10.1007/s11192-018-2958-5.

[2] M. Fire and C. Guestrin, “Over-optimization of academic publishing metrics: Observing Goodhart’s Law in action”, *GigaScience*, vol. 8, giz053, Jun. 2019. DOI: 10.1093/gigascience/giz053.

[3] R. Martinez-Cruz, A. J. Lopez-Lopez, and J. Portela, “ChatGPT vs state-of-the-art models: A benchmarking study in keyphrase generation task”, *Applied Intelligence*, vol. 55, no. 1, pp. 1–25, 2025. DOI: 10.1007/s10489-024-05901-4.

[4] M. Song *et al.*, *Is ChatGPT a good keyphrase generator? A preliminary study*, 2023. arXiv: 2303.13001 [cs.CL].

[5] A. Glazkova, D. Morozov, and T. Garipov, *Key algorithms for keyphrase generation: Instruction-based LLMs for Russian scientific keyphrases*, 2024. arXiv: 2410.18040 [cs.CL].

[6] A. V. Glazkova, D. A. Morozov, M. S. Vorobeva, and A. A. Stupnikov, “Keyword generation for Russian-language scientific texts using the mT5 model”, *Automatic Control and Computer Sciences*, vol. 58, no. 7, pp. 995–1002, 2024. DOI: 10.3103/S014641162470041X.

[7] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey”, *Information*, vol. 10, no. 4, p. 150, 2019. DOI: 10.3390/info10040150.

[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning–based text classification: A comprehensive review”, *ACM Computing Surveys*, vol. 54, no. 3, 2021. DOI: 10.1145/3439726.

[9] S. Garg and G. Ramakrishnan, "BAE: BERT-based adversarial examples for text classification", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6174–6181. DOI: 10.18653/v1/2020.emnlp-main.498.

[10] X. Sun *et al.*, "Text classification via large language models", in *Findings of the Association for Computational Linguistics*, 2023, pp. 8990–9005. DOI: 10.18653/v1/2023.findings-emnlp.603.

[11] K. Kowsari, D. E. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber, and L. E. Barnes, "HDLTex: Hierarchical deep learning for text classification", in *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications*, 2017, pp. 364–371. DOI: 10.1109/ICMLA.2017.0-134.

[12] S. Strydom, A. M. Dreyer, and B. van der Merwe, "Automatic assignment of diagnosis codes to free-form text medical note", *Journal of Universal Computer Science*, vol. 29, no. 4, pp. 349–373, 2023. DOI: 10.3897/jucs.89923.

[13] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings", *Information Sciences*, vol. 471, pp. 216–232, 2019. DOI: 10.1016/j.ins.2018.09.001.

[14] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research", *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[15] Y. Wang *et al.*, "Towards better hierarchical text classification with data generation", in *Findings of the Association for Computational Linguistics*, 2023, pp. 7722–7739. DOI: 10.18653/v1/2023.findings-acl.489.

[16] A. Zangari, M. Marcuzzo, M. Rizzo, L. Giudice, A. Albarelli, and A. Gasparetto, "Hierarchical text classification and its foundations: A review of current research", *Electronics*, vol. 13, no. 7, p. 1199, 2024. DOI: 10.3390/electronics13071199.

[17] M. Kragelj and M. Borstnar, "Automatic classification of older electronic texts into the Universal Decimal Classification-UDC", *Journal of Documentation*, vol. 77, no. 3, pp. 755–776, 2021. DOI: 10.1108/JD-06-2020-0092.

[18] A. Y. Romanov, K. E. Lomotin, E. S. Kozlova, and A. L. Kolesnichenko, "Research of neural networks application efficiency in automatic scientific articles classification according to UDC", in *Proceedings of the International Siberian Conference on Control and Communications (SIBCON)*, IEEE, 2016, pp. 1–5.

[19] O. Nevzorova and D. Almukhametov, "Towards a recommender system for the choice of UDC code for mathematical articles", in *Supplementary Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains*, 2021, pp. 54–62.

[20] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, *Multilingual E5 text embeddings: A technical report*, 2024. arXiv: 2402.05672 [cs.CL].

[21] A. Snegirev, M. Tikhonova, A. Maksimova, A. Fenogenova, and A. Abramov, *The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design*, 2025. arXiv: 2408.12503 [cs.CL].

[22] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.