УДК 004.912

МЕТОДЫ АВТОМАТИЧЕСКОГО ПРИСВОЕНИЯ КОДОВ УДК МАТЕМАТИЧЕСКИМ СТАТЬЯМ: ОЦЕНКА КЛАССИЧЕСКИХ И НЕЙРОСЕТЕВЫХ ПОДХОДОВ

Б. Т. Гизатуллин[1] [0009-0000-6251-9260], О. А. Невзорова[2] [0000-0001-8116-9446]

[1, 2]Казанский (Приволжский) федеральный университет, г. Казань, Россия

[1]gizat.blt@gmail.com, [2]onevzoro@gmail.com

Аннотация

Универсальная десятичная классификация (УДК) – это иерархическая система индексирования, в рамках которой одной публикации могут соответствовать один или несколько кодов. Ручное присвоение кодов УДК трудоемко и нередко оказывается неоднородным. В работе рассмотрена задача автоматического присвоения кодов УДК русскоязычным математическим статьям. Цель исследования – сравнить различные сочетания текстовых представлений и моделей классификации на едином корпусе и определить наиболее эффективные конфигурации. Для этого был сформирован корпус из 4194 статей с ресурса Math-Net.Ru, включающий полные тексты, аннотации, метаданные и коды УДК; были выполнены извлечение текста из PDF-файлов, очистка артефактов верстки и нормализация кодов. В эксперименте сопоставлялись текстовые представления TF-IDF, Word2Vec, SciRus-tiny и SciRus-tiny3.5 в сочетании с моделями логистической регрессии, Complement Naive Bayes (CNB) и CatBoost. Наилучшие результаты в обеих постановках – однозначной (single-label) и многозначной (multi-label) – показала модель TF-IDF + LogReg; близкие результаты продемонстрировала конфигурация TF-IDF + CNB. Полученные результаты могут быть использованы при разработке систем автоматической рубрикации научных публикаций, рекомендательных сервисов для авторов и редакторов, а также средств контроля качества тематической разметки.

Ключевые слова: *автоматическая классификация, универсальная десятичная классификация, УДК, обработка научных текстов, машинное*

обучение, иерархическая классификация, многозначная классификация, математические тексты, цифровые библиотеки, векторизация текста.

ВВЕДЕНИЕ

Универсальная десятичная классификация (УДК) применяется для тематического индексирования научных публикаций, однако ручное присвоение кодов трудоемко и подвержено субъективности, что затрудняет масштабирование на большие массивы текстов. Поэтому актуальна разработка методов автоматического определения УДК по содержанию документа на основе подходов обработки естественного языка и машинного обучения.

Для электронных библиотек и научных архивов задача автоматического индексирования имеет не только исследовательское, но и прикладное значение: от качества тематической классификации документа зависят полнота тематического поиска, корректность навигации по коллекциям и возможность последующего анализа структуры фонда. Ранние работы по автоматической классификации в библиотечно-информационных системах показали, что центральной проблемой остается согласование текстового содержания документа с формальной системой знаний, принятой в конкретной предметной области [1].

Для математических публикаций эта задача осложняется спецификой материала. Помимо общеязыковой и научной лексики, тексты содержат формулы, символические обозначения и устойчивые терминологические сочетания, характерные для отдельных разделов математики. Поэтому границы между классами зависят как от выбора признакового пространства, так и от полноты текстового представления: одна и та же статья может сочетать общетеоретическую лексику и узкоспециальные термы, указывающие на более точный код УДК.

Дополнительную сложность создают иерархическая структура УДК и близость отдельных предметных областей; междисциплинарные тексты могут соответствовать нескольким кодам, а границы между классами часто размыты. Цель настоящей работы – сравнить модели автоматического присвое-

ния кодов УДК для математических публикаций и проанализировать структуру ошибок, включая наиболее информативные термы для различных кодов и те области УДК, которые труднее всего различать между собой.

ИССЛЕДОВАНИЯ, БЛИЗКИЕ ПО ТЕМАТИКЕ

В работах по автоматической классификации кодов УДК одним из наиболее популярных является подход, основанный на алгоритмах машинного обучения. Например, в статье [2] использованы алгоритмы машинного обучения и обработки текста, включая TF-IDF, косинусное сходство, наивный байесовский классификатор и многослойный перцептрон. В [3] исследована эффективность различных архитектур искусственных нейронных сетей для автоматической классификации научных статей по УДК и отмечены возможные области практического применения таких систем. В [4] для определения кода УДК применены алгоритмы SVM и *k*-ближайших соседей.

В более поздних работах все чаще рассматривались нейросетевые и рекомендательные подходы к классификации документов по УДК. В [5] исследовано применение предобученной модели BERT для полуавтоматической предметной идентификации документов и построения рекомендательной системы присвоения кодов УДК. В [6] предложен гибридный подход на основе рекомендательной системы; для его оценки использованы метрики качества ранжирования NDCG, MRR и MAP, а среди наиболее результативных конфигураций были варианты, сочетающие BM25, BERT и дополнительные этапы переупорядочивания рекомендаций. В работе [7] рассмотрена задача классификации научных статей с использованием глубоких нейронных сетей с учетом иерархической структуры УДК. Авторы отмечают, что в некоторых случаях ошибки классификации вызваны некорректно проставленными кодами. В работе [8] рассмотрено использование больших языковых моделей (Large Language Model, LLM) в роли рекомендательной системы для подбора кодов УДК и сравнены различные LLM, что дополняет классические подходы альтернативной парадигмой автоматизации индексирования.

Отдельное направление связано не только с выбором модели, но и с явным учетом иерархической природы меток. В обзоре [9] подчеркнуто, что для иерархической классификации недостаточно плоских метрик: при оценке

качества важно учитывать совпадение предсказанных и истинных меток вместе с их предками в иерархии. Обзор по иерархической классификации текстов [10] и обзор по иерархической многозначной классификации [11] показывают, что методы оценки, декодирования и выбора порогов должны учитывать структуру целевого пространства меток. Для УДК это особенно важно, поскольку иерархия классов задает смысловую близость соседних подклассов и допускает присвоение одному документу нескольких кодов.

Более широкий контекст задают обзоры по автоматической классификации текстов [12–14], в которых систематизируются основные типы текстовых представлений и моделей – от разреженных векторных схем и распределенных представлений до глубоких нейросетевых архитектур. В этих работах показано, что сравнительная эффективность методов определяется не только архитектурой модели, но и свойствами корпуса, включая объем и структуру обучающей выборки, баланс классов и языковые либо доменные особенности данных.

В развитии подходов к представлению текста в задачах автоматической обработки и классификации прослеживается переход от распределенных векторных представлений слов [15] к контекстным моделям на основе трансформеров [16] и далее к специализированным энкодерам, обученным с учетом специфики научного дискурса [17]. Для корпуса математических статей такой переход представляет практический интерес: специализированные модели потенциально лучше учитывают особенности научного дискурса, тогда как лексико-частотные и иные разреженные признаки могут оставаться полезными при различении близких тематик. Именно поэтому в настоящей работе сопоставляются методы, основанные на различных типах текстового представления, но оцениваемые в едином экспериментальном контуре.

Таким образом, анализ литературы показал, что задача автоматического присвоения кодов УДК находится на пересечении нескольких исследовательских направлений: библиотечно-информационного индексирования, иерархической и многозначной классификации, а также обработки научных текстов. Это делает сопоставление методов на математическом корпусе статей на русском языке самостоятельной и практически значимой задачей.

МЕТОД

Корпус формировался автоматически на основе статей с ресурса Math-Net.Ru: для выбранных журналов были отобраны русскоязычные статьи, опубликованные не ранее 2000 г.; для каждой статьи сохранялись PDF-файл полного текста и метаданные (год, заголовок, аннотация, авторы). Итоговый объем корпуса составил 4194 статьи.

Из файлов в формате PDF извлекались коды УДК, год, основной текст и формулы; далее выполнялась очистка от артефактов верстки (служебные блоки, колонтитулы, шапки/подвалы). Для методов TF-IDF и Word2Vec дополнительно выполнялись лемматизация и удаление стоп-слов.

На этапе подготовки корпуса существенным было приведение кодов УДК к сопоставимому виду. В рамках настоящей работы анализ ограничивается глубиной до одного знака после первой точки, что позволяет, с одной стороны, сохранить различимость основных предметных ветвей внутри математической области, а с другой – снизить чувствительность модели к избыточной дробности и единичным вариантам разметки. Такое решение особенно важно для сравнительного исследования, в котором требуется сопоставить модели на едином и достаточно устойчивом уровне детализации.

Для проверки обобщающей способности использовалось временное разбиение: обучение проводилось на статьях до 2020 г., тестирование – на статьях, опубликованных с 2021 г. Временное разбиение выбрано намеренно, поскольку в прикладной системе автоматической категоризации модель чаще применяется к новым публикациям. Такой протокол снижает риск завышенной оценки качества за счет тематического и стилистического сходства текстов одного периода, а также позволяет оценить устойчивость признаков и моделей к возможному смещению распределения корпуса во времени.

Сравнивались четыре типа текстовых представлений: TF-IDF, Word2Vec с усреднением векторов слов с TF-IDF-весами (далее W2V), а также энкодерные представления SciRus-tiny и SciRus-tiny3.5 (mlsa-iai-msu-lab). Для моделей семейства SciRus-tiny текст статьи преобразовывался в последовательность фрагментов, для каждого из которых вычислялось векторное представление; затем эмбеддинги агрегировались во взвешенное среднее и нормализовывались, образуя единый вектор документа. В конфигурации SciRus-

tiny3.5+LogReg(abstract) на вход модели подавались заголовок и аннотация статьи, а в конфигурации SciRus-tiny3.5+LogReg (fulltext) – заголовок и очищенный полный текст; схема кодирования и агрегации эмбеддингов в обоих случаях оставалась одинаковой. В качестве классификаторов использовались логистическая регрессия, Complement Naive Bayes (CNB) и градиентный бустинг на основе CatBoost [18]. Гиперпараметры всех конфигураций подбирались в едином контуре с помощью байесовской оптимизации (Optuna [19]) по кросс-валидации внутри обучающего периода; для каждой конфигурации выполнялось по 100 итераций поиска. В многозначной постановке после обучения для каждого класса дополнительно подбирались индивидуальные пороги бинаризации на валидационной выборке.

Выбор именно этих моделей обусловлен стремлением сопоставить методы, различающиеся как по типу представления текста, так и по способу принятия решения. Логистическая регрессия задает интерпретируемую линейную границу и естественно сочетается с разреженными признаками. Complement Naive Bayes включен в сравнение как сильный базовый метод для высокоразмерных разреженных текстовых представлений; кроме того, он нередко оказывается устойчивым при дисбалансе классов, что особенно важно для задачи присвоения кодов УДК. Градиентный бустинг, в свою очередь, позволяет проверить, дает ли нелинейное моделирование дополнительный выигрыш на тех же входных данных. Благодаря этому сравнение оказывается содержательным не только по итоговым метрикам, но и с точки зрения того, какие свойства корпуса и представления текста оказываются критичными для качества классификации.

Особенность кодов УДК состоит в их иерархичности: коды отражают как общую область, так и более узкую тематику. Поэтому при оценке качества и подборе гиперпараметров учитывается не только точное совпадение класса, но и близость предсказания к истинной ветви УДК.

Разделение на single-label и multi-label постановки отражает два различных прикладных сценария. В первом случае система должна выбрать основной, наиболее представительный код, что ближе к задаче первичной каталогизации. Во втором случае модель должна воспроизвести весь набор темати-

ческих индексов, что ближе к задаче поддержки экспертной разметки и проверки уже существующей классификации. Сравнение этих постановок на одном корпусе по-зволяет оценить, насколько одни и те же признаки по-разному работают в режимах выбора одного класса и выбора набора взаимосвязанных классов.

В постановке с единственной целевой меткой (single-label) оптимизируется критерий

$$S = 0.7 \cdot \mathrm{F1}_{\mathrm{macro}} + 0.3 \cdot \mathrm{F1}_{\mathrm{hier}}.$$

Здесь $\mathrm{F1}_{\mathrm{macro}}$ – макроусредненная F1-мера, а $\mathrm{F1}_{\mathrm{hier}}$ вводится так:

$$L(y) = \{y_1, y_2, \ldots, y_n\}, \qquad L(\hat{y}) = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}.$$

$$P = \frac{|L(y) \cap L(\hat{y})|}{|L(\hat{y})|}, \qquad R = \frac{|L(y) \cap L(\hat{y})|}{|L(y)|}.$$

$$F1_{\mathrm{hier}}(y, \hat{y}) = \frac{2PR}{(P + R)},$$

где $y_1, \ldots, y_n$ – уровни рассматриваемой детализации. Мы используем два уровня для этой метрики: y_1 – код до первой точки, y_2 – код с полной рассматриваемой глубиной (до первой цифры после точки).

В постановке с несколькими метками (multi-label) оптимизируется критерий

$$S = 0.5 \cdot \mathrm{F1}_{\mathrm{micro}} + 0.5 \cdot \mathrm{F1}_{\mathrm{macro}}.$$

Здесь $\mathrm{F1}_{\mathrm{micro}}$ – микроусредненная F1-мера, вычисляемая по агрегированным значениям ошибок и верных результатов по всем меткам, поэтому она преимущественно характеризует качество на частотных классах. Вероятности по классам бинаризовались с использованием индивидуальных порогов, которые для каждого класса подбирались на валидации путем максимизации F1 на заданной сетке значений.

Использование индивидуальных порогов принципиально важно для многозначной постановки, поскольку частота классов и характер их совместной встречаемости различаются. Для устойчивых и частых кодов допустим более низкий порог, если это повышает полноту, тогда как для семантически близких ветвей требуется более осторожная бинаризация, уменьшающая

число ложных срабатываний. Таким образом, подбор порогов становится частью общей адаптации модели к структуре корпуса.

Также рассчитывались метрики: доля правильных ответов (accuracy, acc.), метрики на верхнем уровне – до первой точки, 2-го уровня детализации (top-level) и другие стандартные метрики.

ЭКСПЕРИМЕНТЫ

Экспериментальная часть была построена так, чтобы сравнение моделей опиралось не на одну сводную метрику, а на несколько показателей качества. Для однозначной постановки важны прежде всего устойчивость на редких классах, качество вероятностных оценок и способность модели выводить верные коды в верхние позиции ранжирования. Для многозначной постановки, помимо F1-мер, учитывались показатели качества ранжирования кодов и степень совпадения полных наборов меток на уровне документа.

В однозначной постановке (предсказание основного кода) сравнивались семь моделей: TF-IDF + LogReg, TF-IDF + CatBoost, W2V (усреднение Word2Vec с TF-IDF-весами) + LogReg, SciRus-tiny + LogReg, SciRus-tiny3.5 + LogReg(abstract), SciRus-tiny3.5 + LogReg(fulltext) и TF-IDF + CNB.

Табл. 1. Метрики моделей в задаче предсказания первого кода УДК.

Модель	Acc.	Balanced Acc.	Macro-F1	Weighted F1	Hier-F1	LogLoss	Acc.@3	Top-level Acc.	Top-level Macro-F1
TF-IDF+LogReg	0.857	0.702	0.694	0.855	0.889	0.564	0.970	0.920	0.860
TF-IDF+CatBoost	0.828	0.593	0.638	0.820	0.866	0.597	0.964	0.904	0.781
W2V+LogReg	0.763	0.620	0.595	0.770	0.807	1.148	0.936	0.851	0.755
SciRus-tiny+LogReg	0.705	0.486	0.506	0.699	0.751	1.037	0.928	0.797	0.661
SciRus-tiny3.5+ LogReg(abstract)	0.778	0.557	0.598	0.768	0.818	0.722	0.939	0.859	0.713
SciRus-tiny3.5+ LogReg(fulltext)	0.791	0.559	0.597	0.785	0.830	0.688	0.947	0.867	0.743
TF-IDF+CNB	0.826	0.643	0.629	0.825	0.856	2.708	0.917	0.886	0.807

Наиболее устойчивые результаты на рассматриваемом уровне детализации показала модель TF-IDF + LogReg (см. табл. 1): она обеспечивает наилучший баланс между общей точностью, качеством на редких классах и ранжирующими метриками. Модель TF-IDF + CatBoost близка по общей точности, но сильнее чувствительна к дисбалансу классов. Конфигурация W2V + LogReg уступает на детальном уровне, что, вероятно, связано со сглаживанием различий между близкими терминами при усреднении эмбеддингов. Использование SciRus-tiny без дополнительной адаптации также не дало преимущества. Конфигурация TF-IDF+CNB показала результаты, близкие к TF-IDF + LogReg. Обе версии SciRus-tiny3.5 улучшили показатели относительно исходной SciRus-tiny + LogReg; при этом вариант с полным текстом оказался сильнее варианта с аннотацией.

Таким образом, в однозначной постановке TF-IDF + LogReg демонстрирует не только высокую точность, но и наиболее стабильное качество по совокупности метрик, что делает эту модель перспективной для практических сценариев автоматического подбора основного кода УДК.

Анализ ошибок показал, что труднее всего разделяются близкие подклассы внутри одной ветви УДК; наиболее заметная путаница наблюдается

для пары 517.9 ↔ 517.5. Ошибки между разными ветвями (например, между 517.* и 519.*) встречаются реже и обычно связаны с пересечением терминологии в смежных темах. Характерные термы (табл. 2) подтверждают тематичность признаков: 510.5 соответствует теории алгоритмов и вычислимых функций, 510.6 – математической логике, 512.5 – общей алгебре. Визуализация t-SNE [20] для модели W2V (рис. 1) показывает разделимость на верхнем уровне и частичное смешение на внутреннем, что согласуется с профилем ошибок.

Табл. 2. Наиболее информативные термы TF-IDF для трех кодов УДК по коэффициентам логистической регрессии.

Код	Топ 1 терм	Топ 2 терм	Топ 3 терм	Топ 4 терм	Топ 5 терм	Топ 6 терм
510.5	вычислимый	нумерация	шаг	перечислимый	вычислимость	конструкция
510.6	кортеж	логика	пропозициональный	полигон	язык	категоричный
512.5	подгруппа	группа	идеал	алгебра	кольцо	изоморфный

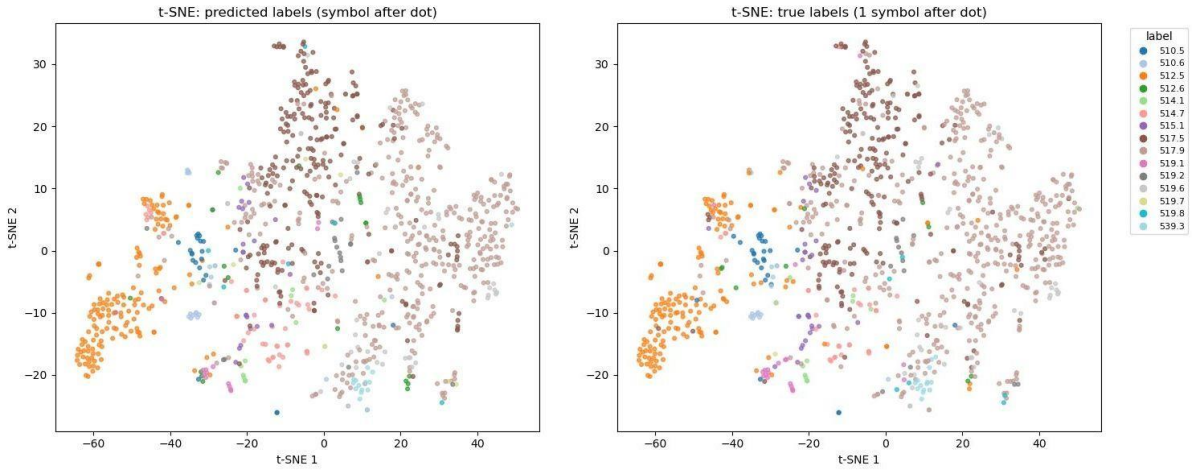

Рис. 1. Визуализация истинных и предсказанных меток кодов УДК с помощью
t-SNE для модели W2V + LogReg.

В свою очередь, такой характер ошибок объясняется иерархической природой задачи. Для прикладной рекомендательной системы смешение

внутри одной тематической ветви обычно менее критично, чем переход в совершенно другую область, именно поэтому наряду с точными метриками в работе анализируются показатели верхнего уровня и иерархическая F1-мера.

В многозначной постановке сравнение проводилось на том же наборе моделей. Вероятности по классам переводились в бинарные решения с использованием индивидуальных порогов, подобранных на валидации для каждого кода. По совокупности метрик наилучшие результаты вновь показала модель

TF-IDF + LogReg: она превосходит альтернативы как по качеству бинарного решения, так и по ранжирующим метрикам. TF-IDF + CNB выступает ближайшей альтернативой, тогда как обе конфигурации SciRus-tiny3.5 улучшают результаты относительно SciRus-tiny+LogReg. При этом вариант с полным текстом устойчиво превосходит вариант, основанный только на заголовке и аннотации.

Распределение числа меток (рис. 2) показывает, что большинство статей имеют один-два кода, а предсказания лучшей модели в целом воспроизводят этот профиль.

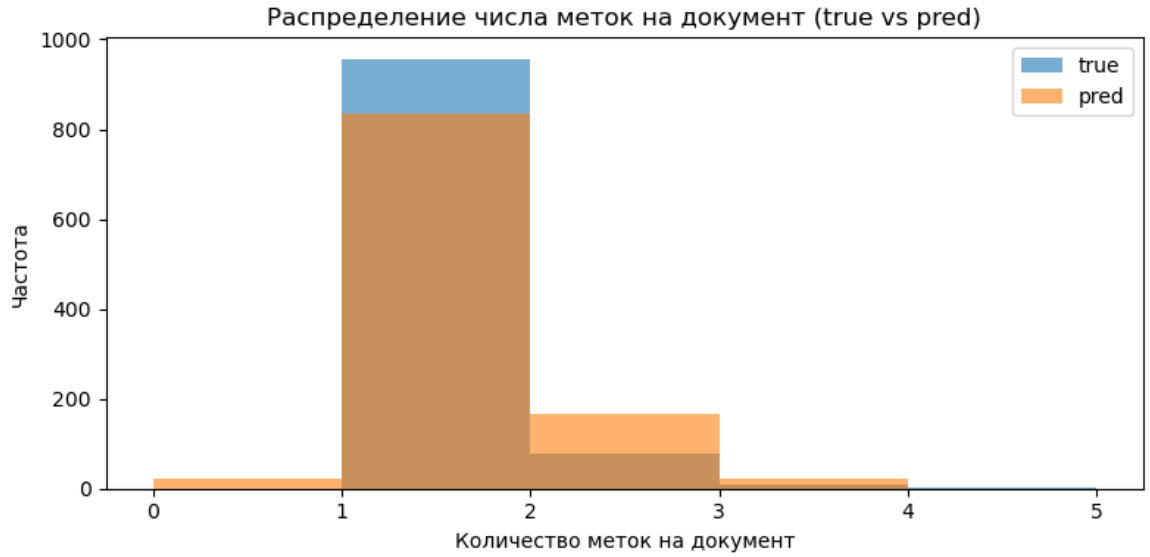

Рис. 2. Распределение числа кодов на документ для модели TF-IDF+LogReg.

Подобранные пороги зависят от частоты и «конкурентности» класса: для устойчивых частых кодов порог может быть ниже, тогда как для семантически близких ветвей он повышается, что уменьшает число ложных срабатываний. Hamming для SciRus-tiny + LogReg оказался ниже остальных моделей из-за того, что модель в целом предсказывала меньше меток и чаще попадала в ответ «такого кода нет». В итоге модель TF-IDF + LogReg превосходит другие рассматриваемые модели и по качеству классификации, и по ранжирующим метрикам (LRAP, mAP), то есть лучше упорядочивает коды (см. табл. 3).

Табл. 3. Метрики моделей в задаче предсказания всех кодов УДК статей.

Модель	Micro-F1	Macro-F1	Samples F1	Hamming	Subset Acc.	LRAP	mAP-micro	mAP-macro	Top-level micro-F1	Top-level macro-F1	Top-level Subset Acc.
TF-IDF + LogReg	0.812	0.646	0.828	0.025	0.717	0.916	0.887	0.727	0.885	0.808	0.820
W2V + LogReg	0.754	0.529	0.756	0.031	0.648	0.871	0.802	0.588	0.833	0.729	0.743
TF-IDF + CatBoost	0.796	0.569	0.796	0.027	0.685	0.891	0.840	0.630	0.871	0.758	0.801
SciRus-tiny + LogReg	0.659	0.149	0.641	0.014	0.531	0.807	0.672	0.302	0.743	0.318	0.640
SciRus-tiny3.5 + LogReg (abstract)	0.724	0.506	0.727	0.035	0.603	0.855	0.797	0.534	0.806	0.621	0.715
SciRus-tiny3.5 + LogReg (fulltext)	0.736	0.544	0.732	0.035	0.599	0.872	0.806	0.592	0.808	0.668	0.694
TF-IDF + CNB	0.790	0.621	0.787	0.027	0.673	0.909	0.848	0.665	0.851	0.776	0.759

Содержательно различие между моделями в multi-label постановке можно интерпретировать так: при множественном присвоении кодов особенно важна способность модели удерживать несколько близких тематических сигналов одновременно.

Разреженные признаки TF-IDF в этих условиях сохраняют различимость частотных и редких терминов, тогда как усреднение плотных векторов сильнее сглаживает различия между соседними тематическими зонами. Это согласуется и с тем, что top-level показатели заметно выше: большая часть ошибок приходится не на выбор неверной верхней ветви, а на выбор соседнего подкласса внутри уже верно определенного направления.

ЗАКЛЮЧЕНИЕ

Исследованы подходы к автоматическому присвоению кодов УДК математическим публикациям на русском языке на корпусе из 4194 статей, собранном на основе ресурса Math-Net.Ru. Оценка проведена на временном разбиении, что позволило более реалистично оценить обобщающие способности моделей.
Задача рассматривалась в однозначной и многозначной постановках на уровне одного знака после первой точки с дополнительным учетом иерархической структуры УДК.

Результаты проведенных экспериментов показали, что при выбранной детализации наиболее стабильные результаты в обеих постановках обеспечивает модель TF-IDF + LogReg. Ее преимущество проявляется не только в основных метриках классификации, но и в показателях, связанных с ранжированием кандидатов, что особенно важно для полуавтоматических сценариев библиотечного индексирования. Конфигурация TF-IDF + CNB выступает сильной альтернативой и дает близкие результаты. Переход от SciRus-tiny к SciRus-tiny3.5 улучшает качество семантических конфигураций, однако даже лучший вариант с полным текстом уступает TF-IDF-базовым моделям на исследуемом корпусе при выбранной агрегации.

С практической точки зрения полученные результаты позволяют рассматривать разработанный подход как основу для нескольких сценариев внедрения: автоматической первичной рубрикации новых поступлений, рекомендаций кодов автору или редактору при подготовке публикации, а также проверки уже существующей разметки в ретроспективных коллекциях. По-

следний сценарий особенно важен для цифровых библиотек, поскольку позволяет использовать модель не только как инструмент предсказания, но и как средство контроля качества метаданных.

В настоящей работе, несмотря на извлечение формульных выражений при обработке документов, модели использовали только текстовые признаки, поэтому вклад формул в качество классификации не оценивался. Дальнейшая работа предполагает включение формульных признаков и оценку их эффекта, поскольку нотация и типовые формульные выражения часто являются предметно-специфичными маркерами разделов математики.

Кроме того, перспективы дальнейших исследований включают расширение коллекции, построение собственных языковых моделей, специфичных для рассматриваемой предметной области, расширение набора методов и ансамблей, более полное сравнение различных источников текста (аннотации и полного текста), а также более глубокий учет иерархии УДК, включая иерархические функции потерь и иерархическое декодирование предсказаний. Эти направления станут продолжением настоящей работы и позволяют перейти от сравнительного исследования к более прикладным библиотечным системам поддержки классификации.

СПИСОК ЛИТЕРАТУРЫ

1. *Tóth E.* Innovative Solutions in Automatic Classification: A Brief Summary // Libri. 2002. Vol. 52, No. 1. P. 48–53. https://doi.org/10.1515/LIBR.2002.48

2. *Romanov A., Lomotin K., Kozlova E.* Automatization of Scientific Articles Classification According to Universal Decimal Classifier // Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017). CEUR Workshop Proceedings. 2017. Vol. 1975. P. 122–133.

3. *Romanov A.Yu., Lomotin K.E., Kozlova E.S., Kolesnichenko A.L.* Research of neural networks application efficiency in automatic scientific articles classification according to UDC // Proceedings of the 2016 International Siberian Conference on Control and Communications (SIBCON 2016), Moscow, Russia, 12–14 May 2016. IEEE, 2016. P. 612–616. https://doi.org/10.1109/SIBCON.2016.7491783

4. *Kragelj M., Kljajić Borštnar M.* Automatic classification of older electronic texts into the Universal Decimal Classification-UDC // Journal of Documentation. 2021. Vol.77, No. 3. P. 755–776. https://doi.org/10.1108/JD-06-2020-0092

5. *Roy A., Ghosh S.* Automated Subject Identification using the Universal Decimal Classification: The ANN Approach // SRELS Journal of Information and Knowledge. 2023. Vol. 60, No. 2. P. 69–76. https://doi.org/10.17821/srels/2023/v60i2/170963

6. *Borovič M., Ojsteršek M., Strnad M.* A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries // IEEE Access. 2022. Vol. 10. P. 85595–85605. https://doi.org/10.1109/ACCESS.2022.3198706

7. *Мамедов В.Ю., Ковалевский Д.А., Морозов Д.А., Столяров С.С., Оспичев С.С.* Иерархическая классификация научных статей при помощи глубокого обучения (на примере иерархии УДК) // Моделирование и анализ информационных систем. 2025. Т. 32. № 1. С. 80–94. https://doi.org/10.18255/1818-1015-2025-1-80-94

8. *Borovič M., Tomovski E., Li Dobnik T., Majninger S.* Evaluating Proprietary and Open-Weight Large Language Models as Universal Decimal Classification Recommender Systems // Applied Sciences. 2025. Vol. 15. No. 14. Art. 7666. https://doi.org/10.3390/app15147666

9. *Silla C.N. Jr., Freitas A.A.* A Survey of Hierarchical Classification across Different Application Domains // Data Mining and Knowledge Discovery. 2011. Vol. 22, No. 1–2. P. 31–72. https://doi.org/10.1007/s10618-010-0175-9

10. *Zangari A., Marcuzzo M., Rizzo M., Giudice L., Albarelli A., Gasparetto A.* Hierarchical Text Classification and Its Foundations: A Review of Current Research // Electronics. 2024. Vol. 13, No. 7. Art. 1199. https://doi.org/10.3390/electronics13071199

11. *Liu R., Liang W., Luo W., Song Y., Zhang H., Xu R., Li Y., Liu M.* Recent Advances in Hierarchical Multi-label Text Classification: A Survey. // 2023. arXiv:2307.16265. https://doi.org/10.48550/arXiv.2307.16265

12. *Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L.E., Brown D.E.* Text Classification Algorithms: A Survey // Information. 2019. Vol. 10, No. 4. Art. 150. https://doi.org/10.3390/info10040150

13. *Li Q., Peng H., Li J., Xia C., Yang R., Sun L., Yu P.S., He L.* A Survey on Text Classification: From Traditional to Deep Learning // ACM Transactions on Intelligent Systems and Technology. 2022. Vol. 13, No. 2. Art. 31. P. 1–41. https://doi.org/10.1145/3495162

14. *Mirończuk M.M., Protasiewicz J.* A Recent Overview of the State-of-the-Art Elements of Text Classification // Expert Systems with Applications. 2018. Vol. 106. P. 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

15. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // 2013. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

16. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. Minneapolis, Minnesota, 2019. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423

17. *Герасименко Н.А., Ватолин А., Янина А., Воронцов К.В.* SciRus: легкий и мощный мультиязычный энкодер для научных текстов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Т. 520, № 2. С. 216–227. https://doi.org/10.1134/S1064562424602178

18. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.* CatBoost: unbiased boosting with categorical features // Advances in Neural Information Processing Systems. 2018. Vol. 31. P. 6638–6648. https://doi.org/10.48550/arXiv.1706.09516

19. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019. P. 2623–2631. https://doi.org/10.1145/3292500.3330701

20. *van der Maaten L., Hinton G.* Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008. Vol. 9, No. 86. P. 2579–2605.

METHODS FOR AUTOMATIC ASSIGNMENT OF UDC CODES TO MATHEMATICAL ARTICLES: AN EVALUATION OF CLASSICAL AND NEURAL APPROACHES

B. T. Gizatullin[1] [0009-0000-6251-9260], O. A. Nevzorova[2] [0000-0001-8116-9446]

[1, 2]*Kazan (Volga Region) Federal University, Kazan, Russia*

[1]gizat.blt@gmail.com, [2]onevzoro@gmail.com

Abstract

Universal Decimal Classification (UDC) is a hierarchical indexing system in which a publication may be assigned one or several codes. Manual UDC indexing is labor-intensive and often inconsistent. This paper addresses the automatic assignment of UDC codes to Russian-language mathematical research articles. The aim is to compare combinations of text representations and classification models on a unified corpus and to identify the most effective configurations. A corpus of 4194 articles was collected from Math-Net.Ru, including full texts, abstracts, metadata, and UDC codes. The preprocessing pipeline comprised PDF text extraction, removal of layout artifacts, and normalization of UDC labels. We compared TF-IDF, Word2Vec, SciRus-tiny, and SciRus-tiny3.5 representations combined with logistic regression, Complement Naive Bayes (CNB), and CatBoost. In both the single-label and multi-label settings, the best performance was achieved by TF-IDF + LogReg, while TF-IDF + CNB showed closely competitive results. The proposed approach can be used in automatic subject indexing systems for digital libraries and scientific archives, in UDC recommendation tools for authors and editors, and in metadata quality control workflows.

Keywords: *automatic classification, Universal Decimal Classification, UDC, scientific text processing, machine learning, hierarchical classification, multi-label classification, mathematical texts, digital libraries, text vectorization.*

REFERENCES

1. *Tóth E.* Innovative Solutions in Automatic Classification: A Brief Summary // Libri. 2002. Vol. 52, No. 1. P. 48–53. https://doi.org/10.1515/LIBR.2002.48

2. *Romanov A., Lomotin K., Kozlova E.* Automatization of Scientific Articles Classification According to Universal Decimal Classifier // Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017). CEUR Workshop Proceedings. 2017. Vol. 1975. P. 122–133.

3. *Romanov A.Yu., Lomotin K.E., Kozlova E.S., Kolesnichenko A.L*. Research of neural networks application efficiency in automatic scientific articles classification according to UDC // Proceedings of the 2016 International Siberian Conference on Control and Communications (SIBCON 2016), Moscow, Russia, 12–14 May 2016. IEEE, 2016. P. 612–616. https://doi.org/10.1109/SIBCON.2016.7491783

4. *Kragelj M., Kljajić Borštnar M*. Automatic classification of older electronic texts into the Universal Decimal Classification-UDC // Journal of Documentation. 2021. Vol. 77, No. 3. P. 755–776. https://doi.org/10.1108/JD-06-2020-0092

5. *Roy A., Ghosh S.* Automated Subject Identification using the Universal Decimal Classification: The ANN Approach // SRELS Journal of Information and Knowledge. 2023. Vol. 60. No. 2. P. 69-76. https://doi.org/10.17821/srels/2023/v60i2/170963

6. *Borovič M., Ojsteršek M., Strnad M.* A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries // IEEE Access. 2022. Vol. 10, P. 85595–85605. https://doi.org/10.1109/ACCESS.2022.3198706

7. *Mamedov V., Kovalevsky D., Morozov D., Stolyarov S., Ospichev S*. Hierarchical classification of scientific articles using deep learning (using the UDC hierarchy as an example) // Modeling and Analysis of Information Systems. 2025. Vol. 32, No. 1. P. 80–94. https://doi.org/10.18255/1818-1015-2025-1-80-94

8. *Borovič M., Tomovski E., Li Dobnik T., Majninger S.* Evaluating Proprietary and Open-Weight Large Language Models as Universal Decimal Classification Recommender Systems // Applied Sciences. 2025. Vol. 15, No. 14. Art. 7666. https://doi.org/10.3390/app15147666

9. *Silla C.N. Jr., Freitas A.A.* A Survey of Hierarchical Classification across Different Application Domains // Data Mining and Knowledge Discovery. 2011. Vol. 22, No. 1–2. P. 31–72. https://doi.org/10.1007/s10618-010-0175-9

10. *Zangari A., Marcuzzo M., Rizzo M., Giudice L., Albarelli A., Gasparetto A*. Hierarchical Text Classification and Its Foundations: A Review of Current Research // Electronics. 2024. Vol. 13, No. 7. Art. 1199. https://doi.org/10.3390/electronics13071199

11. *Liu R., Liang W., Luo W., Song Y., Zhang H., Xu R., Li Y., Liu M.* Recent Advances in Hierarchical Multi-label Text Classification: A Survey // 2023. arXiv:2307.16265. https://doi.org/10.48550/arXiv.2307.16265

12. *Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L.E., Brown D.E.* Text Classification Algorithms: A Survey // Information. 2019. Vol. 10, No. 4. Art. 150. https://doi.org/10.3390/info10040150

13. *Li Q., Peng H., Li J., Xia C., Yang R., Sun L., Yu P.S., He L.* A Survey on Text Classification: From Traditional to Deep Learning // ACM Transactions on Intelligent Systems and Technology. 2022. Vol. 13, No. 2. Art. 31. P. 1–41. https://doi.org/10.1145/3495162

14. *Mirończuk M.M., Protasiewicz J.* A Recent Overview of the State-of-the-Art Elements of Text Classification // Expert Systems with Applications. 2018. Vol. 106. P. 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

15. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // 2013. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

16. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. Minneapolis, Minnesota, 2019. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423

17. *Gerasimenko N., Vatolin A., Ianina A., Vorontsov K.* SciRus: Tiny and Powerful Multilingual Encoder for Scientific Texts // Doklady Mathematics. 2024. Vol. 110, Suppl. 1. P. S193–S202. https://doi.org/10.1134/S1064562424602178

18. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.* CatBoost: unbiased boosting with categorical features // Advances in Neural Information Processing Systems. 2018. Vol. 31. P. 6638–6648. https://doi.org/10.48550/arXiv.1706.09516

19. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019. P. 2623–2631. https://doi.org/10.1145/3292500.3330701

20. *van der Maaten L., Hinton G.* Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008. Vol. 9, No. 86. P. 2579–2605.

СВЕДЕНИЯ ОБ АВТОРАХ

ГИЗАТУЛЛИН Булат Тимурович – магистрант Института математики и механики им. Н. И. Лобачевского Казанского (Приволжского) федерального университета, направление подготовки «Математика и компьютерные науки», профиль «Статистические методы науки о данных».

Bulat Timurovich GIZATULLIN – master's student at the Lobachevsky Institute of Mathematics and Mechanics, Kazan (Volga Region) Federal University, program in Mathematics and Computer Science, track "Statistical Methods in Data Science".

email: gizat.blt@gmail.com

ORCID: 0009-0000-6251-9260

НЕВЗОРОВА Ольга Авенировна – кандидат технических наук, доцент Казанского (Приволжского) федерального университета.

Olga Avenirovna Nevzorova – Candidate of Technical Sciences (PhD equivalent), associate Professor at Kazan (Volga Region) Federal University.

email: onevzoro@gmail.com

ORCID: 0000-0001-8116-9446

Материал поступил в редакцию 14 апреля 2026 года