УДК 81’23
DOI 10.47388/2072-3490/lunn2026-73-1-29-44

ПРОБЛЕМА СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕРМИНОВ БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ: ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ В ПРЕДМЕТНОЙ ОБЛАСТИ «ЦИФРОВОЕ ПРАВО»

Д. И. Галюченко

Одинцовский филиал Московского государственного института международных отношений (университета) Министерства иностранных дел Российской Федерации, Одинцово, Россия

Создание терминологических баз и поддержание их в актуальном состоянии в областях с высокой степенью динамичности, семантической нестабильностью и междисциплинарностью является значительным испытанием для современных терминоведов и лексикографов, т. к. классические методы сбора, описания и обработки терминов не соответствуют текущим вызовам и потребностям.

Оценка особенностей и возможностей применения модели генеративного предобученного трансформера в контексте автоматизации лингвистических исследований позволит сделать работу с данными терминами более эффективной и управляемой.

Цель исследования заключается в оценке эффективности генеративного предобученного трансформера (GPT) на примере языковой модели DeepSeek в решении задач по автоматическому извлечению и семантической классификации терминов на материале текстов в области цифрового права, выявлении перспективности данного направления исследования и поиска возможных путей развития данной области.

В рамках исследования выдвигается гипотеза что генеративные модели благодаря особенностям своего обучения покажут высокую полноту при извлечении терминов-кандидатов, но столкнутся с трудностями на этапе семантической классификации.

Для этих целей было сделано следующее: сформирована методика эксперимента по извлечению и классификации терминов с помощью GPT, собран корпус текстов и эталонный список терминов, относящихся к семантическому полю «цифровое право», проведен эксперимент, в результате которого выполнен анализ и дана оценка эффективности работы по результатам рассчитанных метрик полноты и точности для целей классификации терминов, предложены направления дальнейшего развития данной методики.

Данное исследование подтверждает выдвинутую гипотезу и показывает высокий потенциал стандартных языковых моделей для задач терминологической работы, который полностью может быть реализован при дальнейшей точной настройке модели и ее тренировке на решение конкретных задач.

Ключевые слова: NLP; GPT; лингвистика; компьютерная лингвистика; лексикография; компьютерная лексикография; терминология; цифровое право.

Цитирование: Галюченко Д. И. Проблема семантической классификации терминов большими языковыми моделями: экспериментальное исследование в предметной области «цифровое право» // Вестник Нижегородского государственного лингвистического университета им. Н. А. Добролюбова. 2026. Вып. 1 (73). С. 29–44. DOI 10.47388/2072-3490/lunn2026-73-1-29-44.

Semantic Classification of Terms by Large Language Models: An Experimental Study in the Subject Area “Digital Law”

Danil I. Galyuchenko

Odintsovo Branch of the Moscow State Institute of International Relations (University) of the Ministry of Foreign Affairs of the Russian Federation, Odintsovo, Russia

Creating terminological databases and maintaining their currency in highly dynamic, semantically unstable, and interdisciplinary domains presents a significant challenge for modern terminologists and lexicographers, as classical methods of collecting, describing, and processing terms do not meet current challenges and needs.

Evaluating the features and possibilities of applying the generative pre-trained transformer model in the context of automating linguistic research will make working with these terms more efficient and manageable.

The aim of the study is to assess the effectiveness of the generative pre-trained transformer (GPT) using the DeepSeek language model as an example in solving the tasks of automatic extraction and semantic classification of terms based on texts in the field of digital law, to identify the prospects of this research direction, and to explore possible avenues for the development of this field.

The study hypothesizes that generative models, due to the specific features of their training, will demonstrate high recall in extracting candidate terms but will encounter difficulties at the stage of semantic classification.

To this end, the following was undertaken: an experimental methodology for extracting and classifying terms using GPT was developed; a corpus of texts and a reference list of terms belonging to the semantic field of “digital law” were compiled; an experiment was conducted, based on the results of which an analysis was performed and the effectiveness of the work was evaluated according to the calculated precision and recall metrics for the purpose of term classification; and directions for further development of this methodology were proposed.

This study confirms the proposed hypothesis and demonstrates the high potential of standard language models for terminological work tasks — potential that can be fully realized through further fine-tuning of the model and its training to address specific tasks.

Key words: NLP; GPT; linguistics; computational linguistics; lexicography; computer lexicography; terminology; digital law.

Citation: Galyuchenko, Danil I. (2026) Semantic Classification of Terms by Large Language Models: An Experimental Study in the Subject Area “Digital Law”. *LUNN Bulletin*, 1 (73), 29–44. DOI 10.47388/2072-3490/lunn2026-73-1-29-44.

1. Введение

Современная эпоха цифровой трансформации фундаментально меняет не только технологии, но и социальные институты, включая право. Возникновение и стремительное развитие области цифрового права (*Digital Law*) сопровождается формированием нового понятийного аппарата. Эта терминология характеризуется высокой динамичностью, междисциплинарностью (заимствуя лексику из IT, экономики и юриспруденции) и семантической нестабильностью, когда общеупотребительные слова приобретают узкоспециализированные значения (например, *consideration* как ‘встречное удовлетворение по договору’).

Это порождает фундаментальную проблему для терминоведения и лексикографии: традиционные методы ручного сбора, описания и кодификации терминов не справляются с темпами их появления и изменения (Fuertes-Olivera, Tarp 2014; Kosem, et al. 2014; Карпова 2018; Alguliyev, Gurbanova 2018). Создание и поддержание в актуальном состоянии терминологических баз данных и словарей для таких бурно развивающихся областей становится чрезвычайно трудоемкой задачей, что обусловливает высокую актуальность поиска эффективных методов автоматической обработки специализированной лексики.

В современном мире терминология играет важную роль в представлении и передаче знаний в таких профессиональных областях, как право, медицина, техника и информационные технологии, особенно когда речь идет о письменных текстах. В частности, в связи со стремительным развитием областей и соответствующей терминологии по самому разному ряду причин имеются трудности с пониманием, сбором и обработкой новейшей терминологии. Также развитие технологий обработки текстов на естественном языке создало потребность в практическом применении эффективных методов автоматического извлечения и обработки специализированных терминов.

Ответом на этот вызов стало развитие методов автоматической обработки текстов на естественном языке (*NLP*), в частности инструментов корпусной лингвистики. Платформы, такие как *SketchEngine*, *English Corpora* и программное обеспечение *AntConc*, предоставляют исследователям мощные инструменты для работы с большими текстовыми массивами, позволяя извлекать частотные списки слов, ключевые слова и устойчивые словосочетания (n-граммы, кластеры) (Маник 2019; Гацук 2021; Палийчук 2022; Davies 2023). Эти подходы, основанные на статистических и дистрибутивных методах, доказали свою эффективность в идентификации потенциальных терминологических единиц.

Однако их основной недостаток заключается в ограниченной способности к глубокому семантическому анализу. Они успешно выявляют кандидатов в термины, но задача отличить истинный термин от общеупотребительного слова в специфическом контексте (полисемия) или разграничить близкие по значению понятия остается нерешенной (Bowker, Pearson 2002).

Особое место в этом ряду занимают большие языковые модели (БЯМ, *LLM*), в частности генеративные предобученные трансформеры (*GPT*). В отличие от предыдущих архитектур, они способны улавливать сложные контекстуальные и семантические связи, что открывает новые перспективы для автоматической терминологической работы (Shao, et al. 2017; Маник 2024). Исследования показывают их потенциал для семантического анализа и извлечения информации (Сидорова, Иванов, Овчинникова 2025). Современные GPT-модели используются для извлечения терминов и семантического анализа, но в контексте узкой, относительно новой области (цифровое право) такие

исследования менее распространены. В ходе исследования проверяется следующая гипотеза: предполагается, что генеративные модели, благодаря своей способности улавливать синтаксические и коллокационные паттерны, покажут высокую полноту при извлечении терминов-кандидатов. Однако они столкнутся со значительными трудностями на этапе семантической классификации, что приведет к низкой точности из-за неспособности надежно разграничивать близкие понятия и отличать узкоспециальное употребление слова от общего без дополнительной настройки или специального обучения (Маник 2023).

Из вышесказанного следует, что научная проблема исследования заключается в недостаточной изученности надежности и точности современных генеративных языковых моделей при решении двухэтапной задачи: 1) извлечения терминологических единиц из узкоспециализированного текста и 2) их последующей семантической классификации (отнесения к конкретному семантическому полю). Основная сложность состоит в способности модели разграничивать истинные термины и квазитерминологические единицы или омонимичные лексемы из смежных областей.

2. Характеристика материала и методов исследования

Цель работы заключается в исследовании и оценке эффективности стандартной генеративной предобученной модели (на примере *DeepSeek*) в решении задачи автоматического извлечения и семантической классификации терминологии. Исследование выполнено на материале текстов в области цифрового права.

Для достижения цели были поставлены следующие задачи:

1. Сформировать корпус текстов и эталонный список терминов, относящихся к семантическому полю «цифровое право».

2. Разработать методику эксперимента по извлечению и классификации терминов с помощью *LLM*.

3. Провести эксперимент и рассчитать метрики полноты (recall) для задачи извлечения и точности (precision) для задачи классификации.

4. Проанализировать типичные ошибки модели и определить потенциальные причины низкой эффективности на этапе классификации.

5. Предложить направления для дальнейшего совершенствования методики.

Объект исследования — лексические единицы, функционирующие в сфере цифрового права (*Digital Law*).

Предмет исследования — способы автоматического распознавания, извлечения и классификации специальных терминов при помощи больших языковых моделей.

3. Результаты исследования и их обсуждение

3.1. Специализированная лексики vs термины

Приступая к анализу автоматического извлечения понятий из текстов по цифровому праву, необходимо прежде всего четко разграничить два пусть и взаимосвязанных ключевых понятия: специализированная лексика и терминология. Несмотря на то, что в обыденном употреблении они могут восприниматься как синонимы, в терминоведении их принято дифференцировать, что имеет принципиальное значение для постановки задач машинной обработки текста.

Специализированная лексика, или лексика для специальных целей (*Language for Special Purposes*, *LSP*), представляет собой наиболее широкую категорию. Она охватывает всю совокупность лексических единиц, используемых представителями определенной профессиональной или научной сферы для обеспечения точной и недвусмысленной коммуникации (Cabre 1999). В состав этой лексики входят не только строго определенные термины, но и профессионализмы (профессиональный жаргон), номенклатурные наименования, а также общеупотребительные слова, которые в данном контексте приобретают специфическое значение. Таким образом, возможно обобщить, что специализированная лексика — это весь лексический пласт, характеризующий профессиональный дискурс в той или иной области.

Терминология в свою очередь является ядром, наиболее упорядоченной и кодифицированной частью специализированной лексики. Термин — это не просто специальное слово, а лексическая единица, отвечающая ряду строгих критериев. Согласно подходу отечественной школы терминоведения, представленному в работах С. В. Гринёва-Гриневича, термин должен обладать следующими неотъемлемыми свойствами: системность (термин является элементом целостной системы понятий в данной области знания, и его значение полностью определяется его местом в этой системе); точность и однозначность (моносемичность) (в пределах своего терминологического поля идеальный термин должен иметь только одно значение, что исключает двусмысленность); дефинированность (термин обладает или должен обладать строгим научным определением); стилистическая нейтральность; устойчивость и воспроизводимость (Гринёв-Гриневич, Сорокина, Молчанова 2023).

Между строго кодифицированными терминами и остальной частью специализированной лексики существует «серая зона», в которую входят так называемые терминоиды, или прототермины. Это слова и словосочетания, которые уже используются в специальной литературе, но еще не получили окончательного статуса термина: они могут быть многозначными, не иметь строгой дефиниции или конкурировать с другими вариантами (Хохлова, Еникеева 2020). Область цифрового права, как молодая и бурно развивающаяся, особенно богата такими единицами.

Например, лексическая единица *consideration* в своем общеупотребительном значении может переводиться как «рассмотрение», в области договорного права имеет иное значение, является юридическим термином и переводится как «встречное удовлетворение по договору», также в рамках деловых отношений может иметь перевод «вознаграждение», что вызывает значительные трудности с адекватным восприятием подобных лексических единиц, которые могут в разных случаях являться специализированными терминами.

Актуальность данного разграничения для настоящего исследования является критической: современные большие языковые модели, обученные на гигантских массивах текстов, превосходно улавливают статистические закономерности и контекстуальные связи. Это позволяет им с высокой полнотой извлекать специализированную лексику в широком смысле — т. е. все слова, которые часто встречаются в данном профессиональном дискурсе. Однако основная проблема, исследуемая в данной работе, заключается в их способности отличить истинный термин от остальных элементов этого лексического пласта. Именно на этом этапе, требующем не только статистического, но и, по сути, экспертного знания о системных связях и дефинициях, модели могут давать сбой, что и предполагается проверить в ходе эксперимента.

3.2. Проблемы извлечения и обработки специализированной лексики

Несмотря на предъявляемые требования к инвариантности термина, основной проблемой извлечения специализированных терминов является их вариативность. Термины могут иметь различные формы — сокращения, аббревиатуры, синонимы или другие вариации, — что усложняет их автоматическое обнаружение. Существование нескольких разных слов или словосочетаний для обозначения одного и того же понятия также представляет определенную сложность. Особенно часто это встречается в формирующихся областях, где терминологический стандарт еще не устоялся. Например, понятие права, регулирующего отношения в сети Интернет и цифровом пространстве, может именоваться как *Internet Law*, *Cyber Law* и в более широком смысле как *Digital Law*. Хотя между указанными вариантами есть семантические нюансы, во многих контекстах они используются как взаимозаменяемые, что представляет вызов для автоматической классификации. Вместе с тем известны случаи использования разных частей речи или словообразовательных моделей для выражения схожих понятий. Например, процесс внедрения цифровых технологий может обозначаться терминами *Digitalization* (оцифровка, перевод в цифровую форму) и *Digital Transformation* (более глубокое, системное изменение бизнес-процессов). Модель должна не просто извлечь оба словосочетания, но и понять их семантическую близость и иерархию. Возможно изменение порядка слов или структуры словосочетания без существенного изменения смысла. Так, понятие «право на цифровую

собственность» может быть выражено как *Digital Property Rights* или, в инверсированном виде с предлогом, как *Rights to Digital Property*. В текстах по цифровому праву возможно использование кратких форм наряду с полными, что является стандартом для любой технической и юридической литературы. Например, термин *Artificial Intelligence* повсеместно используется наряду с аббревиатурой *AI*. Аналогично *General Data Protection Regulation* почти всегда соседствует с аббревиатурой *GDPR*. Система автоматического извлечения должна уметь сопоставлять эти формы и понимать, что они относятся к одному и тому же объекту.

Также проблемой является различение специализированных терминов от общеупотребительных слов. Полисемия, при которой одно и то же слово может иметь несколько значений в зависимости от контекста, добавляет сложности. Согласно исследованиям Bowker и Pearson, важной проблемой в терминологии является наличие терминов, которые могут быть многозначными, и необходимо учитывать контекст для их корректной интерпретации. Авторы подчеркивают, что понимание терминов зависит от контекста и для правильного извлечения терминов требуется учитывать специфику каждого конкретного корпуса текстов, а также использовать методы дезамбигуации, чтобы установить правильное значение в конкретном контексте (Bowker, Pearson 2002).

Наконец, непрерывное развитие специализированной лексики требует частого обновления репозиториев терминов. Традиционные методы лексикографии не справляются с такой стремительной динамикой, что требует автоматизированных методов обновления.

Таким образом, для эффективной автоматической обработки текстов по цифровому праву недостаточно просто искать заранее заданный список канонических терминов. Система должна обладать способностью распознавать и группировать (нормализовывать) все эти варианты как относящиеся к одному понятию. В противном случае возникает риск неполноты извлечения данных (пропуск важных упоминаний) и некорректной статистической оценки значимости того или иного понятия в тексте. Учет вариативности является одним из ключевых требований к современным системам терминологического анализа.

3.3. Подходы к автоматическому извлечению и обработке

Задача автоматического извлечения терминологии (*Automatic Term Extraction, ATE*) является одной из классических в области обработки естественного языка (*NLP*). За десятилетия исследований был разработан ряд подходов, которые можно условно разделить на четыре основные группы: статистические, лингвистические, гибридные и подходы на основе машинного обучения.

Исторически первые и наиболее базовые подходы основаны на предположении, что термины обладают особыми статистическими

свойствами. Ключевая идея заключается в том, что терминологические единицы встречаются в специализированном корпусе текстов значительно чаще, чем в общеупотребительной речи (принцип *termhood*). Для их выявления используются такие метрики, как: частотность и *TF-IDF* (*Term Frequency-Inverse Document Frequency*), позволяющие выделить слова, которые являются важными для конкретного документа в коллекции; меры ассоциативности (коллокации), предполагающие применение метрик, оценивающих силу связи между словами, например взаимная информация (*Mutual Information*) и *T-score*. Они помогают определить, является ли словосочетание устойчивым (например, *цифровое право*) или случайным. Несмотря на то, что эти методы просты в реализации, они часто извлекают много «шума» (нерелевантных словосочетаний) и не учитывают семантическую и грамматическую структуру языка.

Лингвистические подходы опираются на знание о грамматической структуре терминов. В большинстве языков термины представляют собой существительные или именные группы (*Noun Phrases*), построенные по определенным синтаксическим шаблонам. Например, «Прилагательное + Существительное» (*Digital law*) или «Существительное + Существительное» (*Data protection*). Этот подход, как показано в работе *Multilingual term extraction from parallel corpora — A methodology for the automatic extraction of verbal structures and their translation equivalents* (Varadi, Héja 2011), позволяет получать более грамматически корректные термины-кандидаты. Однако он требует создания сложных правил для каждого языка и предметной области и может упускать термины с нетипичной структурой.

Стремясь объединить преимущества двух предыдущих методов, гибридные подходы используют статистику для первоначального отбора кандидатов, а затем применяют лингвистические фильтры для удаления нерелевантных результатов (или наоборот). Такой синергетический подход позволяет повысить точность извлечения. Исследования (Хохлова, Еникеева 2020) показывают, что комбинация мер ассоциативности и последующей обработки с помощью методов машинного обучения (которые по сути являются развитием гибридных идей) дает более качественные результаты при выделении коллокаций.

Наиболее современными и распространенными методами использования машинного обучения для автоматической обработки текста являются методы, основанные на мерах ассоциативности и машинном обучении (Хохлова, Еникеева 2020); методы, основанные на семантической близости и синтаксической связности (Seretan 2010); методы, основанные на нейронных сетях и векторных представлениях слов (Shao, et al. 2017).

Методы машинного обучения все чаще применяются для извлечения и обработки терминов. Модели машинного обучения используются для классификации сегментов текста как терминов или общих слов. Например,

условные случайные поля и машины опорных векторов применяются для последовательной разметки текста, используя такие признаки, как морфология слов, контекст и часть речи, для определения специализированных терминов.

В последнее десятилетие обработка *NLP* (*Natural Language Processing* — обработка текстов на естественном языке) приобрела еще большую популярность благодаря успехам в области глубокого обучения и появлению трансформеров. Модели, такие как *BERT* (*Bidirectional Encoder Representations from Transformers*), *GPT* (*Generative Pre-trained Transformer*) и их последующие версии, сделали большой скачок в обработке естественного языка, позволяя обрабатывать тексты с высокой степенью семантической связи.

3.4. Оценка эффективности методов извлечения терминов и их семантической классификации большими языковыми моделями

Оценка эффективности методов извлечения терминов является важной частью исследований в данной области. Основные метрики оценки включают точность, полноту и F-меру (Manning, Raghavan, Schütze 2008). Точность измеряет долю корректных терминов среди всех извлеченных, а полнота — долю извлеченных терминов среди всех имеющихся в тексте, F-мера же представляет собой гармоническое среднее между точностью и полнотой. Баланс между точностью и полнотой остается сложной задачей, особенно в быстро развивающихся областях, где язык постоянно меняется (Баженов 2012).

Точность (*Precision*) высчитывается по формуле $P = \frac{TP}{TP + FP}$

Полнота (*Recall)* высчитывается по формуле $R = \frac{TP}{TP + FN}$

F-мера (точность и полнота) высчитывается по формуле $F1 = 2\frac{P \times R}{P+R}$

TP — истинно положительное решение; *TN* — истинно отрицательное решение; *FP* — ложно положительное решение; *FN* — ложно отрицательное решение.

Обычно используется гибридная оценка, которая сочетает автоматическую проверку с экспертизой специалистов. Эксперты оценивают, насколько извлеченные термины соответствуют спецификам области, что помогает в оптимизации алгоритмов. Стоит отметить, что оценка эффективности различных методов обработки текстов на естественном языке происходит на широкой выборке в рамках одной поставленной задачи. Сначала необходимо поставить данную задачу и оценить эффективность ее выполнения.

Представляется целесообразным поставить эксперимент, как современные трансформерные архитектуры справятся с задачей адекватного извлечения специальной лексики и отнесения конкретных терминов к заданному семантическому полю. Для изучения семантики слова и обнаружения целостных образований, объективно выделенных из массива лексического состава языка как системы, используется когнитивно-

прагматический подход и выделение семантических полей системы понятий (Бочарова 2012).

В юридической терминологии (цифровое право является частью юридической терминологии), по мнению автора, термин будет считаться специальным, если он имеет четкое определение в нормативно-правовых актах или правовой доктрине; несет специфическое значение, используемое именно в юридическом контексте; употребляется в научных или официальных источниках по цифровому праву. Так, например, слово *download* может встречаться в контексте интернет-активности, но не всегда будет считаться специализированным термином именно цифрового права.

В качестве материала исследований возьмем следующую выборку понятий, которые относятся к семантическому полю «цифровое право», из списка терминов, составленного Е. А. Прониной и А. С. Галюченко в материалах, находящихся в разработке, для дисциплины «Иностранный язык (профессиональный)» для студентов первого курса магистратуры: *Digital Law*, *Internet Law*, *Cyber Law*, *Digital platform*, *Digitalization*, *Plagiarism*, *Copyright*, *Copyright Infringement*, *Intellectual Property*, *Cybersecurity*, *Piracy*, *Cyber bullying*, *Digital rights*, *Digital works*, *Smart contract.*

Эксперимент будет проводиться на небольшом корпусе текстов, в котором встречается данная лексика, а именно на текстах из тех же методических материалов объемом 8864 слова. Сначала модели будет представлено определение специальной лексики из публикации Гринева-Гриневича (Гринёв-Гриневич, Сорокина, Молчанова 2022), а затем будет дан запрос извлечь из корпуса всю специальную лексику и определить, какая относится к семантическому полю цифрового права.

В качестве модели генеративного предобученного трансформера будет использоваться разработка *DeepSeek*, владельцем которой являются Hangzhou DeepSeek Artificial Intelligence Co., Ltd. и Beijing DeepSeek Artificial Intelligence Co., Ltd. На момент проведения эксперимента была использована модель *DeepSeek V3* с активной функцией *Reasoning*, которая позволяет модели обосновывать запросы, повышая общее качество ответов.

В модель была загружена публикация «ЕЩЕ РАЗ К ВОПРОСУ ОБ ОПРЕДЕЛЕНИИ ТЕРМИНА» (Гринёв-Гриневич, Сорокина, Молчанова 2022) и отправлен запрос «Перечисли признаки термина из публикации». Автор убедился, что модель точно перечислила все признаки, а именно: специальная лексема, называет общее понятие, дефинированность, точность значения, однозначность, контекстуальная независимость, устойчивость формы и воспроизводимость в речи, номинативность, стилистическая нейтральность.

Далее в модель был загружен один из ранее упомянутых материалов для дисциплины «Иностранный язык (профессиональный)» для студентов первого курса магистратуры и передан запрос «Теперь извлеки всю специальную

лексику из предложенного текста и затем определи, какая относится к семантическому полю цифрового права (Digital Law)».

Всего было найдено 42 лексические единицы, которые, по мнению модели, соответствуют критерию специальной лексики: *Digital Law*; *Internet Law / Cyber Law*; *Digital Technologies*; *Digital Environment*; *Digital Economy*; *Digital Platform*; *Plagiarism*; *Copyright*; *Copyright Infringement*; *Intellectual Property*; *Cybersecurity*; *Piracy*; *Digital Rights*; *Encroachment*; *Cyberbullying*; *Legal Compliance*; *Illegal File Sharing*; *Hacking*; *Identity Theft*; *Software*; *Data Protection*; *Personal Data*; *Data Privacy*; *Outsourcing*; *Cloud Services*; *Blockchain*; *Artificial Intelligence (AI)*; *Cryptocurrency*; *Digital Money / Digital Currency*; *Smart Contract*; *Marketplace*; *Regulatory Legal Acts*; *Regulatory Gaps*; *Hybrid Regulation Model*; *Digital Public Administration*; *G20*; *OECD*; *FATF*; *EAEU*; *Digital Property Rights*; *Personal Non-Property Digital Rights*.

Из них модель выделила 20 лексических единиц, которые, по ее мнению, относятся к семантическому полю цифрового права. А именно: *Digital Law*, *Cybersecurity*, *Intellectual Property*, *Data Privacy*, *Digital Rights*, *Copyright Infringement*, *Plagiarism*, *Cyberbullying*, *Hacking*, *Blockchain*, *Smart Contract*, *Cryptocurrency*, *Cloud Services*, *Legal Compliance*, *Regulatory Gaps*, *Digital Economy*, *Data Protection*, *Digital Public Administration*, *Digital Property Rights*, *Personal Non-Property Digital Rights*.

Дальнейший анализ результатов показывает, что из данных лексических единиц, составляющих эталонную выборку для семантического поля «цифровое право», модель смогла правильно классифицировать лишь шесть. К ним относятся: *Digital Law*, *Plagiarism*, *Copyright Infringement*, *Intellectual Property*, *Cybersecurity* и *Cyber bullying*. Остальные термины из эталонного списка не были отнесены моделью к указанному семантическому полю.

Однако среди общего списка специальной терминологии, отобранной моделью, можно еще обнаружить *Internet Law*, *Cyber Law*, *Digital platform*, *Copyright*, *Piracy*, *Digital rights*, *Smart contract*, которые также присутствуют в выборке.

В результате модель не смогла обнаружить и извлечь лишь два термина — *Digital works* и *Digitalization*. Шесть терминов было корректно отнесено к семантическому полю цифрового права, а всего было правильно найдено 13 терминов из 15 предложенных в выборке.

Теперь подробно проанализируем, что произошло. Считается целесообразным разделить оценку на два независимых этапа: извлечение терминов и классификация терминов.

Оценим этап извлечения терминов, насколько хорошо модель находит термины в корпусе текстов по отношению к заданной выборке. Рассчитаем полноту извлечения. Исходная выборка содержала 15 терминов. Модель извлекла 13 терминов, которые присутствуют в выборке. Два термина, присутствующих в выборке, модель не извлекла.

$$R = \frac{13}{15} \approx 0{,}867$$

Таким образом, полнота извлечения равна 86,7 %. Модель хорошо справилась с извлечением терминов в данном конкретном примере. Среди возможных причин пропуска двух терминов можно выделить следующее: данные термины недостаточно часто встречаются в корпусе; контекст их употребления не был распознан моделью, на сложность чего автор обратил внимание ранее в контексте юридической терминологии.

Далее оценим этап классификации извлеченных терминов, насколько точно модель определяет, какие извлеченные термины относятся к семантическому полю цифрового права. Рассчитаем точность классификации. Среди терминов, которые присутствуют в исходной выборке, модель смогла извлечь 13 терминов. Среди всех классифицированных терминов модель верно классифицировала шесть терминов, семь терминов классифицированы не были.

$$P = \frac{6}{13} \approx 0{,}462$$

Таким образом, точность классификации составляет 46,2 %. Модель верно классифицировала менее половины извлеченных терминов из выборки. Остальные термины (53,8 %) были извлечены, но не отнесены к семантическому полю цифрового права. Среди возможных причин недостаточной эффективности классификации терминов можно выделить следующее: модель не относит термины к семантическому полю цифрового права, даже если они были корректно извлечены; возможно, модели не хватает контекстных признаков для корректной классификации терминов.

Данный эксперимент демонстрирует высокую полноту извлечения терминов и недостаточно высокую точность классификации терминов стандартной моделью генеративного предобученного трансформера, которая не была специально тренирована для выполнения данных задач, что свидетельствует о высоком потенциале модели, если повысить точность классификации терминов. Целесообразно заключить, что модель покажет более высокую эффективность при более тонкой настройке, если выполнить следующее:

1) специально обучить модель на выполнение подобного рода задач на аннотированных данных, точно сообщая модели все TP-, TN,- FP-, FN-результаты;

2) расширить корпус текстов, которые будет обрабатывать модель, чтобы она могла распознать больше вариантных написаний терминов (*Cyber Law — Cyberspace Law*) и учитывала больше контекстных паттернов (например, *Digitalization of law*);

3) задать настройку на большее обращение внимания на контекстный анализ (например, если термин *Digital works* встречается рядом с *Digital rights* или *Compliance* в контексте одного текста, относить его к семантическому полю цифрового права);

4) ввести порог уверенности классификации для отсечения случайных совпадений.

4. Заключение

Автоматическое извлечение и обработка терминов из текстов со специальной лексикой является важным направлением исследований в лингвистике, информационных науках и обработке естественного языка. Эти методы позволяют эффективно управлять специализированными знаниями, улучшать качество перевода и поддерживать информационный поиск. Вариативность терминологии, полисемия и постоянное обновление специализированных словарей создают определенные трудности. Тем не менее внедрение статистических, гибридных и основанных на правилах алгоритмов, а также технологий машинного обучения позволяет значительно повысить эффективность работы со специализированной лексикой в рамках систем управления терминологией (*TMS*) на уровнях извлечения терминов-кандидатов, их классификации и стандартизации.

Эксперимент, проведенный с использованием модели *DeepSeek*, продемонстрировал результаты, полностью подтверждающие выдвинутую гипотезу. Модель показала высокую полноту извлечения (86,7 %), успешно идентифицировав большинство терминологических единиц из эталонного списка в предложенном корпусе. Это свидетельствует о том, что современные *LLM*, обученные на обширных данных, эффективно улавливают лексико-синтаксические паттерны, характерные для специализированной лексики. Однако на этапе семантической классификации — отнесения извлеченных терминов к конкретному семантическому полю «цифровое право» — модель продемонстрировала значительно более низкую точность (46,2 %). Она не смогла корректно классифицировать более половины релевантных терминов, которые сама же успешно извлекла. Этот результат указывает на фундаментальную проблему: без специальной настройки или дополнительных контекстных указаний стандартные БЯМ испытывают трудности с тонкой семантической дифференциацией и разграничением смежных понятий. Модель успешно распознает «особость» слова в дискурсе, но не всегда может точно определить его принадлежность к узкому семантическому классу.

Таким образом, данное исследование вносит вклад в понимание как возможностей, так и ограничений применения «коробочных» БЯМ для задач терминологической работы. Основной вывод заключается в том, что высокий потенциал этих моделей может быть реализован лишь при целенаправленной адаптации к конкретной задаче.

Повышение точности и эффективности автоматического извлечения и классификации терминов позволит создать инструменты, способные в реальном времени поддерживать актуальность терминологических баз данных, что

сделает специализированные знания более доступными и управляемыми в условиях информационной перегрузки.

Дальнейшие исследования должны быть направлены на улучшение существующих подходов к автоматическому извлечению и классификации терминологии, особенно в контексте быстро развивающихся областей, таких как цифровое право. Повышение точности и эффективности автоматического извлечения терминов позволит сделать специализированные знания более доступными и управляемыми в условиях увеличения объемов данных.

Список литературы / References

Баженов А. К. Оценка классификатора (точность, полнота, F-мера) [Электронный ресурс] // Bazhenov.me. 21.07.2012. URL: https://www.bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (дата обращения: 31.01.2025). [Bazhenov, Aleksey K. (2012) (2025, January 31) Otsenka klassifikatora (tochnost’, polnota, F-mera) (Evaluation of Classifier (Accuracy, Completeness, F-score) // Bazhenov.me. Retrieved from https://www.bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (In Russian)].

Бочарова М. А. Семантическое поле как способ системного описания лексики // Вестник РУДН. Серия: Вопросы образования: языки и специальность. 2012. № 4. С. 63–67. [Bocharova, Maria A. (2012) Semanticheskoe pole kak sposob sistemnogo opisaniya leksiki (Semantic Field as a Method of Systemic Description of Vocabulary). *Vestnik RUDN. Seriya: Voprosy obrazovaniya: yazyki i spetsial’nost’*, 4, 63–67. (In Russian)].

Гацук Е. Ю. Sketch Engine как инструмент для выявления специальных номинаций (на примере англоязычных текстов предметной области «Языковая политика») // Язык и межкультурная коммуникация: современные векторы развития: Сборник научных статей по материалам II Международной научно-практической конференции. Пинск: Полесский государственный университет, 2021. № 2. С. 89–96. [Gatsuk, Ekaterina Yu. (2021) Sketch Engine kak instrument dlya vyyavleniya spetsial’nykh nominatsiy (na primere angloyazychnyh tekstov predmetnoj oblasti «Yazykovaya politika») (Sketch Engine as a Tool for Extracting Special Nominations in English-language Texts of the “Language Policy” Subject Area). In *Yazyk i mezhkul’turnaya kommunikatsiya* (Language and Intercultrual Communication), 2, 89–96. (In Russian)].

Гринев-Гриневич С. В., Сорокина Э. А., Молчанова М. А. Еще раз к вопросу об определении термина // Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. 2022. № 3. С. 710–729. [Grinev-Grinevich, Sergey V., Sorokina, Elvira A., & Molchanova, Maria A. (2022) Esche raz k voprosu ob opredelenii termina (Once Again on the Definition of a Term). *RUDN Journal of Language Studies, Semiotics and Semantics*, 3, 710–729. (In Russian)].

Гринев-Гриневич С. В., Сорокина Э. А., Молчанова М. А. Терминоведение. Москва: Наука, 2023. [Grinev-Grinevich, Sergey V., Sorokina, Elvira A., & Molchanova, Maria A. (2023) *Terminovedenie* (Terminology Studies). Moscow: Nauka. (In Russian)].

Карпова О. М. Новые вызовы современной английской лексикографии // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. 2018. № 3. С. 24–28. [Karpova, Olga M. (2018) Novye vyzovy sovremennoy angliyskoy leksikografii (New Challenges of Contemporary English Lexicography). *Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication*, 3, 24–28. (In Russian)].

Маник С. А. Параллельный корпус в практике перевода общественно-политических текстов (с английского на русский) // Современные исследования социальных проблем. 2019. Т. 11. № 4. С. 225–245. [Manik, Svetlana A. (2019) Parallel'nyy korpus v praktike perevoda obshchestvenno-politicheskikh tekstov (Parallel Corpus in Translating Socio-Political Texts). *Modern Studies of Social Issues*, 11 (4), 225–245. (In Russian)].

Маник С. А. Современная лексикография в эпоху Chat GPT // Научно-исследовательская деятельность в классическом университете – 2023: традиции и инновации: Материалы Международного научно-практического фестиваля, Иваново, 10–28 апреля 2023 г. Иваново: Ивановский государственный университет, 2023. С. 519–524. [Manik, Svetlana A. (2023) Sovremennaya leksikografiya v epokhu Chat GPT (Contemporary Lexicography in the Era of Chat GPT). In *Naučno-issledovatel'skaya deyatel'nost' v klassicheskom universitete – 2023: traditsii i innovatsii: Materialy Mezhdunarodnogo nauchno-prakticheskogo festivalya, Ivanovo, 10–28 aprelya 2023 g.* (Science and Research Activities in a Classical University – 2023: Traditions and Innovations: Proceedings of the International Scientific and Practical Festival, Ivanovo, April 10–28. 2023). Ivanovo: Ivanovo State University, 519–524. (In Russian)].

Маник С. А. Корпусные исследования в эпоху нейронных сетей [Электронный ресурс] // Современная филологическая наука: достижения и инновации: Сборник материалов Международного симпозиума, Иваново, 23–25 мая 2024 г. Иваново: Ивановский государственный университет, 2024. С. 304–308. [Manik, Svetlana A. (2024) Korpusnye issledovaniya v epokhu neironnykh setey (Corpus Research in the Era of Neural Networks). In *Sovremennaya filologicheskaya nauka: dostizheniya i innovatsii: Sbornik materialov Mezhdunarodnogo simpoziuma, Ivanovo, 23–25 maya 2024 g.* (Modern Philology: Achievements and Innovations: Proceedings of the International Symposium, Ivanovo, May 23–25, 2024). Ivanovo: Ivanovo State University, 304–308. (In Russian)].

Палийчук Д. А. Корпусные технологии в лингвистических исследованиях // Гуманитарные исследования. История и филология. 2022. № 6. С. 72–79. [Palytchuk, Darya A. (2022) Korpusnye tekhnologii v lingvisticheskikh issledovaniyakh (Corpus Technologies in Linguistic Research). *Humanitarian Studies. History and Philology*, 6, 72–79. (In Russian)].

Сидорова Е. А., Иванов А. И., Овчинникова К. А. Извлечение информации из текстов на основе онтологии и больших языковых моделей // Онтология проектирования. 2025. Т. 15. № 1 (55). С. 114–129. [Sidorova, Elena A., Ivanov, Aleksandr I., & Ovchinnikova, Kristina A. (2025) Izvlechenie informatsii iz tekstov na osnove ontologii i bol'shikh yazykovykh modeley (Information Extraction from Texts Based on Ontology and Large Language Models). *Ontology of Designing*, 15 (1), 114–129. (In Russian)]. DOI: 10.18287/2223-9537-2025-15-1-114-129.

Хохлова М., Еникеева Е. Методы машинного обучения применительно к задаче выделения глагольных и атрибутивных коллокаций // Компьютерная лингвистика и вычислительные онтологии. 2020. № 4. С. 54–60. [Khokhlova, Maria., & Enikeeva, Ekaterina. (2020) Metody mashinnogo obucheniya primenitel'no k zadache vydeleniya glagol'nykh i atributivnykh kollokatsiy (Machine Learning Methods for the Extraction of Verbal and Attributive Collocations). *Computer Linguistics and Computational Ontologies*, 4, 54–60. (In Russian)]. DOI: 10.17586/2541-9781-2020-4-54-60.

Alguliyev, Rasim M., & Gurbanova, Afruz M. (2018) The Conceptual Foundations of National Terminological Information System. *International Journal of Education and Management Engineering (IJEME)*, 8 (2), 31–49.

Bowker, Lynne, & Pearson, Jennifer. (2002) *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Cabre, Teresa M. (1999) *Terminology: Theory, Methods, and Applications*. Amsterdam: John Benjamins Publishing.

Davies, Mark. (2023) Creating and using “Virtual Corpora” to extract and analyse domain-specific vocabulary at English-Corpora.org. In Pan, Jun, & Laviosa, Sara. (eds.) *Corpora and Translation Education: New Frontiers in Translation Studies*. Springer, 89–108 . https://doi.org/10.1007/978-981-99-6589-2_5.

Fuertes-Olivera, Pedro, & Tarp, Sven. (2014) *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110349023.

Kosem, Iztok, Gantar, Polona, Logar, Natasa, & Krek, Simon. (2014) Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies. In *Proceedings of the XVI EURALEX International Congress: The User in Focus, Lexicography and Language Technologies* Leiden: EURAC research, 355–364.

Manning, Christopher, Raghavan, Prabhakar, & Schütze, Hinrich. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Seretan, Violeta. (2010). *Syntax-Based Collocation Extraction* (1st ed.). Dordrecht: Springer-Verlag.

Shao, Louis, Gouws, Stephan, Britz, Denny, Goldie, Anna, Strope, Brian, & Kurzweil, Ray. (2017). Generating High-quality and Informative Conversation Responses with Sequence-to-sequence Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, 2210–2219.

Varadi, Tamas, & Héja, Enikő. (2011) Multilingual Term Extraction from Parallel Corpora: A Methodology for the Automatic Extraction of Verbal Structures and Their Translation Equivalents // *Magyar Terminologia*, 4 (2), 226–237. (2011) Multilingual Term Extraction from Parallel Corpora: A Methodology for the Automatic Extraction of Verbal Structures and Their Translation Equivalents // *Magyar Terminologia*, 4 (2), 226–237.