МИНОБРНАУКИ РОССИИ
**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
«КАРЕЛЬСКИЙ НАУЧНЫЙ ЦЕНТР РОССИЙСКОЙ АКАДЕМИИ НАУК»
(КарНЦ РАН)**

ул. Пушкинская, 11, г. Петрозаводск, 185910
тел. (8142) 76-97-10, 76-60-40, факс 76-96-00 E-mail: krcras@krc.karelia.ru
ОКПО 02700018, ОГРН 1021000531133 ИНН/КПП 1001041594/100101001

УТВЕРЖДАЮ

И. о. Генерального директора Федерального государственного бюджетного учреждения науки Федеральный исследовательский центр «Карельский научный центр Российской академии наук»
член-корреспондент РАН,
доктор биологических наук

_____О.Н. Бахмет

«27» ноября 2024 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного учреждения науки Федеральный исследовательский центр «Карельский научный центр Российской академии наук» на диссертационную работу Тарасова Никиты Андреевича «Гибридные нейросетевые методы анализа понятности текстов юридических документов на русском языке», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.1. – Системный анализ, управление и обработка информации, статистика

Актуальность темы, цель и задачи диссертации

Диссертация Тарасова Никиты Андреевича посвящена разработке методов анализа сложности правовых документов на основе гибридного нейросетевого моделирования текстовых данных. Применение алгоритмов машинного обучения для обработки естественного языка позволяет эффективно анализировать, классифицировать и интерпретировать большие объемы юридических текстов. Определение понятности и сложности юридических документов имеет особенно важное значение в вопросах взаимодействия юристов и неспециалистов в юридической сфере. Анализ сложности, и как следствие, - повышение доступности юридических текстов, является значимым направлением развития технологий в информационно-правовой сфере, в мировой литературе известном как LegalTech.

Целью рецензируемой диссертации является разработка и апробация методов обработки юридических текстов и программном обеспечении процесса определения доступности их восприятия.

Для этого были сформулированы и решены следующие **задачи**:

- изучение существующих подходов в области анализа юридических документов;
- отбор статистически эффективных текстовых показателей, наиболее полно описывающие юридические документы в контексте сложности и доступности восприятия;
- разработка гибридного метода оценки сложности юридических текстов, объединяющего традиционные подходы и гибридные нейросетевые методы;
- разработка комплексного подхода к анализу сложности юридических текстов, основанного предложенном методе;
- тестовые эксперименты с пакетом юридических документов и практическое использование в налоговой сфере, подтверждающие адекватность предложенных решений.

Новизна полученных результатов

Выносимые на защиту результаты являются новыми, к ним можно отнести следующие.

Разработан гибридный метод к оценке понятности юридических документов на русском языке, объединяющем традиционные подходы к ручному построению набора языковых характеристик с нейросетевыми моделями.

Создан наборов юридических текстов, относящихся к широкому кругу стилей и жанров, включающий в себя данные, подходящие для работы методов машинного обучения с учителем.

Применен оригинальный подход к обучению модели, основанный на размеченных данных учебников на русском языке различного уровня сложности по предметам, относящимся к области юриспруденции, и тестированию на реальных юридических документах.

Наконец, новыми являются результаты экспериментов различных типов юридических документов (указы, постановления и другие государственные юридические документы), свободных форм (ответы на юридические вопросы в сфере налогообложения), представляющие несомненный интерес и для юристов-практиков.

Теоретическая и практическая значимость

Теоретическая значимость выражается в разработке методов, ориентированных на анализ юридических документов на русском языке, связанных со сложностью и доступностью восприятия, с использованием нейросетевых методов.

Практическая значимость работы выражается в разработке комплекса методов и программ для автоматизированного анализа русскоязычных юридических текстов с целью оценки сложности и доступности восприятия. Предложенные подходы и инструменты способствуют ускорению внедрения информационных технологий в юридические

процессы, что, в перспективе, способно улучшить качество взаимодействия населения с государственными органами.

Содержание работы

Во введении показана актуальность и новизна исследования, описана теоретическая и практическая значимость, а также поставлены цель и задачи исследования.

Первая глава посвящена методике статистической оценки частотных характеристик юридической лексики в различных типах документов. Предложена методика, включающая получение и статистическую обработку данных, являющаяся важной основой для последующего анализа и создания описательных характеристик документов.

Во второй главе определен набор признаков, оценивающих понятность юридических документов, и проведен анализ их эффективности, а на их основе предложена методика классификации сложности юридических текстов. Оценка понятности документов основана на расчете языковых характеристик и сравнении моделей, использующих эти характеристики, с алгоритмами, основанными на языковых моделях.

В третьей главе представлен гибридный метод оценки сложности, использующий языковые характеристики в сочетании с классическими алгоритмами машинного обучения и большие языковые модели.

Четвертая глава включает сравнительный анализ сложности юридических документов различных стилей и жанров на основе использования гибридной семантической модели предсказания сложности.

В пятой главе приведен пример адаптации методологии для анализа ответов на юридические вопросы в области налогообложения.

Заключение содержит выводы и основные результаты диссертационной работы.

Апробация работы и публикации

Основные результаты по теме диссертации изложены в 9 печатных изданиях, индексируемых Web of Science и Scopus, из которых 4 — в периодических научных журналах, 5 — в тезисах докладов.

Получены 3 свидетельства о государственной регистрации программ для ЭВМ.

Результаты исследований докладывались на ряде международных конференций.

Замечания

Имеется ряд замечаний:

1. Теоретическая значимость работы (стр. 7-8 текста диссертации) объясняется повышением эффективности «... решения задач интеллектуального анализа юридических документов», что является сомнительным объяснением, поскольку эффективность решения таких задач в работе не определена. Кроме того, здесь же теоретическая значимость подтверждается участием в научно-исследовательских проектах, что, скорее, свидетельствует о практической реализации предлагаемых подходов.

2. Заголовки некоторых разделов являются малоинформативными и/или не отражают их содержания. Например, раздел 3.5 «Постановка эксперимента» на самом деле содержит описание построения собственно гибридной модели оценки сложности, о которой заявлено в заголовке главы 3.

3. Глава 3 «Гибридная модель оценки сложности: разработка и применение для российских юридических текстов» изобилует большим количеством описаний лингвистических характеристик, что лучше было бы вынести в отдельную главу. Это изобилие приводит к тому, что собственно модели уделено мало внимания, а хотелось бы видеть отдельную главу, посвященную именно ей.

4. При этом функционирование модели оценки сложности описано в подразделе, который так и называется «4.3.3 Модель оценки сложности», что свидетельствует о не совсем удачном структурировании материала в диссертации.

5. В п.5 перечисляемых задач (стр. 7), по-видимому, ошибочно употреблено словосочетание «программный комплекс», нигде далее не используемое в работе. Как известно, программный комплекс — это набор взаимодействующих программ, согласованных по функциям и форматам, имеющих единообразные, точно определённые интерфейсы и составляющих полное средство для решения больших задач.

На самом деле в диссертации речь идет о комплексе методов и программ, предназначенных для автоматизированного интеллектуального анализа русскоязычных юридических текстов с целью оценки их сложности и доступности восприятия.

6. Обозначения на рисунках (их более 20) даются на английском языке и/или аббревиатурах на латинице без пояснений в тексте, что затрудняет (а иногда и делает невозможным) понимание рисунков. Например, на рис. 2.1 (стр. 34) приводятся обозначения ACW (расшифровка на стр. 96) и SMOG (расшифровка на стр. 47 как Simple Measure of Gobbledygook).

7. Без корректной расшифровки используется аббревиатура НКРЯ, очевидно, для обозначения Национального корпуса русского языка (он упомянут в абзаце выше). Следовало бы отнестись к этому ресурсу более аккуратно и привести ссылку как «Национальный корпус русского языка – URL:https://ruscorpora.ru», а не непонятную ссылку на стр. 108: Corpus, R. N. / R. N. Corpus. — URL: http://www.ruscorpora.ru/new/.

Следующие замечания скорее являются пожеланиями на будущее.

8. Приведенные в работе свидетельства о регистрации программ относятся к программам анализа социальных сетей, а не к программе анализа сложности юридических документов, для которой следовало бы оформить соответствующий документ.

9. Основная публикация по оценке сложности русских юридических текстов написана на английском языке (O. Blinova, N. Tarasov // Frontiers in Artificial Intelligence. — 2022. — Т. 5. — С. 1008530). Возможно, формально это повышает ее международную значимость, но усложняет понимание для практиков.

10. Использованная в диссертации модель BERT была выпущена компанией Google в 2018 году. В настоящее время существует довольно широкий выбор более современных и доступных моделей для работы именно с русским языком, выпущенных значительно позже BERT, на которые следует обратить внимание.

Выводы

Работа «Гибридные нейросетевые методы анализа понятности текстов юридических документов на русском языке» является законченным самостоятельным научным исследованием, выполненным на хорошем научном уровне. Результаты представляют определенную научную ценность в области моделирования сложности юридических текстов и могут найти применение в юридической сфере. Они опубликованы в авторитетных научных журналах и доложены на ряде научных конференций.

Перечисленные в замечаниях недостатки не влияют на окончательную положительную оценку работы.

Диссертация Тарасова Никиты Андреевича на тему «Гибридные нейросетевые методы анализа понятности текстов юридических документов на русском языке» по специальности 2.3.1. – Системный анализ, управление и обработка информации, статистика соответствует основным требованиям, установленным Приказом № 11181/1 от 19.11.2021 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете» (с изменениями и дополнениями), а ее автор – Тарасов Н.А. заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.1. – Системный анализ, управление и обработка информации, статистика.

Доклад Тарасова Никиты Андреевича по теме диссертации заслушан на научном семинаре ИПМИ КарНЦ РАН 29 октября 2024 года.

Отзыв составлен доктором технических наук, ведущим научным сотрудником лаборатории математической кибернетики ИПМИ КарНЦ РАН Печниковым Андреем Анатольевичем.

_____ А.А. Печников

Отзыв на диссертацию Тарасова Никиты Андреевича обсужден и поддержан на заседании Ученого совета Института прикладных математических исследований КарНЦ РАН 26 ноября 2024 г., протокол № 10.

Отзыв на диссертацию Тарасова Никиты Андреевича рассмотрен и одобрен в качестве официального отзыва ведущей организации на заседании Ученого совета Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Карельский научный центр РАН» 26 ноября 2024 г., протокол № 10.

Председатель Ученого Совета КарНЦ РАН
Член-корр. РАН, д.б.н. _____ О.Н. Бахмет

Собственноручные подписи

А.А. Печникова и О.Н. Бахмет удостоверяю:

Ученый секретарь КарНЦ РАН _____ Н.Н. Фокина

27 ноября 2024 г.

Сведения о ведущей организации

Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Карельский научный центр Российской академии наук» (КарНЦ РАН)

Адрес: 185910, Республика Карелия, г. Петрозаводск, ул. Пушкинская, д. 11

Телефон: +7(8142)766040

Факс: +7(8142)769600

Электронная почта krcras@krc.karelia.ru

Web -сайт: http://www.krc.karelia.ru