УДК 004.912:004.4:005

ПОВЫШЕНИЕ КАЧЕСТВА АНАЛИЗА И ОБРАБОТКИ ПРАВОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

[1,2]Ермолатий Д.А., [2]Быстров А.И.

[1]ФГБУ «Федеральное бюро медико-социальной экспертизы» Минтруда Российской Федерации, Москва, e-mail: denis.yermolatiy@yandex.ru;
[2]ФГБОУ ВО «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации», Москва, e-mail: alexandr.jri.bystrov@yandex.ru

Проведены исследования возможностей автоматизации анализа правовой информации, применяемой для повышения точности и качества обработки и анализа, и оперативности принимаемых экспертными и рабочими группами решений. Проведен новый этап исследований, обосновано применение метода машинного обучения в обработке юридических текстов, начата разработка собственного программного продукта на основе описанного средства. Подробно рассмотрен пример метода машинного обучения для обработки и анализа текстов. Описана методика применения нейросетевой модели для работы с юридически значимыми аспектами текстов и приведены некоторые рекомендации по дальнейшему развитию автоматизации процессов деятельности экспертов в законотворчестве. В виде графиков и таблиц приведены результаты эксперимента применения метода машинного обучения к корпусу нормативно-правовых актов. Разработаны рекомендации методологического и методического характера по работе с нормативно-правовой информацией. Предложенная методика работы с применением нейросетевой обработки и дальнейшего анализа нормативно-правовой информации в рамках законотворчества показывает практически значимый результат и способствует более быстрому принятию решений, повышая тем самым качество и оперативность подготовки нормативно-правовых актов.

Ключевые слова: обработка текстов, правовая информация, VBA-скрипт, метод и средство системного моделирования, машинное обучение, Python, нейросети

MPROVING THE QUALITY OF ANALYSIS AND PROCESSING OF LEGAL INFORMATION BASED ON THE USE OF MACHINE LEARNING METHODS

[1,2]Ermolatiy D.A., [2]Bystrov A.I.

[1]Federal State Budgetary Institution Federal Bureau of Medical and Social Expertise of the Ministry of Labor of Russia, Moscow, e-mail: denis.yermolatiy@yandex.ru;
[2]Russian Presidential Academy of National Economy and Public Administration, Moscow, e-mail: alexandr.jri.bystrov@yandex.ru

Research into the possibilities of automating the analysis of legal information used to improve the accuracy and reliability of processing and analysis, and the quality of decisions made by expert and working groups. A new stage of research was conducted; application of machine learning method in processing legal texts was substantiated and development of our own software product on the basis of the described tool was started. The example of machine learning method for processing and analysis of texts is considered in detail. Methodology of neural network model application to address legally relevant aspects of texts is described and some recommendations for further development of automation of expert processes in lawmaking are given. Experimental results of application of machine learning method to corpus of legislative acts are given in the form of tables and diagrams. Methodological and methodological recommendations for working with regulatory information have been developed. *Conclusions.* The proposed methodology of work with the application of neural network processing and further analysis of regulatory information in the law-making shows practically significant results and contributes to a faster decision-making, thereby improving the quality and efficiency of the preparation of regulatory legal acts.

Keywords: text processing, legal information, VBA script, method and tool of system modeling, machine learning, Python, neural networks

В ходе формирования и развития системы комплексной реабилитации в РФ, как и при регулировании любой сферы деятельности, уделяется особое внимание нормативно-правовым актам (далее – НПА). С появлением новых редакций и уточняющих НПА в разы возрастает объем информации, требующий анализа и обработки после поверхностного ознакомления. Увеличение времени вынесения решений приводит к юридическим усложнениям, коллапсам, социальным спорам (напряженности) и судебным процессам. Об устойчивой напряженности сообщают эксперты-реабилитологи, опубликовавшие статистику в рамках тематической конференции «Состояние и перспективы развития системы комплексной реабилитации и абилитации инвалидов и детей-инвалидов в Российской Федерации». Согласно статистике, количество обжалований и социальных споров, в том числе на основании несогласованности законодательных актов за 2019–2020 гг., достигает 87 % и тенденция продолжает сохраняться [1, с. 180]. Приведенные данные подтверждаются также и открытой судеб-

ной статистикой, о чем написано в ранее опубликованной работе [2].

Правовая напряженность напрямую отрицательно влияет на социальную и потенциально экономическую (через недобросовестных граждан, пользующихся несовершенством нормативно-правового регулирования) сферы.

Целью данного исследования является автоматизация процессов анализа и обработки правовой информации посредством использования машинного обучения для повышения эффективности принятия управленческих решений.

Материалы и методы исследования

Исследование проводится на примере свода НПА в областях медико-социальной экспертизы (далее – МСЭ) и реабилитации и абилитации инвалидов (далее – РиАИ). Описано применение методики предварительного обучения и использования нейросетевой модели. Совершенствуются подходы и управленческие методы постановки и решения задач путем углубленного изучения научной проблемы.

Результаты исследования и их обсуждение

Государство каждый год проводит как обновления НПА, так и разработку новых, включая области МСЭ и РиАИ, с учетом того, что РиАИ во многом только начинает комплексно формироваться. Для систематизации работы (в данном случае – анализа и обработки) требуется максимально точно поставить задачу, а также определить варианты при работе с НПА (табл. 1).

Задачи при решении законодательной проблемы ставят формирующиеся рабочие группы.

После определения основной цели при решении поставленной рабочей группой задачи предлагается определенное новшество в управленческом подходе, а именно – формирование дополнительной экспертной группы, основной задачей которой и станет проведение технического анализа и обработки текстовых данных, используя предлагаемую программную среду.

В приведенных в данной статье результатах отражена обработка двух юридических документов:

1. Приказ Минтруда РФ от 29.12.2015 № 1171н «Об утверждении формы протокола проведения медико-социальной экспертизы гражданина в федеральном государственном учреждении медико-социальной экспертизы» (утратил силу).

2. Приказ Минтруда РФ от 04.07.2022 № 389н «Об утверждении формы и порядка заполнения протокола проведения медико-социальной экспертизы гражданина» (действующий).

«Основная сложность при работе с текстом связана с количеством слов, часть из которых не относится к полезной информации либо имеет равное значение <…>. Например, к ним относятся стоп-слова, которые являются вспомогательными (предлоги, союзы, частицы и др.); разные грамматические формы слов» [3].

«Для применения методов машинного обучения и интеллектуального анализа данных текстовые наборы необходимо преобразовать. Очистить от слов и символов, которые могут негативно сказаться на процессе распознавания. Одним из примеров подобной ситуации является использование иностранных слов в русскоязычных текстах» [4].

Таблица 1

Некоторые варианты работы с НПА

Варианты работы с НПА	Краткое описание работы
Выявление методологических изменений и нововведений при входе в силу нового НПА	Производится более точное определение внесенных обновлений в новом НПА, а также составление отчета по методическим и методологическим указаниям
Поиск технических нововведений при обновлении и выпуске новой редакции НПА	Нахождение новых позиций, описываемых в законодательных актах, при выпуске новой или расширенной редакции
Выяснения несогласованности между документами и действующими НПА	Нахождение неточностей и взаимоисключающих элементов и положений в НПА
Поиск обновленных и новых положений при замещении НПА	Нахождение новых и обновленных позиций при отмене НПА и замене такового новым
Взаимосвязи НПА с базой НПА выделенного направления	Ключевой НПА сравнивается с группой НПА определенной сферы
	И др.

Источник: создано Д.А. Ермолатием.

Таблица 2

Прямое сопоставление некоторых смысловых фреймов
после автоматизированной обработки с учетом нововведений

Старая версия (по приказу Минтруда России № 1171н от 29.12.2015 с ред. № 215 от 04.04.2019)	Новая версия (по приказу Минтруда России № 389н от 04.07.2022)	Кол-во заполняемых ячеек выбора
1. Дата подачи заявления: «__» _____ 20__ г.	1. Дата поступления направления на медико-социальную экспертизу медицинской организацией (органа, осуществляющего пенсионное обеспечение гражданина, выехавшего на постоянное место жительства за пределы Российской Федерации, страховщика (территориального органа Фонда социального страхования Российской Федерации, страхователя (работодателя), определения суда (судьи), заявления гражданина о проведении медико-социальной экспертизы (нужное подчеркнуть) (день, месяц, год): «__» _____ 20__ г.	текст
6. Дата рождения: день __ месяц __ год ____ 7. Дата смерти (заполняется в отношении умершего инвалида): день __ месяц __ год ____ 8. Возраст (число полных лет (для ребенка в возрасте до 1 года число полных месяцев)): _____	5. Дата рождения (день, месяц, год): «__» _____ 20__ г. возраст (число полных лет, для ребенка в возрасте до 1 года число полных месяцев): _____ дата смерти (день, месяц, год): «__» _____ 20__ г.	текст
----------	7. Гражданин находится на лечении в стационаре в связи с ампутацией (реампутацией) конечности (конечностей), нуждается в первичном протезировании	1
11. Отношение к воинской обязанности <2>: 11.1. | | военнообязанный |11.2. | | лицо призывного возраста	10. Отношение к воинской обязанности: Гражданин, состоящий на воинском учете Гражданин, не состоящий на воинском учете, но обязанный состоять на воинском учете Гражданин, поступающий на воинский учет Гражданин, не состоящий на воинском учете	4
----------	13. Гражданин находится (нужное отметить и указать):	
----------	13.1. в медицинской организации, оказывающей медицинскую помощь в стационарных условиях	1

Источник: создано Д.А. Ермолатием.

На первом этапе использовался Visual Basic for Applications (далее – VBA) [2], в работе с которым определяются критерии оценки и анализа юридической информации. В случае работы с НПА при использовании VBA на первом этапе и методов машинного обучения на втором, глобальной подготовки и определенной «чистки» текстов не требуется ввиду двух основных причин: редкость стоп-слов (иностранных и латинских) и малое влияние на точность измерений и результатов.

В рамках проводимого опыта первым шагом сравниваются два НПА друг с другом в режиме (формате) «один-один», с выбранной позицией из табл. 1 «Поиск обновленных и новых положений при замещении НПА», т.е. один НПА теряет силу при введении нового.

В пределах фрейма выявляются логико-семантические отношения для построения взаимосвязей и отражения таковых в специализированном отчете, предусмотренном скриптом [2]. Результаты были представлены предметным специалистам в виде таблицы как для проведения экспертной оценки, так и для детального изучения (табл. 2) и вынесения первичной экспертной оценки.

Для разработки программного продукта (при переходе на следующий этап разработки и формирования средства анализа и обработки правовой информации), а также проведения первичного эксперимента по использованию методов машинного обучения выбран язык программирования Python, средой разработки использован JupiterLab. «Функциональные возможности языка Python значительно больше подходят для проектирования сложных структур» [5]. Язык обладает большим количеством библиотек для работы с документами и текстами. В качестве базового инструмента в рамках первичного опыта для векторизации была использована модель Bidirectional Encoder Representation Transformers (далее – BERT) [6], обученная на русскоязычном корпусе текстов и новостных изданиях, подготовленная Лабораторией нейронных систем и глубокого обучения МФТИ, а также для более точного распознавания прагматической составляющей юридических текстов дополнительно проведено самостоятельное «дообучение» на ряде НПА.

С помощью вышеупомянутой модели возможно извлечь вектора слов, из которых получим вектора фреймов.

Для токенизации слов также используем заранее подготовленный токенизатор от создателей модели [7]. Это необходимо, чтобы разбить текстовый ввод на токены (уникальные идентификаторы слов, словосочетаний, буквенных комбинаций) и использовать в качестве входных данных модели BERT.

Изначально мы находим массив слов для всего абзаца, после чего усредняем векторы в массиве, чтобы получить *вектор абзаца (фрейма)*. Далее используется косинусное расстояние для нахождения сходства между фреймами. Вывод распределен от 0 до 1.

Для удобства работы с токенизатором и моделью авторами предлагается использовать функцию pipeline из библиотеки HuggingFace [8], которая позволяет совместить ряд операций в одной функции: предобработка данных, обработка данных моделью (включая вычисление расстояния) и получение вывода (рис. 1).

Для проведения эксперимента был собран корпус юридических текстов, состоящий из 16 похожих пар абзацев и 16 непохожих. Формирование корпуса выполнено с участием экспертов и запущено в обработку после проведения экспертной оценки.

Для каждой пары фреймов с применением предобученной модели BERT и инструментария, предусмотренного вышеупомянутой библиотекой HuggingFace, найдено расстояние (дистанция) (табл. 3).

Среднее косинусное расстояние для схожих текстов составило 0,32. Среднее расстояние для несхожих текстов составило 0,54 (рис. 2).

Далее эмпирически подбирается порог принятия решения – являются ли абзацы схожими или различными по смыслу (прагматически). С учетом экспертной оценки порог в нашем опыте составил 0,5.

Исходя из этого, мы признаем все пары фреймов, где порог не превысил расстояние 0,5, схожими, а где превысил – несхожими. Это дает нам возможность посчитать f-меру для оценки подхода (методики) (рис. 3).

Результаты обработки НПА представлены в табл. 3.

Как можно заметить из значений метрик и матрицы оценки, предлагаемый подход (методика) демонстрирует достаточно точные результаты в определении схожести текстов, на основании чего можно проводить экспертную оценку групп НПА. Получено подтверждение экспертным сообществом о сокращении требуемого времени на анализ нормативно-правовой информации в среднем до 10% в сравнении с применением VBA, что доказывает эффективность. Дополнительное время, полученное путем сокращения временных затрат на первичные обработку и анализ, направляется на углубленную юридическую оценку формирующихся предложений к нововведениям.

При дальнейшем накоплении опыта и развитии средств обработки и анализа будет разрабатываться самостоятельный программный продукт, который в своей финальной реализации сможет предсказывать согласованность разрабатываемого НПА на основе комплексного сравнения с группами НПА заданной предметной области, значение может быть выражено в процентах.

```
def comare_and_get_score(str1: str,
                         str2: str):
    data1 = np.mean(pipeline(str1)[0], axis=0).reshape(1, -1)
    data2 = np.mean(pipeline(str2)[0], axis=0).reshape(1, -1)
    dist = cosine_distances(data1,data2)[0][0]
    return dist
```

Рис. 1. Задание функции сравнения фреймов средствами языка Python (создано А.И. Быстровым)

Измеряем расстояние между сходными по смыслу абзацами

```
In [24]: distances = []
         for _, row in df.iterrows():
             distances.append(comare_and_get_score(row['initial'], row['compare']))
         df['distances'] = distances

In [25]: df['distances'].mean()

Out[25]: 0.32481344615031427
```

Измеряем расстояние между различными по смыслу абзацами

```
In [12]: distances=[]
         for _, row in df.iterrows():
             one_out_df = df[df['initial']!=row['initial']]
             dist = []
             for _, roww in one_out_df.iterrows():
                 dist.append(comare_and_get_score(roww['initial'], row['initial']))
             distances.append(np.mean(dist))
         df['distances_between_initial'] = distances

In [13]: distances=[]
         for _, row in df.iterrows():
             one_out_df = df[df['compare']!=row['compare']]
             dist = []
             for _, roww in one_out_df.iterrows():
                 dist.append(comare_and_get_score(roww['compare'], row['compare']))
             distances.append(np.mean(dist))
         df['distances_between_compare'] = distances

In [14]: df['distances_between_initial'].mean()

Out[14]: 0.5398986338200013
```

Рис. 2. Получение расстояний (создано авторами)

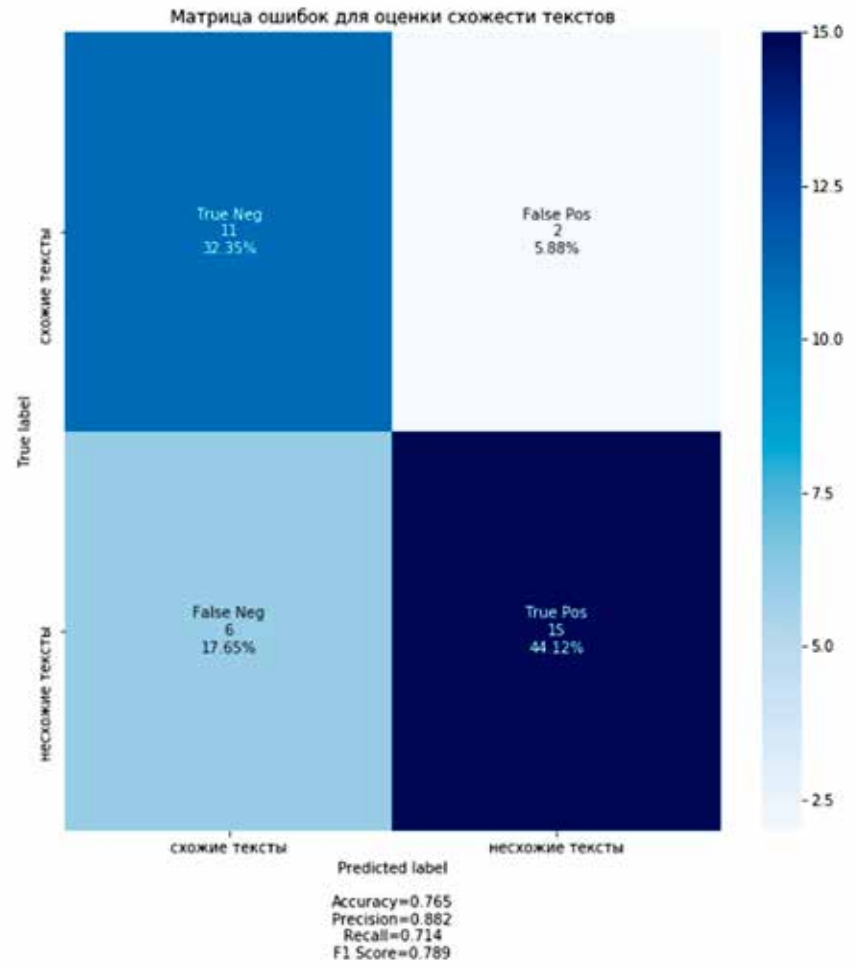

Рис. 3. Матрица оценки схожести текстов (создано А.И. Быстровым)

Таблица 3

Некоторые из результатов проведенного эксперимента

Абзац для сравнения 1	Абзац для сравнения 2	Экспертная оценка	Оценка модели (расстояние)	Решение, основанное на оценке модели
Наименование образовательной организации, в которой получает образование:	Сведения об образовательной организации (полное наименование, юридический адрес), в которой гражданин, в отношении которого проводится медико-социальная экспертиза, получает образование:	Схожи	0,395	Схожи
Курс, класс (указываемое подчеркнуть):	Курс, класс, возрастная группа дошкольной образовательной организации (нужное подчеркнуть и указать):	Схожи	0,315	Схожи
Федеральные органы государственной власти, органы государственной власти субъектов Российской Федерации, органы местного самоуправления (в сфере установленных полномочий), организации независимо от их организационно-правовых форм обеспечивают инвалидам (включая инвалидов, использующих кресла-коляски и собак-проводников): 1) условия для беспрепятственного доступа к объектам социальной, инженерной и транспортной инфраструктур (жилым, общественным и производственным зданиям, строениям и сооружениям, включая те, в которых расположены физкультурно-спортивные организации, организации культуры и другие организации), к местам отдыха и к предоставляемым в них услугам;	Чтобы наделить инвалидов возможностью вести независимый образ жизни и всесторонне участвовать во всех аспектах жизни, государства-участники принимают надлежащие меры для обеспечения инвалидам доступа наравне с другими к физическому окружению, к транспорту, к информации и связи, включая информационно-коммуникационные технологии и системы, а также к другим объектам и услугам, открытым или предоставляемым для населения, как в городских, так и в сельских районах. Эти меры, которые включают выявление и устранение препятствий и барьеров, мешающих доступности, должны распространяться	Схожи	0,192	Схожи
Получение консультативного заключения главного бюро или Федерального бюро	дополнительное обследование в образовательных организациях (в том числе в психолого-медико-педагогических комиссиях)	Не схожи	0,732	Не схожи
Место работы:	Мероприятия психолого-педагогической реабилитации и (или) абилитации	Не схожи	0,767	Не схожи
Сведения об образовательной организации (полное наименование, юридический адрес), в которой гражданин, в отношении которого проводится медико-социальная экспертиза, получает образование:	Максимальный срок ожидания в очереди при подаче заявителем лично заявления о предоставлении государственной услуги и при получении результата предоставления государственной услуги составляет не более пятнадцати минут.	Не схожи	0,402	Схожи

Источник: создано авторами.

Заключение

Применение машинного обучения способствует как сокращению времени на анализ юридической информации аналитиками и экспертами, так и получению более точного сравнения фреймов. Использование модели BERT повышает общее качество принимаемых решений и, как вывод, повышает общее качество НПА и согласованность между ними.

Представленная методика может полноценно использоваться не только в сферах МСЭ и РиАИ, но и во многих других. Дополнительные испытания (например, в сфере высшего образования) будут приведены в новых работах.

Предлагаемая методика обработки и анализа – использование метода предварительного обучения и модели BERT, позволяет ускорить процесс оценки и согласования различных вопросов в рамке действия рабочих и экспертных групп при работе с правовой информацией в законотворческой деятельности. Также и повысить качество рассматриваемых законодательных поправок и редакций при формировании предложений к дальнейшим нововведениям. Важной особенностью методики является относительная доступность для всех участников законотворческого процесса.

В последующих статьях также будет описан ход создания собственного продукта на основе описанной модели. Будут рассмотрены и уточнены инструментарий автоматизации определения вариантов обработки и выбора целевых функций для анализа юридической информации с привлечением экспертов в качестве «учителей» для нейросетевых моделей. Такой опыт может лечь в основу для создания отдельной информационной системы (в том числе государственной).

Список литературы

1. Морозова Е.В., Жукова Е.В. Повышение информированности пациентов и их законных представителей в рамках социально-ориентированной технологии «школа социальной жизни» // Состояние и перспективы развития системы комплексной реабилитации и абилитации инвалидов и детей-инвалидов в РФ: сборник материалов и докладов. М., 2022. 366 с.

2. Ермолатий Д.А. Анализ правовой информации прикладными средствами при использовании VBA-скриптов // Современные наукоемкие технологии. 2022. № 9. С. 22–26.

3. Валиев А.И., Лысенкова С.А. Применение методов машинного обучения для автоматизации процесса анализа содержания текста // Вестник кибернетики. 2021. № 4 (44). С. 12–15.

4. Томашевская В.С., Старичкова Ю.В., Яковлев Д.А. Использование машинного обучения для распознавания текстовых шаблонов литературных источников // Известия высших учебных заведений. Поволжский регион. Технические науки. 2022. № 3. С. 15–25.

5. Леметюйнен Ю.А., Дударов С.П. Сравнительный анализ возможности нейросетевого моделирования на языке программирования Python и в среде Matlab // Успехи в химии и химической технологии. 2021. № 3 (238). С. 6–8.

6. QuData. Портал разработчиков ИИ/МО [Электронный ресурс]. URL: https://qudata.com/ml/ru/NN_Attention_BERT.html (дата обращения: 08.01.2023).

7. DeepPavlov Github Блог-портал экспертов разработки решений машинного обучения. [Электронный ресурс]. URL: https://github.com/deeppavlov/DeepPavlov (дата обращения: 08.01.2023).

8. The AI Community. Open source in machine learning. [Электронный ресурс]. URL: https://huggingface.co/docs/transformers/main_classes/pipelines (дата обращения: 08.01.2023).