Санкт-Петербургский государственный университет

На правах рукописи

Тарасов Никита Андреевич

Гибридные нейросетевые методы анализа понятности текстов юридических документов на русском языке

Научная специальность 2.3.1.
Системный анализ, управление и обработка информации, статистика

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
канд. физ.-техн. наук, доцент
Блеканов Иван Станиславович

Санкт-Петербург — 2024

Оглавление

Стр.

Введение

Использование современных методов сбора, обработки и анализа данных ведет к развитию существующих областей и созданию принципиально новых технологий в информационно-правовой сфере (LegalTech). В настоящее время к технологиям LegalTech, как правило, относятся технологические решения, осуществляющие автоматизацию различных юридических процессов: сбор, обработка и анализ больших объемов юридических данных, информационное сопровождение различных правовых процессов и т.п.

Автоматизированная обработка больших массивов юридических текстов с применением нейросетевых моделей и технологий позволит эффективно и качественно решать целый ряд задач правового процесса. В частности, современные методы языкового моделирования могут использоваться для решения задач определения сложности юридических документов, актуальных не только для отдельных компаний, но и в масштабах государства. Таким образом, повышсится эффективность правового процесса за счет увеличения доступности восприятия больших объемов юридической информации.

Актуальность темы исследования. Автоматическая обработка юридических текстов представляет все больший научный и практический интерес. Современные методы обработки данных и искусственный интеллект значительно улучшают способы работы с юридическими текстами. Применение алгоритмов машинного обучения и обработки естественного языка позволяет эффективнее анализировать, классифицировать и интерпретировать большие объемы правовой информации.

Интеллектуальные методы анализа текстовых данных дают возможность как структурировать содержимое отдельных документов, так и категоризировать корпуса документов различных типов, с учетом семантики, а также эффективно выявлять признаки, описывающие разнообразные языковые характеристики содержимого. Методы на основе алгоритмов машинного обучения и технологий обработки естественного языка способны осуществлять более глубокий анализ текстов и извлекать семантически значимую информацию из объемных корпусов документов.

Применение современных методов обработки текстов в юридической сфере помогут в значительной степени минимизировать риски ошибок при анализе

правовых текстов и обеспечить более точное выполнение юридических процедур.

Большие языковые модели (Large Language Models, LLM) - эффективный современный подход для решения различных задач в области интеллектуальной обработки текстов, в том числе юридических. Однако для эффективного применения таких моделей для русского языка и с учетом различных юридических контекстов необходимо дообучение и точная настройка моделей. В программах, представленных в свидетельствах о регистрации [1—3] приводятся примеры возможностей дообучения языковых моделей и их адаптации для работы с текстами на русском языке. Для создания программных комплексов исследовались методологические ограничения языковых моделей в контексте анализа пользовательского контента в социальных сетях. Специфичность языка и нестандартные размеры документов объединяют задачи анализа юридических документов и пользовательских сообщений.

Сложность изложения правовых документов часто создает преграды к эффективной коммуникации различных сторон юридического процесса. В данном случае определение понятности документов особенно актуально для повышения качества взаимодействия юристов и лиц, не являющихся профессионалами в юридической сфере. Так, выявление нечетких языковых конструкций способствует предотвращению возможных двоечтений.

Таким образом, определение сложности и понятности юридических документов дает возможность увеличить доступность правовой информации, позволяет выявить потенциально неоднозначные и чрезмерно усложненные части документов различных типов - от соглашений и заявлений до указов и постановлений.

Целью диссертационной работы является разработка и апробация методологических и инструментальных средств интеллектуальной обработки юридических текстов и алгоритмическом обеспечении процесса определения доступности их восприятия.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Изучить современное состояние юридических, лингвистических исследований в области анализа юридических документов, выявить актуальные проблемы и определить возможные методы их решения.

2. Разработать методологические подходы для сбора, обработки и семантического анализа русского правового языка.
3. Разработать методологию статистической оценки частотных характеристик юридического языка.
4. Выявить и отобрать языковые характеристики юридических документов, наиболее полно описывающие их в контексте сложности и доступности восприятия.
5. Разработать программный комплекс для интеллектуального анализа сложности юридических документов на основе гибридных нейросетевых методов использования языковых моделей.
6. Провести сравнительный анализ сложности документов различных подстилей и жанров с использованием гибридной оценки сложности.
7. Провести практический анализ доступности восприятия юридических текстов с использованием представленных моделей и методов.

Научная новизна выполненных исследований заключается в следующем:

- Основываясь на современных лингвистических и юридических исследованиях, а также экспертных оценках выявлен и реализован наиболее полный список понятностных характеристик русского правового языка.
- Опираясь на современные научно-технические методы была разработана система интеллектуальной обработки данных в задачах оценки сложности и доступности восприятия юридических текстов.
- Разработан комплекс подходов, адаптированных для русского языка, созданы специализированные базы данных юридических текстов различных видов и направленностей.
- Представлена и протестирована методология гибридной нейросетевой оценки сложности юридических документов.
- Протестирована и апробирована система оценки сложности документов для различных типов юридических документов, как стандартизированных (указы, постановления и другие государственные юридические документы), так и в свободной форме (ответы на юридические вопросы в сфере налогообложения).

Теоретическая значимость. Разработанный комплекс подходов и программ существенно повысит эффективность решения задач интеллектуального анализа юридических документов, связанных со сложностью и доступностью

восприятия. Теоретическая значимость работы подтверждена участием в следующих научно-исследовательских проектах:

- №19-18-00525 “Понятность официального русского языка: юридическая и лингвистическая проблематика”, 2020-2023 гг. (Российский научный фонд, исполнитель)
- №96417361 “Юридико-лингвистическая неопределенность в текстах правовых актов с учетом их коммуникативных особенностей и юридических функций”, 2023-2024 гг. (Госзадание - Грант на НИР за счет средств СПбГУ, исполнитель)
- №93825201 Проект «Научно-исследовательский институт проблем государственного языка», 2022 г. (Санкт-Петербургский государственный университет, исполнитель)
- №5-6-01/79 “Выполнение работ по исследованию уровня доступности восприятия письменных ответов налоговых органов на обращения физических лиц и организаций”, 2023 г. (ФНС России, исполнитель)
- №92564627 “Центр международных медиаисследований”, 2023 г. (Госзадание - Грант на НИР за счет средств СПбГУ, исполнитель)
- №16-18-10125-П “Кривое зеркало конфликта: роль сетевых дискуссий в репрезентации и динамике этнополитических конфликтов в России и за рубежом”, 2019-2020 гг. (Российский научный фонд, исполнитель)
- №21-18-00454 “Медиатизированная коммуникация и современный деливеративный процесс”, 2023 г. (Российский научный фонд, исполнитель)

Практическая значимость. На основе проведённых исследований разработан комплекс методов и программ, предназначенных для автоматизированного интеллектуального анализа русскоязычных юридических текстов с целью оценки их сложности и доступности восприятия. Предлагаемые подходы и инструменты позволяют анализировать различные типы правовых документов, способствуя ускорению внедрения информационных технологий в реальные юридические процессы. Разработанные методы могут найти применение в научной сфере (например, в лингвистике и юриспруденции), а также повысить эффективность работы профильных специалистов и улучшить качество взаимодействия населения с государственными органами.

Апробация работы. Основные результаты работы были представлены на следующих конференциях:

- Международная 15-ая конференция SCSM 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Дания, 23.07.2023
- Международная конференция IAMCR Annual conference 'Inhabiting the planet: Challenges for media, communication and beyond', Франция, 13.07.2023
- Международные пятнадцатые международные научные чтения в москве «сми и массовые коммуникации–2023» : эпоха неопределенности в современных сми и журналистике: вызовы больших данных и искусственного интеллекта, Российская Федерация, 09.11.2023 - 10.11.2023
- 25-я Международная конференция по человек-компьютерному взаимодействию : HCI International - 2023 ('hybrid' conference), Дания, 23.07.2023
- 27-й Мировой конгресс политической науки (IPSA/AISP'2023), Аргентина, 15.07.2023 - 19.07.2023
- Международная конференция «Диалог 2022», Российская Федерация, 15.06.2022 - 18.06.2022
- Всероссийская международная конференция по естественным и гуманитарным наукам с международным участием "Наука СПбГУ – 2021", Российская Федерация, 28.12.2021 - 28.12.2021
- Международная конференция Networks in the Global World 2022, Российская Федерация, 22.06.2022 - 24.06.2022
- Международная 13-я конференция Social Computing and Social Media, SCSM 2021, held as part of the 23rd International Conference, HCI International 2021, Online, 24.07.2021 - 29.07.2021
- Корпусная лингвистика - 2021: международная научная конференция, Российская Федерация, 30.06.2021 - 03.07.2021
- Международная 12-я конференция Social Computing and Social Media, SCSM 2020, held as part of the 22nd International Conference on Human-Computer Interaction, HCII 2020, Дания, 19.07.2020 - 24.07.2020
- 26-я Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», Российская Федерация, 17.06.2020 - 20.06.2020

Публикации. Основные результаты по теме диссертации изложены в 9 печатных изданиях, из которых 4 — в периодических научных журналах, индексируемых Web of Science и Scopus [4—7], 5 — в тезисах докладов [8—12]. Получены 3 свидетельства о государственной регистрации программ для ЭВМ [1—3].

Объем и структура работы. Диссертация состоит из введения, 5 глав и заключения. Полный объём диссертации составляет 124 страницы, включая 21 рисунок и 14 таблиц. Список литературы содержит 156 наименований.

Во введении сформулированы критерии, показана актуальность и новизна исследования, описана теоретическая и практическая значимость, обозначена цель и задачи исследования.

В первой главе описана методология статистической оценки частотных характеристик юридической лексики среди различных типов документов. Статистические данные, методология их получения и обработки являются важными компонентами дальнейшего анализа, создают основу описательных характеристик документов.

Во второй главе приводится набор признаков, характеризующих юридические документы по критериям понятности, проводится анализ их эффективности и предлагается методика применения для решения задачи классификации сложности. Расчет языковых характеристик является наиболее распространенным способом оценки понятности документов. Модели, основанные на данных характеристиках в дальнейшем сравниваются с алгоритмами, построенными на основе языковых моделей.

В третьей главе представлен гибридный метод оценки сложности, основанный на совместном применении языковых характеристик и больших языковых моделей. Использование языковых моделей является ключевым элементом методологии. Их эффективность в задачах анализа естественного языка была рассмотрена, в том числе, в задачах анализа пользовательских текстов - специфических данных с нетиповой лексикой.

В четвертой главе проводится сравнительный анализ сложности юридических документов различных подстилей и жанров, основанный на использовании гибридной семантической модели предсказания сложности.

В пятой главе приводится пример адаптации предложенной методологии для решения задачи анализа ответов на юридические вопросы в сфере налогообложения.

В заключении подведены итоги, сформулированы основные результаты диссертационной работы.

Основные научные результаты:

1. Формализация и разработка гибридной нейросетевой модели оценки сложности юридических текстов. Представлена в работе [4], см. разделы 2-6. (метод разработан лично автором диссертации).
2. Разработаны и адаптированы методы анализа текстовых данных на основе языковых моделей, см. работы [5; 6; 8—11] (автором диссертации разработаны методы и проведены вычислительные эксперименты).
3. Выявлены текстовые параметры, наиболее эффективно моделирующие сложность юридических текстов, см. работы [10; 12] (автором диссертации разработаны методы тестирования и проведены вычислительные эксперименты).
4. Проведена оценка эффективности моделирования частотных зон, в контексте оценки сложности текстов, см. работу [11] (автором диссертации проведены вычислительные эксперименты).
5. Проведен анализ сложности юридических текстов различных жанров, см. работу [7] (автором диссертации проведены вычислительные эксперименты).
6. Разработаны программные методы адаптации языковых моделей для решения задач анализа нестандартных текстов [1—3] (автором диссертации разработаны методы и программная реализация).

Основные положения, выносимые на защиту:

1. Комплекс современных гибридных нейросетевых методов на основе больших языковых моделей в информационном обеспечении данными правовых процессов.
2. Методологические подходы для сбора, статистической и семантической обработки юридических текстов различной природы.
3. Методологические основы адаптации больших языковых моделей в задаче определения понятности юридических текстов.
4. Комплекс программ для информационного обеспечения научно-исследовательской и опытной работы с русскоязычными текстовыми данными правовых процессов, включающий компоненты сбора, интеллектуального анализа и визуализации.

Глава 1. Моделирование частотных диапазонов лемм для оценки лексической сложности текстов на русском языке

1.1 Вступление

Данный раздел посвящен проблеме формирования сводного частотного списка лемм на основе частотных списков крупных российских корпусов. Такой список можно использовать для оценки лексической сложности русских текстов (например, можно будет оценить количество низкочастотных, т.е. незнакомых, слов текста и использовать эти значения в формулах читаемости). Такой список должен содержать интерпретируемые значения частот, которые позволят нам разделить список частот на полосы и различать высокочастотные, среднечастотные и низкочастотные леммы.

Существует достаточно давняя традиция применения методов оценки читабельности к текстам на русском языке; обзор см. в [13]. В частности, используются метрики читабельности, то есть формулы, в переменные которых входит количество сложных слов. Сложные слова можно понимать как длинные (многосимвольные или многосложные) единицы, так и как незнакомые единицы.

Хотя, как отметил K. Collins-Thompson, «списки слов, используемые в измерениях читаемости на основе словаря, таких как Dale-Chall, могут рассматриваться как упрощенная языковая модель» [14], см. также [15], использование таких формул является распространенным методом оценки сложности документа. В настоящее время он используется в сочетании с другими, более сложными методами, подробнее см., например, [16]. Точнее, количество сложных (длинных, незнакомых/редких/низкочастотных) слов текста или средняя длина слов в буквах или слогах используется в различных моделях классификации текста как один из многих признаков, см., например, [17]. Понятно, что, за исключением некоторых особых случаев, применение критерия знакомых слов трудно или невозможно реализовать без использования информации о частоте слов.

1.2 Частота слов как параметр оценки сложности текста

По данным [18], частота слов тесно связана как с фактической сложностью слова (измеряемой тем, насколько хорошо читатели могут выбрать правильное определение слова), так и с трудностью его чтения.

Исследования сложности русского текста для носителей языка или изучающих второй язык также показывают, что лексические характеристики, включая информацию о частоте слов и/или включение в словарные списки для каждого уровня CEFR («лексические минимумы»), успешно предсказывают сложность. Например, по данным [19], именно эти характеристики показали наибольшую корреляцию со сложностью. В [20] метрики, основанные на лексических признаках (в том числе частотности слов, средней частотности существительных и т. д.), оцениваются как достоверные, см. также [21; 22].

Информация о частоте может применяться различными способами. В качестве мер лексической сложности можно использовать среднюю абсолютную частоту слов или среднюю логарифмическую частоту [23], общую частоту содержательных слов [20] и т. д. Кроме того, при оценке сложности текста можно учитывать количество слов, не вошедших в списки высокочастотных слов, подробнее о более сложных моделях см. [24].

Частоту леммы можно оценить с помощью частотных словарей или репрезентативных корпусов. Данный раздел посвящен проблеме общеязыкового частотного моделирования на основе данных крупных российских корпусов.

1.3 В поисках общеязыковой частоты

По мнению K. Collins-Thompson, «широко используемым признаком лексической сложности слова является, таким образом, относительная частота этого слова в повседневном использовании, измеряемая его относительной частотой в большом репрезентативном корпусе или его присутствием/отсутствием в список справочных слов» [14]. Для оценки общеязыковой частоты слов следует использовать некий «общеязыковый корпус», см. исследования по проектированию и балансировке корпусов и репрезентативности корпусов, например, [25].

Как указано в [26], репрезентативный корпус «может содержать примерно 90% разговорного языка».

В [24] эта проблема учета фактической компетентности носителя языка также обсуждается, ср.: «списки частот, принятые в этих исследованиях, были в основном взяты из письменных корпусов. Разговорный язык редко принимался во внимание при составлении списков частот. Это сопряжено с риском того, что значения частоты не будут точным представлением фактического языкового опыта читателя и, следовательно, будут неоптимальными для прогнозирования легкости восприятия и извлечения информации». Соответственно, при моделировании общеязыковой частоты русского языка было бы разумно придавать больший вес значениям частот, полученным из разговорного корпуса (например, Корпуса разговорного русского языка в Национальном корпусе русского языка).

1.4 Методы моделирования общеязыковых частот и диапазонов частот

Исследования эффекта частоты слов показывают, что высокочастотные слова обычно воспринимаются и производятся эффективнее и быстрее, чем низкочастотные, см., например, [27].

Между тем, при использовании классических методов прогнозирования сложности текста с использованием частотной информации, усредняя по всем значениям частот, то вклад низкочастотных слов становится минимальным [24]. Поэтому стоит задача выявления полос частот, в которых явно показаны высокочастотные, низкочастотные и среднечастотные единицы.

Для разделения полос используются различные пороговые значения (для частот или рангов). Условное пороговое значение для низкочастотных слов в корпусе из 100 миллионов слов составляет 5 ipm (элементов на миллион) [28]. Для рангов также используются разные пороговые значения. Высокочастотными единицами являются слова рангом до 2000 [29][60]; среднечастотные единицы – это слова рангами от 2000 до 8000–9000 [29][70]. Редкими единицами в Новом частотном словаре русского языка являются леммы рангом 10 000 и выше [28][229]. Весь список частот можно разбить на квартили (например, в

[30] слова из нижнего квартиля ранжированного списка частот считаются низкочастотными); Для этой цели также можно использовать проценты, см. [31].

1.5 Источники данных о частоте

В этой главе сравниваются списки частот, полученные из трех крупных веб-корпораций: ruTenTen11 [32; 33], Araneum Russicum III Maximum [34; 35],, Taiga [36] и Новый частотный словарь русской лексики на основе данных Национального корпуса русского языка [37; 38].

Списки частот были получены с сайтов корпусов или от создателей корпусов. Список возможных комбинаций получается с помощью НКРЯ. Для однобуквенных лемм проводился отдельный поиск. Данные представлены в Таблице 1.

1.6 Методы сравнения частотных списков

Существует несколько способов сравнения списков частот и методов измерения расстояния между ними. В частности, существуют меры, основанные на геометрических понятиях (евклидово расстояние, манхэттенское расстояние, косинусное расстояние и т. д.), меры, основанные на известных статистических тестах и процедурах (меры на основе хи-квадрата, логарифмическое правдоподобие, ρ Спирмена и т. д.), теоретико-информационная мера неопределённости, мера расстояния по ключевым словам Simple Maths) и другие, см. [39—41] и многие другие. Были выбраны три меры, которые указывают на различия между списками частот с разных точек зрения (сравнение рангов лемм, значений относительных частот или оценка перекрытия между списками).

Сначала был применен ранговый корреляционный анализ, вычисливший значения коэффициентов ранговой корреляции Спирмена и Кендалла для пар списков частот. Списки сравнивались с помощью пересекающихся лемм, уравнивающих их длину.

Таблица 1 — Источники частотных данных

Корпус	Состав	Размер	Число лемм	Анализатор
RNC (NFDR)	genre-balanced RNC subcorpus	91,982,416 граф. слов	52,138 с более чем 37 вхождениями	Mystem
ruTenTen11	Интернет: новостные и коммерческие сайты, блоги, социальные сети.	18 млрд. токенов	457,473 лемм с более 5 вхождениями	Treetagger
Araneum Russicum III Maximum	Интернет: новостные и коммерческие сайты, блоги, социальные сети.	15,961,200,372 слов	8,893,947 лемм с более чем 5 вхождениями	Treetagger
Taiga	Интернет: статьи из литературных журналов, наивная поэзия, новости популярных новостных сайтов и другие тексты	почти 5 млрд слов	2,988,610 лемм с более чем 1 вхождением	UDPipe

Во-вторых, были применены два показателя перекрытия, рассмотренные в [42] (“Coverage” и “Enrichment”). Показатель охвата рассчитывается по формуле:

$$Coverage(X,Y) = \frac{N1 \cap N2}{N1}$$

где X, Y — корпуса, $N1$ — количество лемм с абсолютной частотой, большей или равной заданному пороговому значению, в корпусе X, $N2$ — количество лемм с абсолютной частотой большей или равной заданному значению среза в корпусе Y. Мера Enrichment рассчитывается по формуле:

$$Enrichment(X,Y) = \frac{M2}{M1}$$

где $M2$ — количество лемм с частотой выше порога в корпусе Y и ниже порога в корпусе X, $M1$ — количество лемм с абсолютной частотой ниже порога в корпусе . В качестве порогового значения, согласно [42]), использовалась абсолютная частота в 20 случаев. Это так называемый «порог Синклера». Этот (произвольный) порог был выбран под влиянием утверждения Дж. Синклера о том, что опытному лексикографу потребуется не менее 20 вхождений однозначного слова, чтобы дать описание его поведения, см., например, [43][818].

В-третьих, была применена мера «Сумма минимальных частот» (СМЧ), предложенная А. Я. Шайкевича в [44], см. также [45]. СМЧ рассчитывается по формуле:

$$SMF(X,Y) = \frac{\sum_{min}(pX_i,pY_i)}{\sum_{0.5}(pX_i,pY_i)}$$

где pX_i — относительная частота появления леммы в корпусе X, pY_i — относительная частота появления леммы в корпусе Y.

1.7 Результаты сравнения

Рассматриваемые списки частот не подвергались какой-либо специальной предварительной обработке. В Таблице 2 представлены результаты применения рангового корреляционного анализа.

Коэффициент ранговой корреляции ρ принимает значение больше 0,7 только в паре ruTenTen11-NFDR ($\rho = 0{,}828$). Это можно объяснить тем, что эти

Таблица 2 — Значения ρ Спирмена и τ Кендалла

Spearman's ρ				**Kendall's τ**			
X/Y	ruTenTen	Taiga	NFDR	X/Y	ruTenTen	Taiga	NFDR
Araneum	0.033	0.081	0.223	Araneum	0.022	0.006	0.157
ruTenTen		0.071	0.828	ruTenTen		0.048	0.648
Taiga			0.095	Taiga			0.065

списки самые короткие и не содержат очень длинных низкочастотных хвостов. В парах веб-корпусов значения коэффициентов корреляции не превышают 0,3, то есть различия в ранжировании между этими корпусами значимы.

В Таблице 3 показаны результаты сравнения с использованием показателей Coverage и Enrichment. Coverage — это мера доли слов, для которых «достаточно» информации в корпусе X и «достаточно» информации в корпусе Y [42]. Другими словами, это «(очень грубая) мера того, насколько X «заменяем» на Y». Enrichment позволяет оценить долю слов среди тех слов, которые засвидетельствованы в корпусе X и для которых недостаточно информации в корпусе X, но достаточно информации в корпусе Y.

Таблица 3 — Значения мер перекрытия, порог= 20^{10}

Coverage				**Enrichment**			
X/Y	Araneum	ruTenTen	Taiga	X/Y	Araneum	ruTenTen	Taiga
Araneum		53	51.5	Araneum		0.9	0.2
ruTenTen	7.8		23.1	ruTenTen	3.4		1.9
Taiga	4.6	14.1		Taiga	13.9	0.2	

При интерпретации представленных значений метрик следует учитывать, что меры способны оценивать соотношение списков частот как X/Y или как Y/X. Показатель Coverage имеет наибольшее значение для пар Araneum (X)-ruTenTen11 (Y) (53) и Аранеум (X)-Тайга (Y) (51,5); пропорция показывает, что только около половины слов выше границы в Araneum находятся также выше границы в ruTenTen11 и Taiga. Таким образом, словари сравниваемых веб-корпусов существенно различаются. Значения Enrichment позволяют оценить, насколько списки частот способны дополнять друг друга. Наибольшее значение имеет пара Taiga—Araneum (13,9). Таким образом, если рассматривать весь рассматриваемый диапазон частот, то использование различных веб-корпусов не столь необходимо.

В целом оценка перекрытия позволяет сделать вывод, что списки частот не взаимозаменяемы и при составлении сводного списка частот лемм следует использовать все сравниваемые списки частот.

Наконец, в Таблице 4 показаны результаты сравнения всех четырех списков с использованием показателя SMF. Эта мера сравнивает относительные частоты всех пересекающихся элементов (лемм) в списках попарно.

Таблица 4 — Значения меры SMF

X/Y	ruTenTen	Taiga	NFDR
Araneum	0.056	0.024	0.264
ruTenTen		0.116	0.756
Taiga			0.197

Особое внимание следует уделить результатам сравнения веб-корпусов с НКРЯ. Высокое значение мы наблюдаем в паре НКРЯ—ruTenTen11 (SMF=0,756). Ранее было показано, что коэффициенты ранговой корреляции для этой пары также принимают наибольшее значение из наблюдаемых значений. Значительно менее похожи НКРЯ и Araneum (SMF=0,264), НКРЯ и taiga (SMF=0,197). Это также можно объяснить тем, что в списках частот Araneum и Taiga присутствуют длинные хвосты низкочастотных единиц.

Таким образом, применив три меры, было обнаружено, что между списками существуют значительные расхождения в рейтинге и относительной частоте. Использование показателя Coverage показало, что списки частот ни в коем случае не являются взаимозаменяемыми. Поэтому ни один из рассматриваемых корпусов не может быть исключен при составлении сводного частотного списка.

1.8 Сравнение по диапазонам частот

Для более детального сравнения списков частот по разным диапазонам частот список НКРЯ был разделен и ранжирован на 4 равные части, затем с использованием рангов были сформированы 4 случайные выборки (содержащие по 20 лемм из каждого квартиля). Для каждой леммы из 4 случайных выборок значения относительных частот присваивались по всем сравниваемым спискам.

Видно, что даже для лемм из верхнего квартиля имеются существенные различия в значениях ipm по разным корпусам. Итак, диапазон значений ipm для самой часто встречающейся леммы в выборке (существительное «центр») составляет 390,80.

Важно, что общий разброс значений IPM значителен. В НКРЯ присутствуют леммы с относительными частотами от 35 801,8 (союз «и») до 0,4 ipm, в Taiga — леммы с частотой от 18 710,7 (предлог «в») до 0,0017 ipm. Значительное количество лемм имеют частоты $<$1 ipm. Например, частотный список Taiga объемом 2 988 608 строк содержит всего 28 500 лемм с частотой $\geqslant$1 ipm (а это менее 1/100 всего списка). Наблюдаемая доля редких слов является следствием закона Ципфа.

Из-за широкого диапазона значений наблюдаемые значения относительной частоты трудно интерпретировать. Кроме того, не существует надежных порогов, разделяющих высокочастотные, среднечастотные и низкочастотные слова. Между тем полезно иметь удобный способ назначения лемм определенным диапазонам частот.

Поэтому (вслед за Chen [24]) используется подход Van Heuven [46], где предложена новая мера частоты «значение Ципфа». Значение этой меры рассчитывается по формуле

$$Zipf - value = log_{10}(ipm \times 1000)$$

Мера имеет следующие преимущества:

– Используется логарифмическая шкала.
– Значения легко интерпретировать.
– Шкала позволяет отделить среднечастотные слова от высокочастотных и низкочастотных.
– Значения Ципфа легко вычислить, зная значения IPM.

Обсуждаемый подход не является единственно возможным. [47] предлагает другую логарифмическую меру частоты «FClass», где $freq(max)$ — абсолютная частота наиболее частого слова (MFW) в конкретном корпусе, $freq(w)$ — абсолютная частота слова в конкретном корпусе, для которого рассчитывается значение меры).

$$FClass(w) = log_2 \frac{freq(max)}{freq(w)}$$

Мера FClass также имеет небольшой диапазон значений. Например, лемма «субпопуляция» из нижнего квартиля списка частот НКРЯ примет значения FClass, равные 16 и 21 (см. Таблицу 5).

Таблица 5 — Значения FClass

	$freq(w)$	MFW	$freq(max)$	FClass
NFDR	37	и "and"	3,293,765	16
Taiga	5	в "into"	11,076,749	21
Araneum	194	и "and"	563,822,183	21

Верхнее значение FClass можно оценить как $freq(w) = 1$, диапазон значений меры для сравниваемых корпусов — [0;22], или [0;23], или [0;29], см. Таблицу 6.

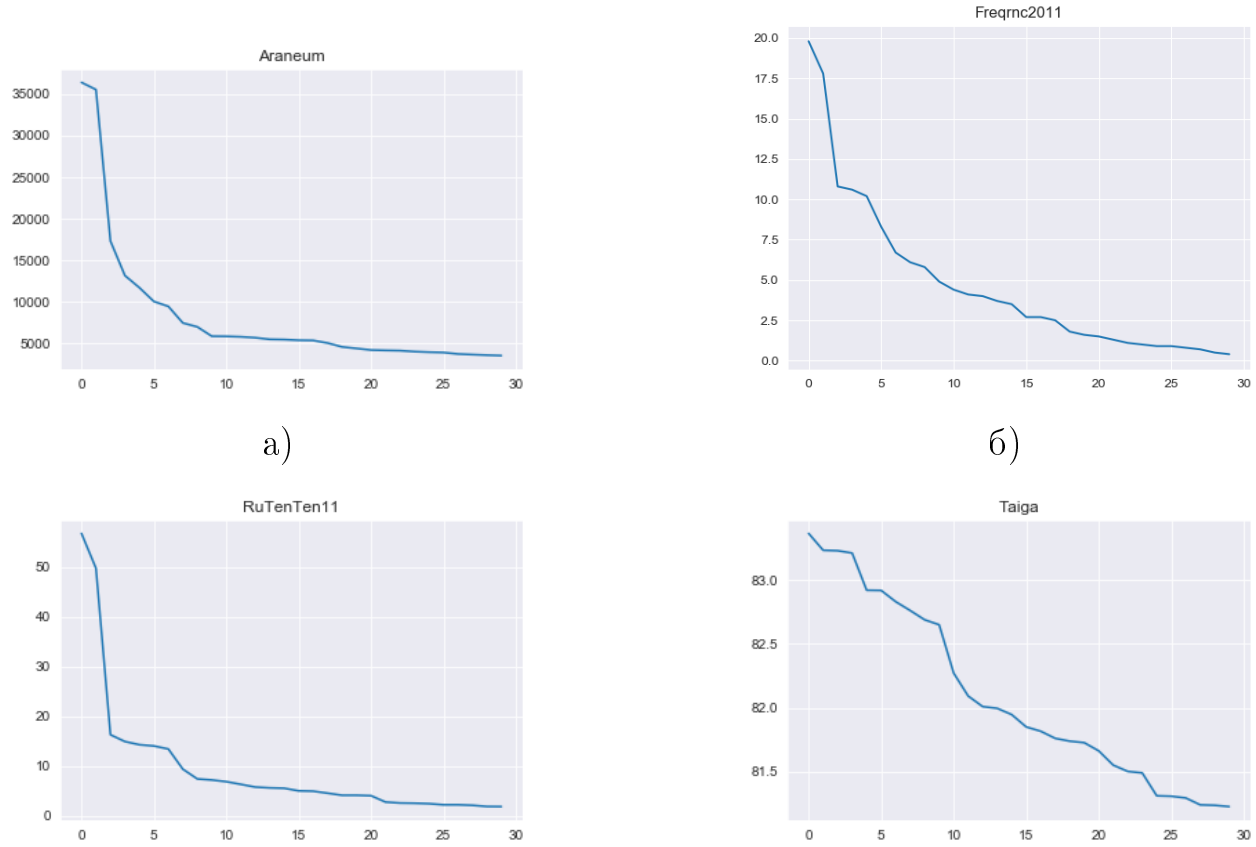
Таблица 6 — Максимальные значения FClass

	$freq(w)$	$freq(max)$	FClass
NFDR	1	3,293,765	22
Taiga	1	11,076,749	23
Aranuem	1	563,822,183	29
ruTenTen	1	503,894,565	29

Диапазон значений FClass превышает диапазон значений Ципфа. Шкала FClass не похожа на типичную шкалу оценок[48]. Соответственно, интерпретация значений Ципфа становится более простой задачей.

Сравниваемые списки частот, как показано ниже (см. рис. 1), подчиняются экспоненциальному закону. Следовательно, значение Ципфа можно использовать в качестве меры частоты.

Следует отметить, что лемматизаторы присваивают формам русских глаголов разные леммы, ср. превращаться (Pf) — превращаться (Impf), об этой проблеме см. Ляшевскую [28]. Это одна из причин расхождений между списками частот. Лемма конвертироваться присутствует во всех списках частот, но в списке Taiga конвертируется (Pf) имеет ipm=0,49, а лемма конвертироваться (Impf) имеет ipm=55,36, что гораздо ближе к значениям, демонстрируемым другими корпусами. Аналогичные расхождения в значениях ipm наблюдаются для лемм взорваться (взрываться) и прибить (прибивать).



а) б)

в) г)

Рисунок 1.1 — Распределение частот

Список лемм второго квартиля можно комментировать так же, как и список лемм первого квартиля. В списке ruTenTen11 леммы подоспеть (Pf) «прийти вовремя» не нашлось, но была лемма подоспеть (Impf). Леммы из второго квартиля (три из которых имеют среднее значение Ципфа, равное 4, 16 имеют значение Ципфа, равное 3, 1 (окрылить) имеет значение Ципфа, равное 2) по большей части могут можно рассматривать как среднечастотные. Список лемм третьего квартиля также достаточно однороден: 15 из 20 лемм (75%) имеют значение Ципфа 3.

Некоторые низкочастотные леммы из нижнего квартиля невозможно найти ни в двух частотных списках из четырех (послепожарный, шине), ни в одном частотном списке (несолоно, экономразвитие, напряг, поубавить, промельк, субпопуляция). Этот факт можно объяснить ошибками лемматизации. Например, в частотном списке Araneum присутствуют репрезентации леммы роздых в различных падежах (кроме именительного падежа).

Соответственно, перед предварительной обработкой списков частот с целью формирования сводного списка необходимо решить, как быть с такими

явлениями, как роздыха, роздыху и т.п. По-видимому, таким явлениям следует присвоить нормированные формы, а частоты различных словоформ, относящиеся к одной и той же лемме, следует обобщить.

1.9 Выводы главы

В этом разделе сравнивались частотные списки, полученные из четырех русскоязычных корпусов. Целью было не само сравнение, а разработка методологии создания сводного списка частот и моделирования общеязыковой частотности. Представляется, что включение в такой список значения Ципфа сделает частотные данные интерпретируемыми, так как диапазон значений меры невелик (наиболее часто встречающиеся леммы будут иметь значения Ципфа, равные 7 и 8, наименее частотные леммы будут иметь значения Ципфа, равные 1 и 2).

Глава 2. Метрики сложности российских юридических текстов: выбор, использование, первоначальная оценка эффективности

2.1 Вступление

Юридические тексты (особенно – тексты законов) безотносительно правовой традиции и языка характеризуются как сложные, тёмные, запутанные и для неюриста непонятные, см. [49—52] и мн. др. Настоящий раздел посвящен описанию модели, разработанной для измерения объективной сложности правовых текстов на русском языке. Модель, основанная на 130 метриках, разработана с учётом опыта исследований языковой сложности (в том числе – сложности юридических текстов), стилеметрических исследований, а также экспериментальных работ в области восприятия правовых текстов.

Задачи определения сложности текстов решаются достаточно давно. В том числе существует традиция применения методов оценки сложности к русским текстам, обзоры см., например, в [13; 53]. Наряду с понятием «сложность» в литературе используется понятие «читабельность». Читабельность понимается как оценка текста, полученная с применением параметров, которые в [54] названы латентными, в частности, формул читабельности и мер лексического разнообразия. Латентные параметры поддаются измерению, хотя и не поддаются непосредственному наблюдению в форме отдельных языковых сущностей, присутствующих в текстах. Соответственно, сложность понимается как более комплексное явление, она оценивается и с обращением к скрытым параметрам, и с применением формально-статистических (поверхностных) параметров [12].

2.2 Мотивации выбора метрик

Сложность может пониматься как переменная, значение которой измеримо для любого (связного) текста на естественном языке. Модели оценки сложности развивались от простых (подразумевающих использование формул читабельности) к изощрённым (подразумевающим использование разнообраз-

ных метрик, обращающихся к лексике, морфологии, синтаксису, информации о частотности единиц текста и т. д.).

В представленной модели используются в том числе традиционные метрики сложности; прежде всего сказанное относится к категории базовых метрик и категории “формулы читабельности” (подробнее см. Раздел 3). Опыт, накопленный в работах по функциональной стилистике, стилеметрии и в психолингвистических исследованиях перцептивной сложности (трудности) при разработке модели также учитывался.

Если учитывать, что стилевая принадлежность текста коррелирует с его сложностью (то есть в некотором общем случае деловые и научные тексты сложнее новостных, публицистических, разговорных), включение стилеспецифичных метрик получит обоснование.

Например, в работах упоминается свойственный деловым текстам рост доли существительного и падение доли глагола в личной форме. Например, в [55] указано, что “употребление глагольных форм сводится до минимума в официально-деловом стиле, который отличается наиболее ярко выраженным именным характером речи”, см. также [56] и мн. др. Повышение доли существительных может объясняться по-разному.

Во-первых, принято упоминать о частотности в текстах официально-делового стиля (далее – ОДС) “глагольно-именных сочетаний” с “расщеплением сказуемого”, см. [55] и мн. др., то есть о конструкциях с лёгкими глаголами типа “оказывать содействие”, “производить замену”. В модели не только подсчитываются доли слов различных частей речи, но и учитываются вхождения конструкций с лёгкими глаголами.

Во-вторых, в литературе встречается суждение о частотности в текстах ОДС отглагольных номинализаций (безотносительно к их вхождениям в состав конструкций с лёгкими глаголами). Эта черта в составе модели учтена в словообразовательной метрике и (частично) в лексической метрике, учитывающей вхождения абстрактных слов.

В-третьих, увеличение доли существительного может объясняться через употребительность в ОДС неоднословных терминоподобных сочетаний типа “товарищество собственников жилья” [57]. Эта черта учтена в лексической метрике, подсчитывающей вхождения юридических терминов (в том числе неоднословных).

В-четвёртых, доля существительных растёт за счёт неоднословных производных предлогов, компоненты которых размечаются как существительные, ср. “в соответствии”, “в связи”. Эта черта также учтена в лексических метриках. Представляется, что все четыре объяснения релевантны. Пример показывает, что учёт работ по стилистике позволяет анализировать сложность более подробным образом. Практическая стилистика рекомендует не злоупотреблять пассивными конструкциями в некнижных стилях, см. [55], а также [58] и мн. др. В работах о восприятии правовых текстов показано, что пассивные конструкции труднее активных, см., например, [59].

Соответственно, в модели среди метрик категории “отдельные граммемы” присутствует доля словоформ в творительном падеже (т.к. творительный падеж кодирует агенса в пассивных конструкциях). Кроме того, среди синтаксических метрик имеется доля вхождений пассивного подлежащего главной или зависимой клаузы. Наконец, учитывается доля личных форм глагола на -ся, а также (в составе частеречных метрик) – доля полных страдательных причастий и доля кратких страдательных причастий.

Важно заметить, что экспериментальные работы о сложности (точнее, перцептивной трудности) демонстрируют, что диагностическая сила некоторых традиционных метрик сложности для измерения актуальной понятности по данным эксперимента невысока. Например, в [59] показана невысокая предсказательная сила формул читабельности. Показано также, что длина предложения в стимуле практически не оказывала влияния на то, насколько успешно испытуемые справлялись с экспериментальным заданием по перефразированию, и что предложения одинаковой длины могут сильно различаться по фактической понятности.

Таким образом, сопоставление выводов квантитативных исследования сложности текстов и выводов экспериментальных исследований позволяет смотреть на предсказательную силу метрик сложности более трезво. В то же время эффективность метрик может быть проверена тестированием.

2.3 Набор метрик

В модели для оценки сложности используется 130 метрик, разделённых на следующие категории:

1. базовые метрики;
2. формулы читабельности;
3. доли слов разных частеречных классов;
4. частотность лемм;
5. словообразование;
6. отдельные граммемы;
7. лексические и семантические признаки, неоднословные выражения, гипертекстовые связи;
8. синтаксические признаки;
9. оценки связности.

Модель предусматривает использование 28-ми базовых метрик. Их можно разделить на базовые квантитативные и базовые лексические. Первые нацелены прежде всего на измерение длины слов и предложений (ср. ASL — "средняя длина предложения в словах", ASW — "средняя длина словоформы в слогах", S — "среднее число предложений на 100 словоформ" и пр.). Базовые лексические метрики подразумевают подсчёт индексов лексического разнообразия, а также подсчёт долей гапаксов.

В модели используется 5 адаптированных для русского формул читабельности: формула Флеша-Кинкейда [60], SMOG, ARI, индекс Дейла-Чейл, индекс Колман-Лиау, см. [61].

22 метрики, учитывающие доли вхождений слов разных частей речи, разработаны с учётом различий между использованными в модели инструментами разметки. Для лемматизации, частеречной и синтаксической разметки использовался UDPipe (модель "ru-syntagrus") [62]. Для второго слоя более подробной частеречной разметки и морфологической разметки использован pymorphy2 [63]. Под влиянием [64] в модель введены: индекс аналитичности (отношение числа служебных слов к общему числу слов в тексте); индекс глагольности; индекс субстантивности; индекс адъективности; индекс местоименности; индекс автосемантичности (отношение числа значащих слов к общему числу слов; "незначащими" считаются все служебные слова и местоимения). Кроме того,

учитываются: отношение числа существительных к числу глаголов; доли сочинительных и подчинительных союзов; доли полных и кратких прилагательных; доли полных и кратких причастий; доля местоимений-существительных; доли предикативов, деепричастий, инфинитивов; доли числительных; доля частиц; доля однословных предлогов, а также доля форм компаратива.

Введены 13 метрик, обращающихся к представленности в текстах n-грамм частеречных тегов. Об эффективности метрик, учитывающих частеречную сочетаемость, см., например, [65]. Отдельно стоит прокомментировать биграммы вида “NOUN + NOUN”, триграммы вида “NOUN + NOUN + NOUN” и биграммы вида “NOUN + NOUN,*gent”. Их использование нацелено в том числе на выделение именных групп с несколькими генитивными аргументами, которые в литературе по стилистике эксплицитно оцениваются как трудные для восприятия, ср., например, цитату из [55]: «Затрудняет восприятие текста нанизывание одинаковых грамматических форм, которые последовательно зависят друг от друга <...>. Эпифора часто возникает при нанизывании форм родительного падежа, что обычно связано с влиянием официальноделового стиля» и следующий пример из Бюджетного кодекса РФ: «для обеспечения необходимой степени конфиденциальности рассмотрения отдельных разделов и подразделов расходов федерального бюджета и источников финансирования дефицита федерального бюджета Государственная Дума утверждает персональный состав рабочих групп <...>».

Добавлена предложенная в [66] “формула динамичности / статичности”, призванная отделить тексты, в которых описывается множество событий (“динамические тексты”) от текстов “статических”. Эта метрика хорошо противопоставляет деловые тексты текстам других стилей (тексты официально-делового стиля более “статичны”).

Использованы 9 метрик, учитывающих вхождения лемм с разной общеязыковой частотностью, принадлежащих 9-ти частотным диапазонам. Для подсчёта значений этой метрики на базе больших русских корпусов создан сводный частотный список лемм с индексами частотности Zipf value, см. [11]. Значение Ципфа в этом списке принимает значения от 0 (наиболее низкочастотные леммы) до 8 (высокочастотные леммы). При оценке сложности учитываются доли вхождений в тексты лемм каждого из девяти частотных диапазонов.

Для диагностики сложности под влиянием [57] введена одна словообразовательная метрика. При подсчёте значений этой метрики модель обращается к уровню лемм, учитывая леммы вида *ция, *ние, *вие, *тие, *ист, *изм, *ура, *ище, *ство, *ость, *овка, *атор, *итор, *тель, *льный, *овать (то есть подсчитывая вхождения некоторых отглагольных и отадъективных существительных, отглагольных прилагательных и производных глаголов). Заметим, что осложнённая когнитивная обработка производных слов по сравнению с непроизводными подтверждается в экспериментах на принятие лексического решения, см., например, [67].

17 метрик категории “отдельные граммемы” заслуживают подробного обсуждения. Род существительных учитывается, так как абстрактные существительные, употребительные в правовых текстах, часто среднего рода. Граммема родительного падежа хорошо диагностирует сложность, это известно из литературы вопроса, см., например, [57]. Творительный падеж кодирует агенса в пассивных конструкциях. Набор личных форм глагола стилеспецифичен и жанровоспецифичен.

Согласно литературе вопроса, в ОДС частотны формы 3-го лица, формы 2-го лица практически не встречаются, а формы 1-го лица употребимы в ограниченном наборе жанров [55]. 11 метрик категории “лексические и семантические признаки, неоднословные выражения, гипертекстовые связи” также обращаются к описанным чертам текстов официально-делового стиля. Среди метрик категории: доля средств текстового дейксиса, обеспечивающих связность; доля графических сокращений; доля аббревиатур; доля леммы “являться”; доля юридических терминов; доля абстрактных лемм; доля лексических показателей деонтической возможности и необходимости; доля неоднословных предлогов; доля неоднословных оборотов в функции союза или союзного слова; доля конструкций с лёгкими глаголами, а также доля указаний на федеральные законы типа “231-ФЗ” (метрика призвана учитывать гипертекстовые связи).

В 21-й синтаксической метрике учитываются:

– признаки, описывающие организацию отдельных синтаксических групп (именной группы – доля адъективных модификаторов имени; глагольной группы – доля наречных модификаторов предиката); признак, описывающий вхождения аппозитивных именных групп (“Appos”);
– признаки, показывающих наличие сочинённых рядов (будь то сочинённые клаузы или однородные члены предложения; имеются в виду

признак “Cc”, описывающий союзные средства, а также признак “Conj”, описывающий количество конъюнктов, в том числе вводимых бессоюзно);

- признаки, описывающие вхождения сентенциальных определений (причастий и причастных оборотов “Acl” и относительных клауз “Acl:relcl”), сентенциальных обстоятельств (деепричастий и зависимых клауз с личными формами глагола, “Advcl”), различных сентенциальных дополнений (“Ccomp”, “Xcomp”), а также так называемых конструкций с сентенциальным субъектом (“Csubj”, “Csubj:pass”); отдельно учитываются единицы, способные вводить зависимые клаузы (“Mark”);
- признак, описывающий вхождения клауз со связочными элементами (“Cop”);
- признаки, с разных точек зрения описывающие вхождения пассивных конструкций (“Aux:pass”, “Nsubj:pass”, “Csubj:pass”).

Наконец, 2 метрики связности оценивают количество повторов существительных в соседних предложениях и количество повторов граммем времени и вида у глаголов в личной форме (в соседних предложениях).

2.4 Тестирование модели

Для определения качества выбранных 130 метрик, их способности предсказывать сложность текстов произведены такие тесты и сравнения:

- классификация с использованием полученных метрик в качестве параметров
- классификация с использованием в качестве параметров векторов языковой модели

2.4.1 Тестирование на текстовом наборе “plainrussian”

Тесты проводились на стандартном текстовом наборе “plainrussian” И. Бегтина, включающем тексты, распределённые на группы по уровню образования

(с 3-го класса начальной школы до 6 курса вуза) [61]. Из-за ограниченного размера тестового набора (68 текстов) для тестирования данные были разбиты на 3 класса: “простые тексты” – до 6-го класса, “средние по сложности тексты” – с 6 по 11 классы, “сложные тексты” – тексты уровня высшего образования. Итоговое число документов для каждой группы: “простые” – 14, “средние” – 32, “сложные” – 22. В качестве тестовой модели классификации использован XGBoost [68].

2.4.2 Классификация с использованием в качестве параметров векторов языковой модели

Сравнение производилось с языковой моделью USE (Universal Sentence Encoder) [69] с использованием современной нейросетевой архитектуры “Transformer”, показавшей высокую эффективность в решении задач классификации текстов [5; 9]. Оно позволило получить представление об эффективности выбранных метрик в задаче классификации по сложности. Таким способом проверено качество кодирования сложности текстов в описанном подходе по сравнению с подходом, кодирующим тексты на основе выбранных 130 метрик, отражающих знания о естественном языке.

Модель тестировалась с предварительным разбиением на тестовую и тренировочную выборки с последующим подбором гиперпараметров с помощью библиотеки “Hyperopt” [70]. Для подбора параметров было обучено 1000 моделей с различными параметрами. Цитируемые выше показатели качества (см. Таблицу 7) приводятся для оптимизированной модели с использованием кросс-валидации [71] с разбиением данных на 10 групп. Этот подход даёт возможность показать результаты более объективно, учесть генерализацию модели для ранее не использованных данных, что особенно важно в случае работы с небольшими наборами данных.

Таким образом, метрики позволяют получить более точные оценки сложности текстов. Наиболее успешно выделяются “сложные тексты”, несколько менее успешно – “простые тексты”, наименее успешно – “тексты средней сложности”.

Таблица 7 — Оценки классификации в эксперименте с “plainrussian”.

USE кодировки			
Тип текста	Точность	Полнота	F-мера
Простой текст	0.506	0.583	0.524
Текст средней сложности	0.667	0.333	0.419
Сложный текст	0.634	0.736	0.679
Кодировки метриками			
Тип текста	Точность	Полнота	F-мера
Простой текст	0.778	0.806	0.775
Текст средней сложности	0.567	0.733	0.622
Сложный текст	0.849	0.778	0.811

2.4.3 Тестирование на текстовом наборе учебников обществознания

Вторая итерация тестов проводилась проводились на наборе учебников обществознания, распределённые на группы по классам общеобразовательной школы (5 – 11 классы) [72]. Данные также были распределены на 3 категории: “более простые тексты” – 5, 6, 7 классы, “средние по сложности тексты” – 8, 9 классы, “более сложные тексты” – 10, 11 классы. Итоговое число документов для каждой группы: “более простые” – 5, “средние” – 4, “более сложные” – 5, размер датасета – 716 тыс. слов, средняя длина документа – приблизительно 1200 строк (по предложению на строку).

Все документы были случайным образом разбиты на фрагменты длиной в 100 строк. Затем данные были размечены с использованием UDPipe и pymorphy2, для каждого фрагмента вычислены значения 130 метрик. После этого была выполнена классификация. В качестве тестовой модели классификации использован XGBoost [68].

Итоговые показатели качества для кодирования с использованием метрик приведены в Таблице 8. В описанных экспериментах получены данные об эффективности работы 130 метрик в задаче классификации по сложности. Тестирование проводилось на наборах данных, существенно отличающихся от наших. Между тем, некоторые метрики были целенаправленно разработаны для применения к текстам ОДС. В текстах других стилей, по крайней мере, некоторые признаки могут описывать редкие или сверхредкие явления.

Таблица 8 — Оценки классификации в эксперименте с учебниками

Кодировки метриками			
Тип текста	Точность	Полнота	F-мера
Простой текст	0.929	0.867	0.897
Текст средней сложности	0.793	0.920	0.852
Сложный текст	0.971	0.895	0.932

2.4.4 Эффективность отдельных метрик

Эксперимент с “plainrussian” показал, что в задаче классификации эффективны 72 метрики. В эксперименте с учебниками обществознания выяснилось, что для классификации важна прежде всего формула Флеша-Кинкейда, коэффициенты (константы) которой вычислялись как раз на датасете с учебниками обществознания его создателями [72], а также 94 других признака.

В число десяти наиболее эффективных метрик в эксперименте с “plainrussian” вошли: средняя длина словоформы в буквах, доля полных прилагательных, доля слов длиной 4 и более слога, доля словоформ в родительном падеже, доля прилагательных, доля биграмм тегов существительного и существительного в род. п., формула Флеша-Кинкейда, доля вхождений пассивного подлежащего главной или зависимой клаузы, формула динамичности / статичности и средняя длина предложения в слогах, см. Рис. 2.1.

На классификацию текстов учебников (см. Рис. 2.2) лучше других метрик сработали: формулы читабельности (FRE GL, SMOG, ARI), а также индекс именной лексики, доля неодушевлённых существительных, индекс Колман-Лиау, доля лемм с «хвостами» типа *ция, *ние, *вие, *тие, *ист (см. о них Раздел 3 выше), доля полных прилагательных, доля кратких прилагательных и доля адъективных модификаторов имени.

На Рис. 2.3 представлены метрики, эффективные в обоих экспериментах. Они ранжированы по суммарной значимости и отобраны так: вес каждого из элементов (т. е. метрики для определённого набора данных) не превышаает 70% от общей суммы. Среди них (в порядке убывания значимости): индекс Колман-Лиау, доля адъективных модификаторов имени, доля лемм с «хвостами», включающими определённые словообразовательные аффиксы, доля среднечастотных лемм (Значение Ципфа = 6), индекс глагольности, доля сред-

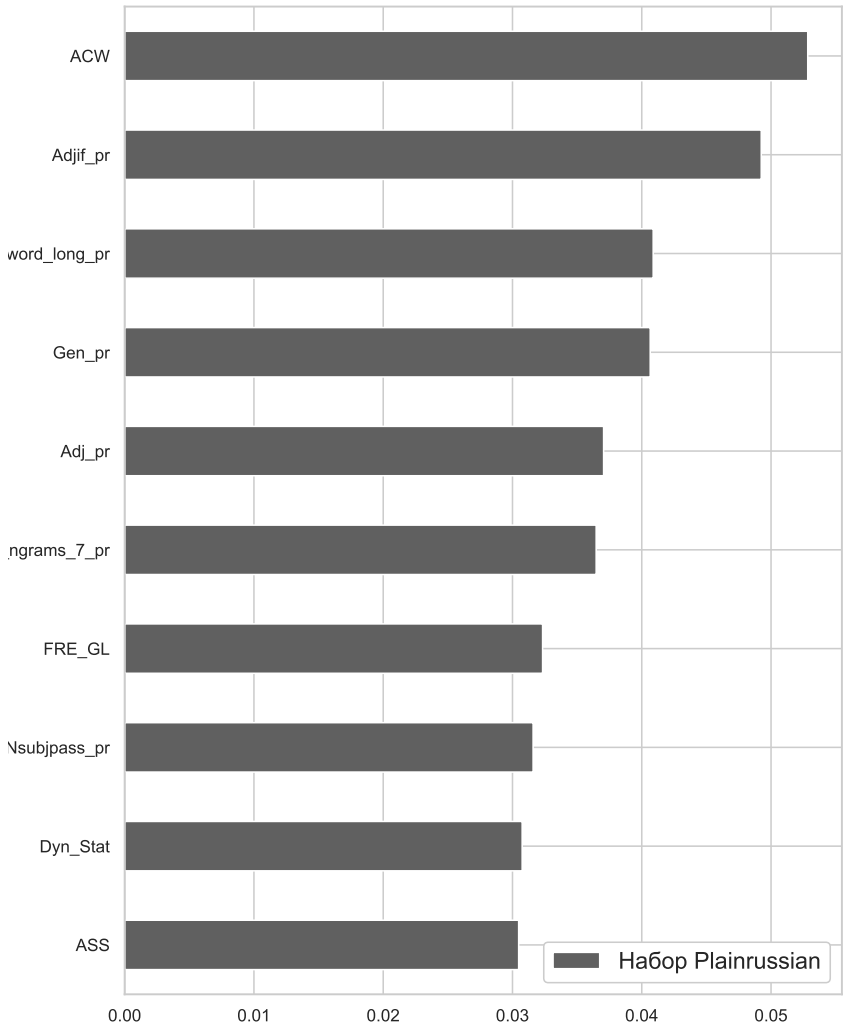


Рисунок 2.1 — Топ-10 метрик, “plainrussian”.

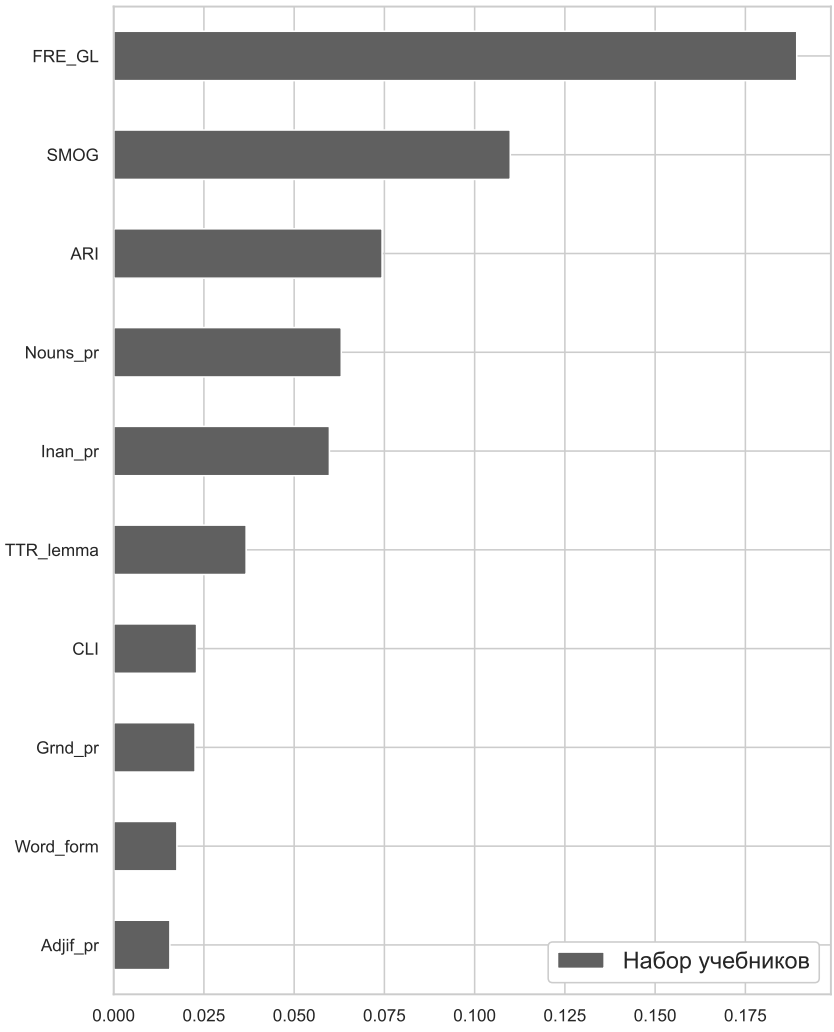


Рисунок 2.2 — Топ-10 метрик, учебники.

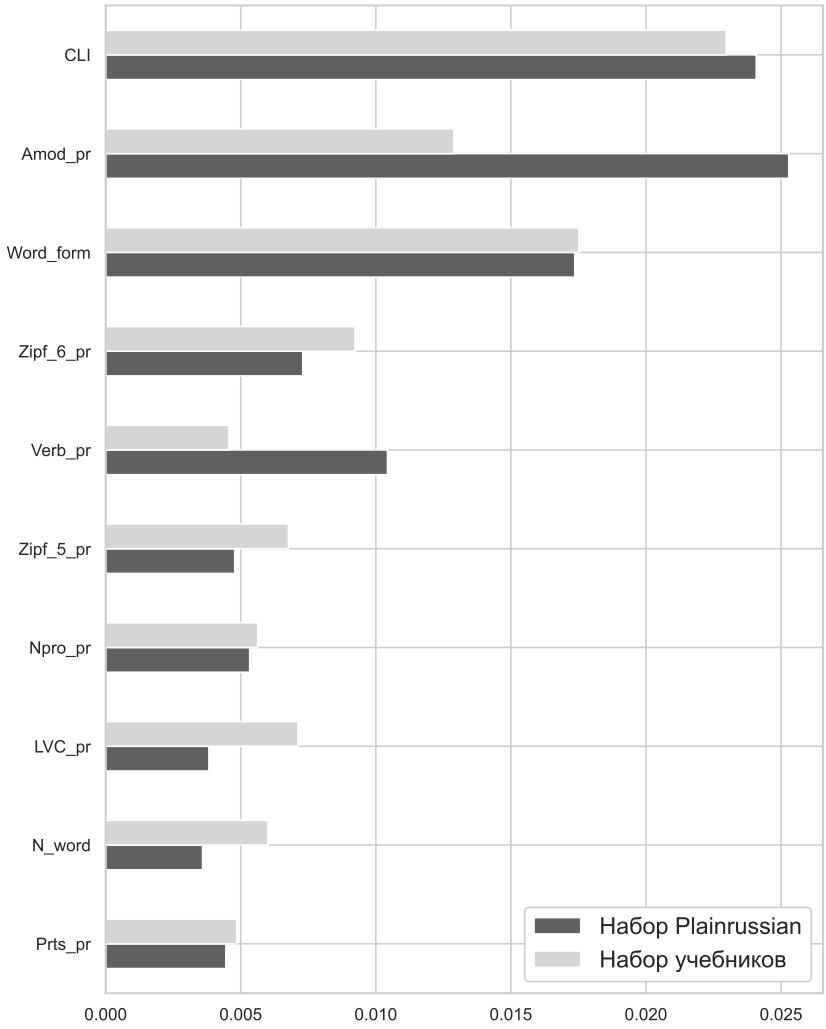


Рисунок 2.3 — Топ-10 метрик по суммарной значимости

нечастотных лемм (Значение Ципфа = 5), доля местоимений существительных, доля конструкций с лёгкими глаголами, количество словоформ и доля кратких причастий.

2.5 Выводы главы

В настоящей главе охарактеризована модель оценки сложности, в которой учитывалось 130 параметров, в том числе – стилеспецифичных (т.е. целенаправленно выделенных для русских текстов ОДС). Одновременно с этим выявленные лигвистические метрики показывают высокую эффективность в задаче представления текстов явными языковыми параметрами [8]. Продолжением работы, при этом, становится превращение модели, основанной на метриках, в гибридную. Использование метрик в совокупности с эффективным кодированием языка позволит оценивать сложность как по языковым параметрам, так и по неявным признакам. При тестировании модели была выявлена нехватка доступных русскоязычных текстовых наборов с оценками сложности (читабельности), содержащих изучаемые тексты. Был использован набор “plainrussian”, содержащий в общей сложности 68 текстов, а также существенно более обширный датасет из 14-ти учебников [73]. Таким образом, тестирование проводилось на наборах данных, существенно отличающихся от целевых.

Глава 3. Гибридная модель оценки сложности: разработка и применение для российских юридических текстов

3.1 Вступление

В этой главе описывается создание гибридной модели оценки сложности, которая включает 130 метрик в сочетании с нейросетевыми кодировками. Лингвистические особенности учитывают лексические, семантические, синтаксические свойства текста, его связность, а также последовательность частей речи, некоторые словообразовательные закономерности и общеязыковую частотность лемм. Кроме того, учитываются внутритекстовые ссылки на другие правовые документы (что особенно важно при анализе законов).

Использование метрик в сочетании с эффективными языковыми кодировками позволяет оценивать сложность как по лингвистическим параметрам, так и по латентным свойствам текстов. Исследование [74] показало успех такого подхода в его самом базовом варианте, т.е. добавление нейросетевого кодирования в качестве отдельного параметра для оценки сложности.

С точки зрения сложности лингвистические исследования сравнивают языки и диалекты; языковые регистры (или стили) и определенные единицы (особенно слова и предложения). Используется различие между так называемой «глобальной» и «локальной» сложностью [75]: первое направление исследований заинтересовано в изучении языков «как таковых»; второй измеряет сложность в конкретных языковых поддоменах и касается фонологической, морфологической, синтаксической, семантической, лексической и прагматической сложности. Межъязыковым сравнением занимаются типологи (О. Даль [76], Дж. Николс [77] и др.), социолингвисты и контактологи (П. Трудгилл [78], Дж. Маквортер [79] и др.). Перцептивную сложность изучают психолингвисты (см., например, [80]). Компьютерные лингвисты также участвуют в исследованиях сложности; обзор подходов см., например, в [81]. Существует довольно давняя традиция применения методов оценки сложности к русским текстам, обзор см. напр. [13; 82].

Интерес к сложности юридического языка вполне естественен. Lingua legis уже давно подвергается критике за многословие, избыточность, удлинение,

чрезмерное синтаксическое усложнение, архаичную лексику и необоснованные повторы, см., например, [83; 84].

Ряд исследований направлен на выявление особенностей юридических документов, обуславливающих их сложность, на разработку подходов к «движению простого языка», составление рекомендаций по «простому написанию». Популярные руководства, такие как [85], дают юристам практические советы, такие как «опускать лишние слова», «использовать глаголы для выражения действия», «предпочитать активный залог», «использовать короткие предложения». и т. д. Для российского научного сообщества проблемы, связанные с открытым языком, получили свое развитие совсем недавно.

Российские юридические тексты привлекли внимание исследователей сложности, которые, во-первых, были сосредоточены в основном на оценке законодательных документов, а, во-вторых, использовали только формулы читабельности или другие достаточно простые и немногочисленные меры.

Например, в [86] тексты решений Конституционного Суда изучались с помощью простой метрики оценки читабельности — формулы Флеша-Кинкейда, адаптированной И.В. Оборневой [87]. Д. Савельев и Р. Кучаков также занимаются исследованием сложности, см. [88; 89]. В цитируемых статьях авторы использовали только одну меру лексического разнообразия (TTR, значение которой зависит от длины текста, поэтому результаты применения метрики могут быть подвергнуты сомнению) и одну синтаксическую меру («Максимальная длина зависимости», расстояние между головой и зависимым по дереву зависимостей, рассчитываемое следующим образом: ”для каждого конкретного текста берется одно значение, максимальное для всех предложений текста”.

В новой книге [90] о сложности законодательных текстов выделено 9 факторов, среди них: ”доля глаголов в пассивном залоге”, ”доля глаголов по отношению к общему числу слов в тексте”, ”среднее количество слов в именной группе”, ”среднее количество придаточных придаточных, расположенных в предложениях после определяемого слова, на одно предложение”, ”среднее количество наречий причастных предложений в предложении”, ”среднее количество слов в предложении”, ”среднее расстояние между зависимыми словами в предложении”, ”среднее количество корней в предложении”, ”среднее количество слов на абзац”. К сожалению, авторы не объясняют в явном виде причины своего выбора параметров, которые впоследствии не всегда понятны читателю. Например, не совсем понятно, что имеется в виду под «значением». доля глаго-

лов в страдательном залоге», вероятно, только доля страдательных причастий (поскольку граммемы залога на слое морфологической разметки не закреплены за конечными формами глагола).

Так, авторы исследований русского юридического языка акцентировали внимание на сложности законодательных текстов и текстов судебных решений. Кроме того, для оценки сложности использовалось либо только формулы читаемости, либо другие, относительно немногочисленные меры.

В этой главе предлагается модель оценки сложности, основанная на сочетании различных лингвистических характеристик и нейросетевой языковой модели, обученной на больших данных и протестированной на трех разнообразных по жанрам юридических корпусах. Цель состоит в том, чтобы протестировать различные модели машинного обучения, обученные на наборе лингвистических признаков, и сравнить их с результатами, достигнутыми с помощью подхода глубокого обучения. Здесь можно предположить, что гибридный подход потенциально может обеспечить лучшее качество, чем любая отдельная модель, за счет использования как явного кодирования мер сложности, так и неявного представления языковых моделей.

3.2 Обзор литературы

Недавние разработки в области обработки естественного языка открыли новые возможности для разработки функций и представили новые контролируемые и неконтролируемые методы оценки сложности. В целом современные подходы можно разделить на две отдельные категории: традиционные подходы машинного обучения и модели глубокого обучения.

Классические подходы к машинному обучению обычно используют набор конкретных прописанных характеристик в сочетании с алгоритмом классификации. Внедрение классификационных моделей позволило превзойти традиционные оценки читабельности, такие как шкала Флеша-Кинкейда с использованием функций униграмм и наивного классификатора Байеса [91]. Более поздние наборы признаков были расширены и теперь включают более сложные лексические, грамматические и дискурсивные функции [92]. В [93] была предложена модель оценки читаемости для лиц, изучающих второй язык. Авторы использовали

лексико-семантические функции, функции дерева разбора (например, грамматические отношения), функции n-грамм и функции, основанные на дискурсе. Результаты показали эффективность этих функций и классификатора SVM. Аналогичные результаты можно найти в исследовательских работах [94] для текстов на немецком языке и [95], где авторы достигли наилучших результатов для итальянского языка, используя набор лингвистических функций в сочетании с классификатором Random Forest. [96] показала эффективность лингвистических функций для задачи оценки сложности текстов, написанных русскими, изучающими английский язык. Авторы сравнили классификатор случайного леса, классификатор k-соседей и логистическую регрессию и пришли к выводу, что классификатор случайного леса с добавленными векторами TF-IDF в качестве признака дает наилучший результат. Этот результат, в частности, показывает потенциал объединения лингвистических особенностей и моделей кодирования текста.

Подходы, основанные на нейронных сетях, можно разделить на три основные категории: общие подходы к глубокому обучению (такие как нейронные сети с прямой связью — FNN и сверточные нейронные сети — CNN), рекуррентные сети — RNN (включая долговременную краткосрочную память — подходы LSTM [97]) и языковые модели на основе Transformer. В [98] сравниваются традиционные машинные алгоритмы с общими подходами глубокого обучения, такими как FNN и CNN. Подходы на основе нейронных сетей превзошли традиционные подходы, такие как случайные леса, в большинстве тестов. Авторы провели эксперименты на трех наборах данных на русском языке, собранных из учебников. В [99] предлагается метод, позволяющий связать нейронные прогнозы сложности текста с лингвистическими свойствами данных.

Кроме того, некоторые модели используют нейронные кодировки в качестве представления документов вместо традиционных лингвистических функций, кодировок n-грамм или кодировок TF-IDF. Известно, что Word2vec [100], GloVe [101], FastText [102] обеспечивают общеприменимое высококачественное кодирование. В [103] сравниваются эти методы кодирования в сочетании с RNN, чтобы оценить сложность итальянского языка. Однако эти подходы могут быть ограничены с точки зрения применения к конкретной задаче. Нейронные сети на основе Transformer обходят эту проблему, предоставляя возможность дообучить модель для повышения ее эффективности при выполнении конкретной задачи. В [104] обсуждается применимость модели BERT

на основе Transformer [105] для задачи оценки читаемости немецкого языка. Авторы сравнивают регрессию случайного леса с лингвистическими функциями, модель на основе RNN с базовыми кодировками BERT и дообученный BERT для регрессии. Результаты показывают эффективность дообученной модели BERT.

Таким образом, предыдущие исследования демонстрируют потенциал как лингвистических функций, так и кодировок BERT. Различные исследовательские работы показывают неубедительные результаты по вопросу выбора модели для задач оценки сложности — классификация и регрессия случайных лесов, RNN и FNN, модели SVM — все они демонстрируют потенциал для достижения результатов высокого качества.

3.3 Данные

Из-за отсутствия доступных размеченных данных по теме оценки читабельности и сложности на русском языке конкретно юридических документов, в целях обучения и тестирования модели были собраны различные наборы данных. В исследованиях сложности русского языка, в частности, обычно используют данные учебников, см. [106]. Таким образом, данные учебников используются для обучения извлечению общих закономерностей сложности текста для языковой модели. Кроме того, эти данные были использованы для обучения окончательной гибридной модели и оценки ее качества. Для финального тестирования был использован набор юридических документов. Эти тексты используются для проверки эффективности окончательной модели данных, в частности, связанной с основной задачей данного исследования — оценкой сложности юридических документов.

3.3.1 Обучающие данные

Данные учебников были собраны с целью дообучения модели BERT и обучения окончательной гибридной модели. Данные состоят из блоков тек-

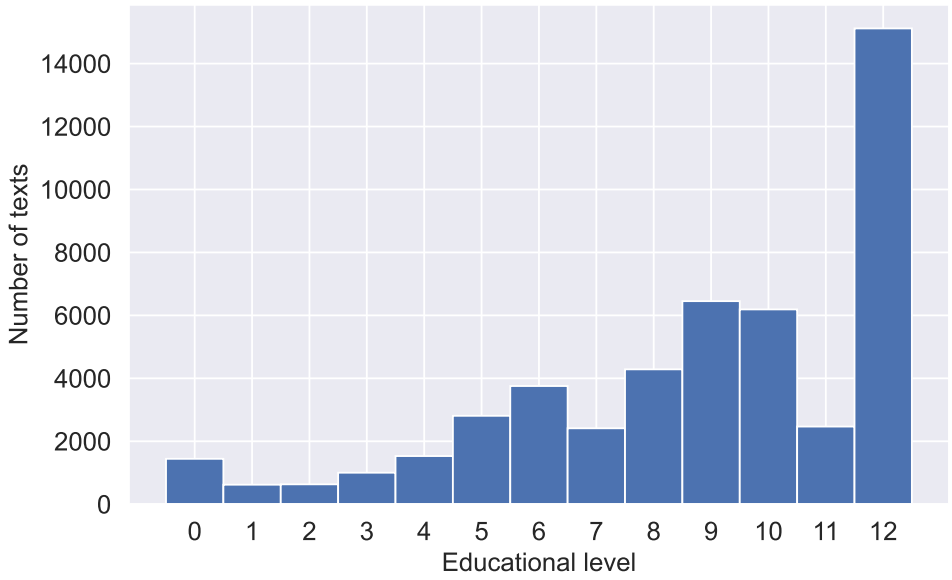


Рисунок 3.1 — Распределение текстов по уровням образования: 0 для текстов из книг для дошкольников, 1–12 для школьных учебников и 12 для текстов из книг университетского уровня.

стов, случайно выбранных из 1448 учебников на русском языке. Учебники были разбиты по параграфам, чтобы получить большой объем обучающих данных и предоставить языковую модель с сокращенными текстами. Ограничение размера текстовых блоков важно из-за того, что языковые модели на основе Transformer имеют максимальную длину входной последовательности обычно от 128 до 1024 токенов. Данные прошли дополнительную предобработку: были удалены оглавления, дополнительная конечная информация и любая нетекстовая информация (таблицы, изображения и т. д.). Также были удалены специальные символы (кроме знаков препинания), встречающиеся либо естественным образом в тексте, либо из-за ошибок кодировки текстовых файлов. Данные обучения собирались с учетом разнообразия и актуальности. Сборники учебников варьируются по сложности: от книг для дошкольников и начальной школы до книг для старших классов и университетов. В таблице 9 показаны статистические характеристики обучающих данных. На рисунке 3.1 показано количество текстов для каждого уровня образования в диапазоне от 0 для текстов дошкольного уровня, от 1 до 11 для школьного обучения и 12 для текстов университетского уровня. На рисунке 3.2 показаны темы и соответствующие объемы текстов.

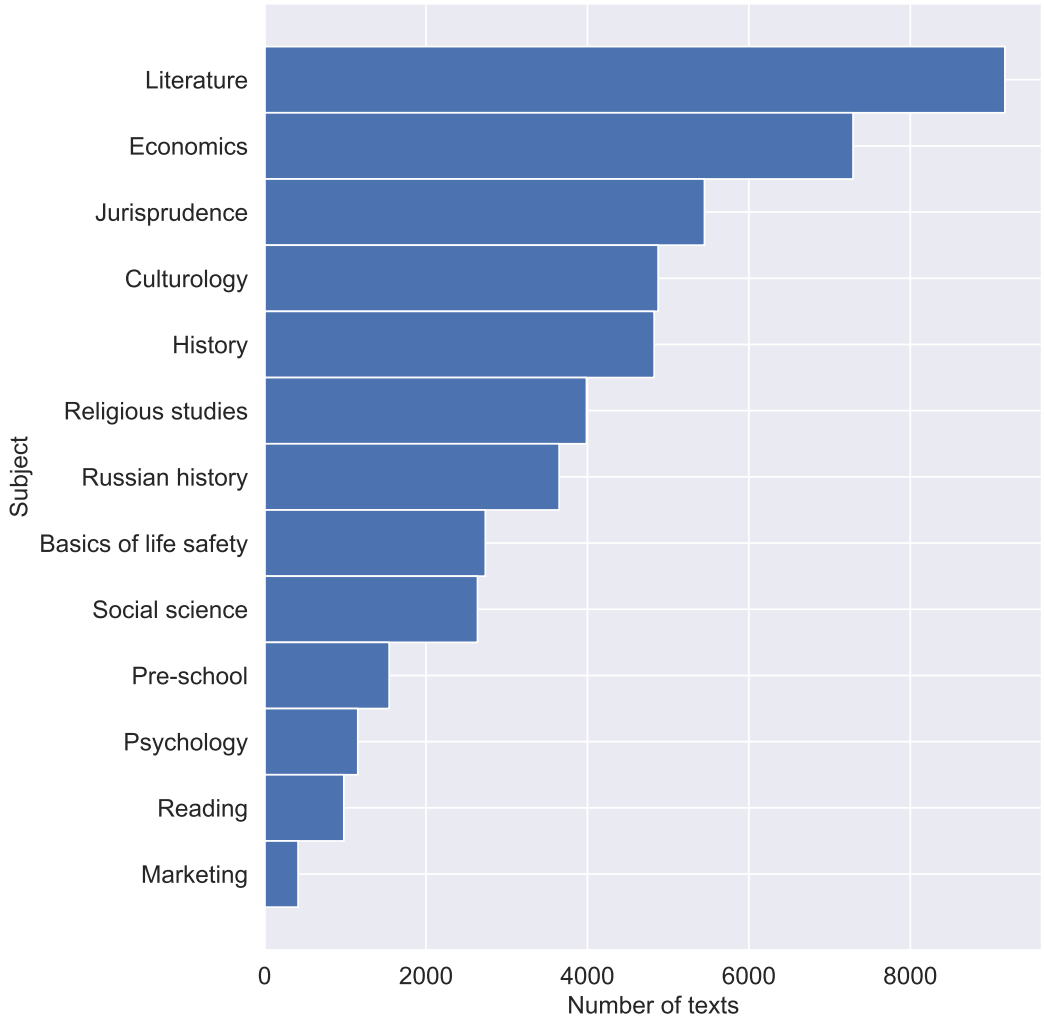


Рисунок 3.2 — Распределение текстов по дисциплинам

Дисциплины были выбраны из-за ожидаемого сходства с юридическими документами (т. е. в набор данных входят учебники по юриспруденции, общественным наукам, экономике) и из-за возможности представления образцов текстов на русском языке разного уровня сложности (т. е. в набор данных входят учебники по литературе, культурологии и истории).

Таблица 9 — Характеристики обучающих данных

	Всего	Среднее для текстового блока	Отклоенение
Предложение	526 935	11	7
Токены	9 939 730	204	151
Уникальные токены	7 012 687	144	97

3.3.2 Тестовые данные

В цифровом мире существует значительное количество российских юридических документов; они доступны, например, через правовые информационные системы «КонсультантПлюс»[107], «Гарант» [108]. Это дает возможность создавать обширные корпуса.

Данными тестирования предлагаются три юридических корпуса [10]. Во-первых, это корпус внутренних документов России "CorRIDA", состоящий из 1546 документов и содержащий 1784 тыс. токенов. Во-вторых, это корпус решений Конституционного Суда РФ "CorDec"объемом 3,427 тыс. токенов, в том числе 584 документа. В-третьих, это корпус законодательных документов CorCodex, содержащий 278 текстов кодексов, федеральных законов (всего 3227 тыс. токенов).

Известно, что синтаксические особенности хорошо прогнозируют сложность текста, см., например, [109]. Корпуса UD (Universal Dependencies - универсальных зависимостей) в последнее время все чаще используются для оценки морфосинтаксической сложности как при межъязыковом сравнении, так и при сравнении коллекций текстов на одном языке [110]. Поэтому в качестве основного инструмента разметки был выбран UDPipe. В качестве инструмента морфологического анализа использовался pymorphy2 [111]. При выборе предварительно обученной модели UDPipe использовалась статистика точности из [112] с выбранной моделью «russian-syntagrus».

После предварительной обработки выполнялась автоматическая лемматизация, морфологическая разметка и синтаксический анализ. Каждой словоформе был присвоен двойной тег части речи в терминах UDPipe и в терминах pymorphy2. Набор PoS-тегов pymorphy2 позволяет, в частности, различать 'ADJF' (полные формы прилагательных), 'ADJS' (краткие формы прилагательных), 'VERB' (конечные формы глагола), 'INFN' (инфинитивы), 'PRTF' (полная форма причастия), 'PRTS' (краткая форма причастия) и 'GRND' (наречное причастие). Это удобно для оценки сложности, в частности, потому, что существует положительная корреляция между количеством полных прилагательных (а также причастий и деепричастий) и сложностью и отрицательная корреляция между количеством конечных глаголов и сложностью, см. [113].

3.4 Лингвистические характеристики

Для оценки сложности российских юридических текстов было выбрано 130 характеристик. Лингвистические свойства русских официальных текстов (ср. понятия «официально-деловой стиль»), описанные в исследованиях по функциональной стилистике, а также учитывались признаки, способные разделить такие тексты из текстов других стилей при решении задачи автоматической классификации по стилям.

Все используемые метрики условно разделены на следующие категории:

1. основные признаки;
2. формулы читабельности;
3. слова разных классов частей речи;
4. n-грамм тегов частей речи;
5. общеязыковая частота лемм;
6. словообразовательные закономерности;
7. отдельные граммемы;
8. лексические и семантические особенности, многословные выражения;
9. синтаксические особенности;
10. оценки связяности.

3.4.1 Базовые метрики

Модель предусматривает использование 28 базовых метрик. Некоторые из них традиционно используются в задачах классификации текстов по сложности. Все базовые метрики можно разделить на «базовые количественные» и «базовые лексические». Первые направлены, в том числе, на учет доли длинных слов и длинных предложений («длинными словами» в модели считаются слова, состоящие из 4 и более слогов). Базовые лексические метрики подразумевают расчет индексов лексического разнообразия (простой TTR для словоформ и лемм; производный от метрик TTR "Yule's K" и "Yule's I", значения которых не зависят от длины текста) и расчет долей гапаксов (hapax legomena и hapax dislegomena).

3.4.2 Формулы читаемости

Использование формул читаемости — распространенный метод оценки сложности. Сейчас он используется в сочетании с другими методами (см., например, [114]) и встроен в различные текстометрические ресурсы. В описываемой модели используются пять формул: адаптированная формула Флеша-Кинкейда [115], адаптированная формула SMOG (Simple Measure of Gobbledygook), адаптированная формула для расчета индекса автоматизированной читаемости ARI, формула Дейла-Чейла, индексная формула Коулмана-Лиау [116]. Формулы были адаптированы Бегтиным с использованием текстового набора, включающего 68 документов, классифицированных по уровню образования (от 3-го класса начальной школы до 6-го курса высшего образования).

3.4.3 Слова различных классов частей речи

Метрики, учитывающие доли вхождений слов различных классов частей речи, разработаны с учетом различий между используемыми инструментами разметки — UDPipe и pymorphy2, то есть различий между наборами PoS-тегов. [111; 117]. Следуя [118], в модель были введены такие индексы:

- “индекс аналитичности” (отношение количества служебных слов к общему количеству слов);
- “индекс вербализации” (отношение количества глаголов к общему количеству слов);
- “показатель содержательности” (отношение количества существительных к общему количеству слов);
- “индекс адъективности” (отношение количества прилагательных к общему количеству слов);
- “показатель местоимения” (отношение количества местоимений к общему количеству слов);
- “индекс автосемантичности” (отношение количества содержательных слов к общему количеству слов).

Кроме того, использовалось соотношение количества существительных к числу глаголов; Отдельно рассматриваются вхождения кратких и полных прилагательных, кратких и полных причастий.

3.4.4 Частеречные N-граммы

Информацию о n-граммах PoS-тегов было решено использовать для анализа сложности под влиянием исследований по количественному анализу стиля [119; 120]. Там же была предложена так называемая «динамическая/статическая формула», позволяющая отделить «динамические тексты», описывающие последовательность событий, от «статических», содержащих описательные отрывки, подробнее см. [121]. Эта метрика позволяет успешно отличать официальные документы (они более «статичны»).

3.4.5 Общеязыковая частота

При оценке сложности принято учитывать длину слов текста и их «знакомость» читателю. «Знакомство» можно операционализировать через информацию об общеязыковой частоте текстовых лемм. В рамках модели для точного учета частотных данных на базе крупных российских корпусов был создан список частот. Этот список содержит около 1 миллиона лемм, распределенных по 9 частотным диапазонам с использованием значений Ципфа, см. метод [11]. Модель оценки сложности способна рассчитать долю лемм, принадлежащих каждому из 9 диапазонов частот, и различать высокочастотные, среднечастотные и низкочастотные леммы.

3.4.6 Словообразование

Производные слова, образованные с помощью аффиксов, обычно длиннее производящих. Кроме того, производные более сложны морфологически. Это усложняет восприятие производных слов, что подтверждается экспериментально, см. [122]. В модели словообразовательные данные извлекаются с уровня лемм, в каждом документе доля лемм с окончаниями типа **ция*, **ние*, **вие*, ** тие*, **ист*, **изм*, **ура*, **ство*, **ост'*, ** овка*, **атор*, **итор*, **тель*, **льный*, **овать*. Это позволяет учитывать употребление девербативных и прилагательных существительных, глагольных прилагательных и некоторых производных глаголов.

3.4.7 Граммемы

В модели используется 17 метрик, учитывающих, в частности: словоформы в родительном, творительном, дательном падеже, существительные среднего рода, глаголы 3-го лица, полные и краткие формы страдательных причастий, а также конечные формы глаголов с *-ся* .

3.4.8 Лексико-семантические особенности, многословные выражения

Список характеристик, оцениваемых через слой лемм или словоформ, выглядит следующим образом.

- доля тексто-дейктических выражений типа *настоящий*, *нижеследующий*; *вышеупомянутый* и т.д.;
- доля графических сокращений;
- доля буквенных сокращений;
- доля юридических терминов;
- доля абстрактных лемм;

- доля лексических показателей деонтической возможности и необходимости типа *запрещать'*; *противоправный*, *надлежащий* и т.д.;
- доля многословных предлогов типа *в соответствии с*;
- доля многословных выражений, используемых в качестве союзов или союзных слов, таких как *ввиду того что*; *вследствие чего*;
- доля легких глагольных конструкций типа *оказывать содействие*, *осуществлять подготовку*;
- доля внутритекстовых ссылок на законодательные акты, в частности, на федеральные законы типа *231-ФЗ* «Федеральный закон #31».

Для расчета значений соответствующих метрик применяется набор пользовательских словарей, то есть значение метрики рассчитывается как доля единиц, совпавших с единицей из словаря.

3.4.9 Синтаксические признаки

Высокая синтаксическая сложность — характерное свойство официальных текстов. В обширной литературе описаны параметры оценки сложности предложения, сложности предложения и сложности фразы. Актуальный обзор представлен в [123]. Важным исследованием в этой области является [124]. [74] использовало большое количество показателей синтаксической сложности.

В русском языке признаками сложности считаются, прежде всего, причастные и деепричастные придаточные предложения, сложные и сложноподчиненные предложения, см., например, [109; 125].

Очевидно, что возможности анализа синтаксического анализа ограничены форматом парсинга. Модель использует UDPipe для анализа зависимостей (подробности см. в разделе 3.1.2 выше), использует 21 синтаксическую метрику и учитывает, среди прочего: модификаторы именного предложения, модификаторы наречного предложения, различные дополнения к предложениям.

3.4.10 Связность

Для оценки референциальной связности использовалась мера “Cohes_1” (количество повторов существительных в соседних предложениях). Кроме того, была использована метрика “Cohes_2”, учитывающая количество повторов граммем времени и вида для конечных глаголов (также в соседних предложениях).

В конце стоит отметить, что некоторые параметры оценки сложности не являются независимыми друг от друга, в частности, согласно закону сокращений Ципфа, длина слова коррелирует с частотой слов, см., например, [126]. При этом репрезентация в текстах различных перечисленных выше признаков может иметь как положительную, так и отрицательную корреляцию с целевой сложностью.

3.5 Постановка эксперимента

Полученная модель состоит из трех основных модулей, как показано на рисунке 3.3. Процесс обучения осуществляется в два этапа. На первом этапе модель BERT на основе Transformer дообучается для получения начального прогноза сложности для каждого текста. Тексты дополнительно кодируются с использованием набора метрик, описанных в разделе 3.4. Начальные прогнозы сложности на основе языковой модели и кодировки функций на основе предопределенных метрик объединяются и передаются в модуль окончательного тестирования — выбор между различными моделями регрессии и классификации.

3.5.1 Предсказания языковой модели

Архитектура Transformer использовалась для ряда различных задач обработки естественного языка как в качестве отдельного подхода, так и как

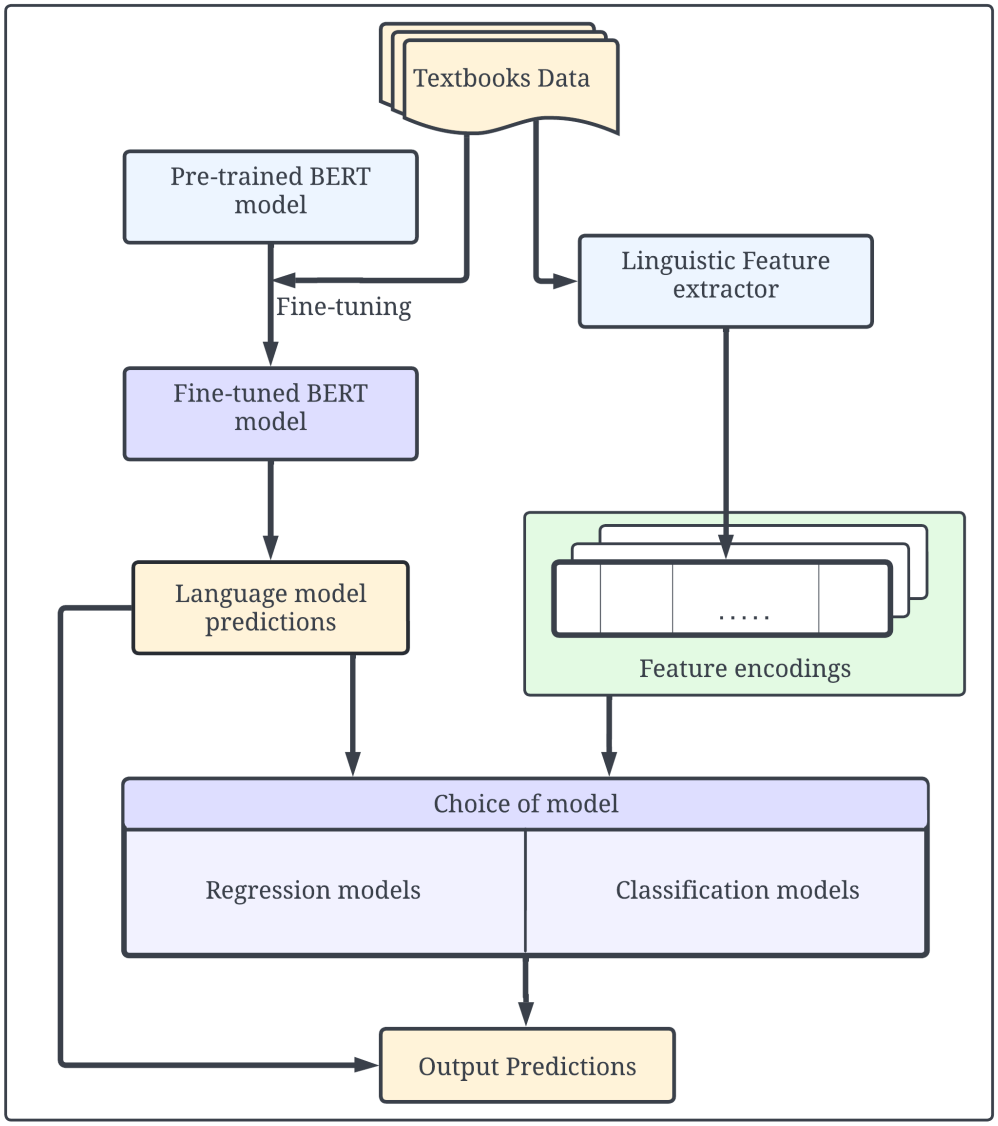


Рисунок 3.3 — Предлагаемый процесс обучения и тестирования, включающий три основных модуля: языковая модель, анализатор признаков и окончательная гибридная модель. Окончательная модель выводит как результат нейронной модели, так и окончательный результат гибридной модели.

часть более сложных комбинационных решений. Основная идея этого подхода заключается в замене повторяющихся слоев слоями внимания. Это привело к значительно более быстрому процессу обучения и лучшему использованию ресурсов благодаря возможностям распараллеливания, ранее невозможным для рекуррентных сетей и LSTM. По сути, Transformer — это быстрый и надежный метод языкового моделирования, который служит основой для других, более сложных и специализированных алгоритмов.

Bidirectional Encoder Representations from Transformer – BERT. Модель BERT улучшает эту идею, вводя двунаправленную архитектуру и процедуру дообучения. С момента своего создания трансферное обучение стало неотъемлемой частью большинства решений для анализа текста. Этот подход состоит из двух основных этапов, т.е. начальное предварительное обучение модели в крупномасштабном и универсальном наборе задач (предсказание следующего предложения и моделирование скрытых слов для BERT) и этап тонкой настройки, предназначенный для адаптации модели к конкретной задаче.

Было показано, что метод дообучения моделей на основе Transformer, предварительно обученных на крупномасштабных данных, обеспечивает высококачественное текстовое представление для различных задач NLP. Этот процесс выполняется путем добавления дополнительного линейного слоя в конце предварительно обученной модели и ее обучения в течение нескольких эпох. Интуиция этого подхода заключается в том, что исходная предварительно обученная модель изучает общие языковые шаблоны, а процесс дообучения позволяет модели изучать особенности, специфичные для конкретных задач [127].

В этом исследовании использовалась базовая версия RUBERT [128], полученная из библиотеки Huggingface [129]. Модель предварительно обучена для русского языка на данных, полученных из различных наборов данных социальных сетей. Исходная предварительно обученная модель состоит из 12 слоев, 768 скрытых единиц на слой и 12 элементов внимания.

Из-за большого количества категорий сложности в наборе данных и их упорядоченного характера можно предположить, что регрессионный подход может быть более применимым. Определив задачу как регрессию, можно достичь потенциально более качественных прогнозов в крайних случаях. В то время как классификация предсказывает один из результатов без учета их близости друг к другу, регрессионная модель может предоставить полезную информа-

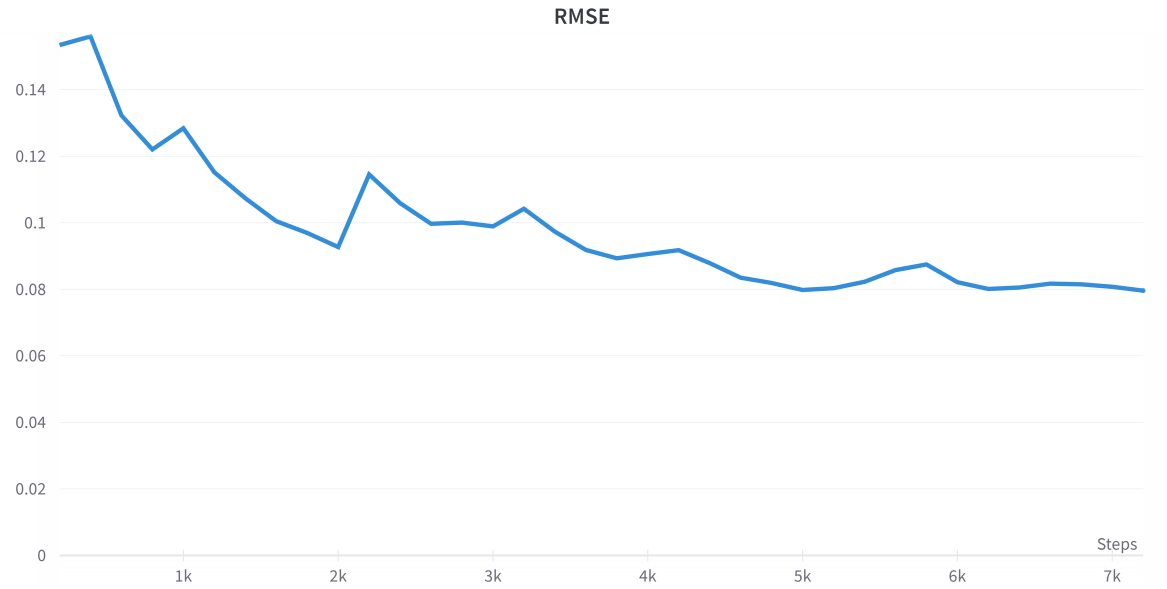


Рисунок 3.4 — Повышение качества при дообучении языковой модели, на что указывает метрика RMSE.

цию, делая прогнозы, которые ближе к реальным значениям, даже если они не полностью точны.

Этот подход использует стандартный процесс тонкой настройки. Он использует предварительно обученный токенизатор RUBERT для разделения текстовых блоков на токены и добавления специальных полей и токенов [CLS]. Затем кодировки передаются через модель до последнего уровня, где скрытое состояние токена [CLS] извлекается и передается через полносвязный слой с функцией активации гиперболического тангенса. Для тонкой настройки использовался оптимизатор AdamW [130] со скоростью обучения 2e-5, размером пакета 16, 3 эпохами и затуханием веса 1e-2. Модель оптимизирована для поиска наилучшего результата с точки зрения потерь RMSE для подмножества проверочных данных — 10% от исходных текстов. На рисунке 3.4 показано улучшение качества в процессе дообучения.

3.5.2 Комбинированный подход

Чтобы объединить лингвистические функции с языковой моделью, мы получаем выходные данные дообученной модели BERT и используем их как

функцию в сочетании с лингвистическими функциями. Это окончательное векторное представление передается в другую модель. [74] использует классификатор SVM для выбора окончательной модели из-за его простоты и частого использования в задачах, связанных с добавлением числовых признаков.

Дополнительно оценивается потенциал других типов моделей, в том числе регрессионных. При большом количестве классов сложности (в данном случае их 13 категорий) существует вероятность того, что регрессионные модели могут дать лучший результат благодаря их способности получать оценку сложности, а не прямому прогнозированию классов. Это может улучшить качество и удобство использования модели. Хотя модель классификации может спутать любой класс во время вывода, ошибки модели регрессии все равно могут быть близки к целевому значению.

Было проверено качество шести моделей: линейной регрессии, XGBoost [131] для регрессии, FNN для регрессии, SVM для классификации, классификации случайного леса и XGBoost для классификации. Линейная регрессия и классификатор SVM были выбраны для обеспечения базовой оценки качества с использованием простых подходов. Классификатор SVM также является моделью, обычно используемой для задачи оценки сложности. Регрессионная модель FNN представляет собой плотную нейронную модель, которая в данном случае состоит из 3 скрытых слоев, по 128 скрытых блоков в каждом. Модель была обучена с помощью оптимизатора Adam со скоростью обучения 1e-3. Случайный лес — это широко используемый ансамблевый подход, который обучает несколько более слабых деревьев решений на подмножествах данных и объединяет их в более сильный предиктор, уменьшая переобучение. Extreme Gradient Boosting или XGBoost — это библиотека машинного обучения с использованием дерева решений с градиентным бустингом (GBDT). Он использует метод, при котором вводятся новые модели для исправления ошибок, допущенных существующими моделями. Гиперпараметры этого алгоритма были настроены с использованием библиотеки Hyperopt [70] для построения 500 оценщиков для задач классификации и регрессии и поиска набора оптимальных параметров модели для каждого.

3.6 Результаты экспериментов

Для сравнения эффективности каждого метода используется набор метрик. Точность классификации измеряется как базовый процент правильных предсказаний. Для регрессионных моделей этот и все будущие показатели классификации определяются путем округления предсказаний до ближайшей категории. Точность для текстов университетского уровня (AUT) измеряет точность классификации текстов с максимальным рейтингом сложности. Он измеряется для обеспечения качества прогнозов для текстов более высокой сложности, предположительно составляющих большой объем юридических текстовых данных. Точность, полнота и f-мера рассчитываются с использованием средневзвешенного значения для каждого класса. Среднеквадратическая ошибка измеряется, чтобы найти разницу между прогнозами и истинными значениями в задачах регрессии. Более низкие значения указывают на более высокое качество. Для алгоритмов классификации прогнозы отображаются в пространстве от 0 до 1. Оценка R2 — коэффициент детерминации — это более простая оценка регрессии, обычно варьируется от 0 до 1, однако может быть сколь угодно хуже. В таблице 10 показаны результаты тестирования каждой модели.

Во всех случаях внедрение предсказаний BERT обеспечило улучшение по сравнению с моделями, обученными только на лингвистических особенностях. Почти во всех случаях результаты были улучшены по сравнению с базовыми предсказаниями BERT. Как показано в таблице, модель классификации XGBoost, обученная на лингвистических особенностях и прогнозах языковой модели, достигла наилучших результатов почти по всем показателям. Это справедливо даже для показателей, основанных на регрессии, что указывает на то, что неверные прогнозы были близки к реальным оценкам. Для регрессионных моделей внедрение предсказаний языковой модели обеспечило более существенное улучшение качества, причем наивысшее качество достигалось за счет трехслойной нейронной сети. Модель линейной регрессии с предсказаниями языковой модели позволила добиться наилучшего качества предсказаний для текста университетского уровня и в целом получить точные предсказания.

Таблица 10 — Результаты тестирования, показывающие качество различных моделей и комбинаций моделей.

	Accuracy	AUT	Precision	Recall	F1	RMSE	R2
Fine-tuned BERT	0.6308	0.9502	0.6366	0.6308	0.6311	0.0762	0.9173
Regression models							
Linear Regression with features	0.2095	0.2793	0.3821	0.2095	0.2333	0.1985	0.4399
Linear Regression combined	0.7053	**0.9873**	0.7163	0.7053	0.7028	0.0621	0.9451
XGBoost features	0.1491	0.2531	0.3871	0.1491	0.1378	0.2005	0.4283
XGBoost combined	0.5782	0.8055	0.6273	0.5782	0.5946	0.0728	0.9246
FNN with features	0.4918	0.8334	0.4834	0.4918	0.4839	0.1786	0.5465
FNN combined	0.7358	0.9741	0.7317	0.7358	0.7308	0.0654	0.9391
Classification models							
SVM features	0.3738	0.9455	0.3161	0.3738	0.2731	0.3226	-0.4787
SVM combined	0.3741	0.9462	0.3162	0.3741	0.2732	0.3226	-0.479
Random Forests with features	0.6002	0.9422	0.5952	0.6002	0.573	0.2179	0.3252
Random Forests combined	0.7775	0.9814	0.7814	0.7775	0.7723	0.0863	0.894
XGBoost features	0.6039	0.9137	0.5888	0.6039	0.5867	0.1968	0.4493
XGBoost combined	**0.7855**	0.9834	**0.7839**	**0.7855**	**0.7835**	**0.0605**	**0.9479**

3.7 Обсуждение

Полученная модель была протестирована на данных юридических документов. Первоначальные прогнозы были получены с использованием точно настроенной модели BERT в сочетании с лингвистическими функциями и использованы в модели XGBoost.

Для набора данных «CorDec» все документы были идентифицированы как имеющие наибольшую сложность. Для данных CorCodex 95% документов получили максимальную оценку сложности. Данные «CorRIDA» оказались

наиболее разнообразными: 83% данных идентифицированы как документы высокой сложности. На рисунке 3.5 показано распределение остальных файлов.

Наблюдаемые различия между тремя наборами данных в целом соответствуют нашим ожиданиям. В корпус российских внутренних документов и актов "CorRIDA" входит малоизученная категория юридических текстов — так называемые «внутренние документы». Они создаются в конкретной государственной организации и регулируют только деятельность этой организации. В корпусе присутствуют документы, адресованные «обычному гражданину»: абитуриенту в вузе, посетителю музея или театра, пациенту поликлиники и т. д. Видимо, именно такие официальные тексты мы (т. е. русскоязычные, не являющиеся профессиональными юристами) периодически имеем дело. Например, мы подписываем «Согласие на обработку персональных данных», «Информированное согласие на медицинское вмешательство» или «Договоры на оказание услуг». Внутренние документы не всегда пишутся юристами, для их формирования используются стандартные шаблоны, но самое главное – они адресованы «обычным ораторам». Неудивительно, что набор данных «CorRIDA» состоит не только из текстов максимального уровня сложности.

Решения Конституционного суда, напротив, пишутся высокопрофессиональными юристами, описание см. в [132]. Такие документы номинально адресованы широкому кругу граждан. Однако самих юристов беспокоит излишняя сложность формулировок решений Конституционного суда. Таким образом, [86] приходит к выводу, что «среднее решение Суда написано слишком сложным языком, рассчитанным на читателя с высшим образованием».

Третий набор данных (корпус "CorCodex") состоит в основном из текстов федеральных законов и кодексов. Жалобы на сложность и непонятность законов можно считать трюизмами, ср. остроумная цитата из [133]: «Жалобы на чрезмерную сложность закона так же стары, как и сам закон». Существующие исследования показывают, что сложность законодательных текстов с годами возрастает, см. [88]. Действительно, по нашим результатам, только 11 из 278 текстов корпуса «CorCodex» не получили оценку, кроме максимальной, при этом 6 документов относятся к периоду с 1993 по 1999 гг., 4 написаны в период с 2000 по 2003 гг., 1 текст был подготовлен в 2010 г.

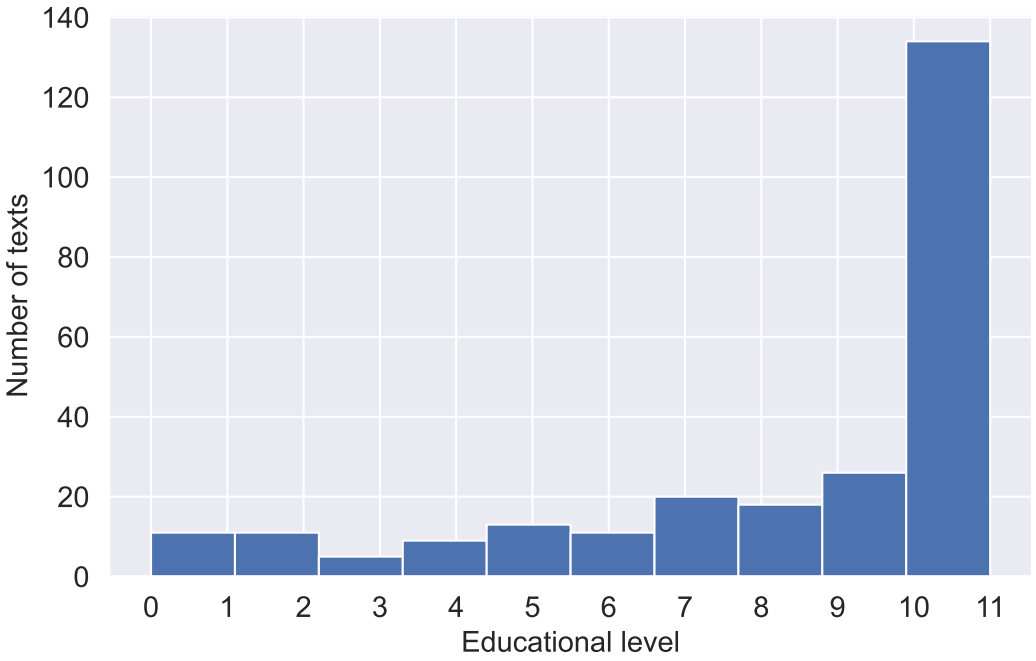


Рисунок 3.5 — Распределение сложности данных CorRIDA, за исключением текстов университетского уровня.

3.8 Выводы главы

В этой главе был предложен метод гибридизации модели прогнозирования сложности. Обучающий набор данных с текстами был собран из учебников на русском языке различного уровня сложности по предметам, как относящимся к области юриспруденции, так и дающим общеязыковые характеристики. Исследование демонстрирует эффективность модели глубокого языка BERT как самой по себе, так и в сочетании с заранее заданными лингвистическими функциями. Качество моделей измерялось по набору метрик, направленных на поиск модели, способной обеспечить высокую точность в целом, высокое качество прогнозов для сложных текстов в конкретных условиях и малое расстояние между прогнозируемыми и фактическими значениями даже в случае ошибок. Эти результаты показывают, что дополнительные прогнозы языковой модели обеспечивают повышение качества всех моделей, основанных на регрессии и классификации. Модель XGBoost с настроенными параметрами, обученная на основе функций и прогнозов языковой модели, дала наилучший результат на обучающих данных и использовалась на заключительном этапе тестирования.

Дополнительные испытания юридических документов показали эффективность этого подхода при идентификации сложных текстов, но выявили его самый большой недостаток, т.е. зависимость от данных.

Глава 4. Языковая сложность русских юридических подстилей и жанров

4.1 Вступление

В данной главе основное внимание уделяется лингвистической сложности юридических подстилей и жанров в современном русском языке. Как указывает S. Goźdź-Roszkowski, «За выражением «юридический язык» скрывается множество специфических классов текстов (жанров), используемых различными профессиональными группами, работающими в разных правовых контекстах. Юридический дискурс охватывает континуум от законов, принятых на разных уровнях, судебных решений, юридических отчетов, справок, различных договорных инструментов, завещаний, доверенностей и т. д. через устные жанры, такие как, например, допрос свидетелей, подведение итогов присяжных, заключение судьи и т. д. Этот список ни в коем случае не является исчерпывающим. Это лишь указывает на необычайное разнообразие юридического дискурса» [134].

Mattila et. al. [50] особо отмечают, что в некоторых правовых областях некоторые национальные правовые традиции используют «очень сложные конструкции предложений», научную лексику, формальный и архаичный язык и т. д. Таким образом, юридические жанры можно охарактеризовать по уровню языковой сложности языка, напр. [135]) по двум международным документам, [136] по контрактам, [137] по различным подвидам юридического языка.

Цель данного раздела — выяснить различия в лингвистической сложности между юридическими документами, отличающимися предметной областью, подстилем и жанром.

Для этого используются подходы к классификации стилей, подстилей и жанров, предложенные русской функциональной стилистикой. Под юридическими текстами понимается разновидность текстов «официально-делового стиля».

Функциональная стилистика выделяет законодательный, судебный и административный подстили официально-делового стиля. Первый подстиль относится к сфере законодательства, второй – к юрисдикционной, третий –

к административной, см., например, [56][329]. Кроме того, выделяется дипломатический подстиль. Документы этого подтипа регулируют правоотношения между государствами.

Во-первых, в этой главе разделяются документы национального права и международно-правовые документы. Это различие имеет смысл, поскольку многие документы международного права переводятся, т.е. в лингвистическом отношении они могут иметь существенные отличия от документов, составленных на русском языке.

Во-вторых, рассматриваются синхронные документы. Понятие синхронистичности формализуется следующим образом. «Синхронными» принято считать все документы, выпущенные в Российской Федерации в 1991 году и после (независимо от того, имеют ли эти документы юридическую силу или нет). Таким образом, анализируются документы Российской Федерации, а не СССР, не Российской империи, не Киевской Руси и т. д. Исключением из этого определения синхронности являются действующие международные документы, которые (независимо от даты их составления) также включены в анализируемый российский правовой корпус.

В-третьих, изучаются отдельные юридические жанры. Каждый из подстилей – законодательный, юрисдикционный и административный – имеет отдельный набор жанров. При этом в набор документов административного подстиля не вошли разнообразные служебно-деловые документы, относящиеся к учетной документации, товаросопроводительной документации и т.п. Такие документы не вошли в изучаемую выборку, поскольку они явно не относятся к категории юридических текстов. Более подробную информацию о создании корпуса юридических текстов, образец которого анализируется в этой главе, см. раздел 4.3.1.

В-четвертых, рассматриваются только письменные юридические жанры; устные жанры остаются за рамками рассмотрения.

4.2 Обзор литературы

4.2.1 Жанровые исследования

В западной лингвистике существуют три основные научные традиции жанроведения, а именно риторическое (RGS), системная функциональная лингвистика (SFL) и английский для специальных целей (ESP), см., например, Wang (2019). Первая традиция понимает жанры как риторические действия, утверждая, что «жанр возникает в результате повторяющихся социальных действий в повторяющихся ситуациях, которые порождают закономерности в форме и содержании» (Wang, 2019: 457). Жанровые исследования в рамках нового риторического подхода больше фокусируются на отношениях между текстом и контекстом, чем на его особенностях. Исследователь SFL J. Martin определяет жанр как «постановочную, целенаправленную, целенаправленную деятельность, в которой говорящие участвуют как представители нашей культуры», соответственно, тексты с одинаковой общей целью принадлежат одному и тому же жанру [138][456]. Определение жанра в рамках ESP было предложено J. Swales, который рассматривает жанр как «класс коммуникативных событий, члены которых разделяют некоторый набор коммуникативных целей» [139][58].

Основываясь на идеях трех жанровых теорий, V. K. Bhatia предложил следующее определение жанра: «Жанр по существу относится к использованию языка в конвенционализированной коммуникативной среде с целью выражения определенного набора коммуникативных целей дисциплинарного или социального института, которые порождают устойчивые структурные формы, накладывая ограничения на использование лексико-грамматических, а также дискурсивных ресурсов» [140][27].

Помимо самого жанра в качестве основной таксономической единицы, исследователи используют объединяющую текстовую категорию (супержанр или макрожанр) и разделяющую текстовую категорию (поджанр). Таким образом, говоря о юридическом языке, Mattila et. al. [50] предлагает различать юридические поджанры в соответствии с различными подгруппами авторов-юристов (среди которых, в частности, судьи, законодатели, администраторы и адвокаты).

Как указано в [141][13], «не существует фиксированного списка юридических жанров, хотя можно выделить ряд известных типов юридического текста». К основным типам относятся: «законодательные» документы (например, договоры, конституции, статуты, нормативные акты, подзаконные акты (иногда «подзаконные акты»), нормативные кодексы); документы «частного права» (например, контракты, приказы, акты, завещания, договоры аренды, перевозки, ипотечные документы, строительные контракты); и «процессуальные» документы (например, вступительная речь на суде, перекрестный допрос, итоговая речь, указания присяжным).

Активные исследования юридических жанров начались в 1980-х годах, см. [142][13]. Имеются исследовательские работы по законодательству и правовым жанрам под авторством Bhatia [140; 143], по адвокатским запискам Kurzon [144], по контрактам Tiersma [145] и Trosborg [146], по законодательству тексты и контракты Trosborg [147], о профессиональной аргументации юристов Howe [148], об ученичестве в академическом дискурсивном сообществе и степени лингвистической сложности Iedema [149].

4.2.2 Исследования сложности

Существует множество исследовательских работ, связанных с анализом сложности языка. [82]. Исследователи русскоязычных юридических документов акцентировали внимание на сложности текстов того или иного типа, а точнее, даже документов с типовым заглавием, выданных тем или иным учреждением, см. работу Дмитриевой (2017)[86] о сложности решений Конституционного Суда РФ и другие исследовательские работы, о которых говорится далее. В статье Дмитриевой[86] сложность оценивалась по единой формуле читаемости. Савельев, Кучаков [89] проанализировали решения арбитражных судов субъектов РФ с использованием двух метрик сложности: простого TTR, значение которого зависит от длины текста, и максимальной длины зависимости - расстояния от корня до зависимого компонента из дерева синтаксических зависимостей, рассчитываемое следующим образом: «для каждого конкретного текста берется одно значение, которое является максимальным для всех предложений текста» [88]. В то же время авторы интерпретировали значения TTR в противоречии

с общепринятым подходом, ср. следующая цитата: «Множество формальных повторов одних и тех же слов, обозначающих предметы права и различные юридические термины, мешают восприятию смысла предложения. В данном случае можно сказать, что сокращение <лексического – О. Б., Н. Т.> разнообразия не только не приводит к упрощению текста, но и вызывает противоположный эффект»[88].

Наиболее жанрово разнообразная выборка русских юридических текстов проанализирована Савельевым[150]; В исследовании автор сравнивает акты Конституционного Суда РФ, законы и кодексы, приказы министров и указы президента. Савельев подсчитывает «количество трудночитаемых предложений» по «теме» текстов (см., например, следующие темы: «Правила, инструкции, указания, приказы и другие решения», «Акционерное общество», «Центральный банк РФ», «Пенсионный фонд РФ»). При этом тематика текстов получается не в результате их анализа, а согласно «Общеправовому классификатору отраслей законодательства». Таким образом, читателю не предоставляется сравнительный анализ жанров или типов текста по сложности.

Можно резюмировать, что для русского языка по сложности рассматривались следующие категории документов: законодательные тексты, то есть законы [88; 90], и судебные решения (см. цитированные выше исследования).

Основной вывод состоит в том, что юристы, изучающие тексты российского права, игнорируют жанровые различия как нетрадиционные и неактуальные. То есть авторов совершенно не интересует жанровый анализ и связь жанра текста с его сложностью, поскольку они применяют другие (легальные, не жанровые) классификации текстов или не применяют никаких классификаций вообще. Между тем было продемонстрировано, что игнорирование жанра может существенно повлиять на адекватность анализа текстов юридической предметной области, см., например, [151] о юридической терминологии. [152] показали, что «на оценку читабельности сильно влияет текстовый жанр, и по этой причине необходимо жанрово-ориентированное понятие читабельности <...> при использовании классификационных подходов к оценке читабельности надежные результаты могут быть достигнуты только с использованием моделей, учитывающих жанр».

4.3 Материалы и методы

4.3.1 Юридические документы

Чтобы понять, какие документы подлежат включению в правовой корпус, были рассмотрены таксономии из российских правовых баз и баз документации: Консультант Плюс [107], Гарант [108], Континент [153] , Техэксперт [154]. На основе этой информации был сформирован предварительный перечень видов документов, содержащий 591 позицию (далее – «list-591»). Для оценки этого списка при содействии юристов был проведен эксперимент по параллельному аннотированию типов документов пятью экспертами. Эксперты (один доктор наук и четыре аспиранта) рассматривали список по строкам и ответили на вопрос: «Является ли этот <конкретный пункт списка, тип документа> юридическим документом или нет?». Согласованность ответов оценивалась по каждой строке (т.е. по каждому «виду документа» отдельно) с использованием простого процента согласия. Таким образом был получен список из 108 «видов документов», соотнесенных с письменными юридическими жанрами.

Следующим шагом в формировании списка жанров стал анализ словарей юридических терминов Борисова [155] и Додонова [156]. Последовательно рассматривались все строки «list-591» (независимо от оценок адвокатов), содержащие «виды документов». Затем в словарях искали термин, соответствующий типу документа. На основании интерпретации значения термина было принято решение включить вид документа в список жанров для формирования корпуса. Данная процедура позволила выявить виды документов, не упомянутые в «list-591», а также уточнить понимание рассматриваемых жанров. Следующие категории документов не подлежали включению в корпус юридических текстов: «бухгалтерские документы» (например, авансовый отчет, аудиторское заключение, баланс, коносамент), «платежные документы» (например, долговое требование, дорожный чек). , счет-фактура), «внешнеторговые документы» (например, отпечаток), «транспортные документы» (например, коносамент, ордер на отпуск товара), «грузовые документы» (например, грузовая квитанция, грузовая манифест, доковая квитанция, погрузочная накладная) , «денежные

документы» (например, кассовый чек), «складские документы» (например, складское свидетельство).

Последним этапом формирования перечня видов документов стал анализ Российского классификатора управленческой документации, с помощью которого перечень наименований документов был вновь расширен. Объединенный список юридических «видов документов» (612 позиций) затем использовался для получения текстов документов с сайтов юридических баз данных и сайтов органов государственной власти.

4.3.2 Анализ данных

С помощью списка типов документов (см. предыдущий раздел) были получены юридические документы и сформированы в текстовую коллекцию. Затем названия документов из этой текстовой коллекции были нормализованы, в результате чего получился список жанров, состоящий из 306 позиций. Все жанры были разделены на следующие категории: международные документы и документы национального права (документы административного подстиля, документы законодательного подстиля и документы юрисдикционного подстиля; далее соответствующие документы будут обозначаться с использованием аббревиатур АССД, LSSD и JSSD). На следующем этапе были проанализированы выбранные жанры (всего 68 жанров, в том числе 14 административных, 24 законодательных и 30 судебных). Основанием для отбора послужило количество документов той или иной жанровой категории и общественная значимость документа (например, в выборку LSSD вошла Конституция Российской Федерации).

Перечни анализируемых жанров документов национального права приведены в таблице 11. В таблице также указано количество рассматриваемых жанров (по подстилям), общее количество документов каждого подстиля и объем выборок. в словах.

Таблица 11 — Жанры национальных правовых документов

SS	#Genres	List of Genres	#Docs	#Words
ASSDs	14	Ministerial Declaration of Goals and Objectives, Interaction Agreement, Ministerial Rules, Ministerial Agreement, Ministerial Minutes (Extract), Agreement on Information Interaction, Cooperation Agreement, Territorial Agreement, Performance Standard, Priority Project Change Request, Code of Ethics and Service Conduct, Ministerial Minutes, Ministerial Regulations, Ministerial Letter	938	3,798,795
LSSDs	24	RF Government Decree, Ministerial Order, RF Presidential Edict, Federal Law, Ministerial Decree, Labor Protection Instruction, Ministerial Instruction, RF Subject's Law, Ministerial Resolution, Ministerial Decision, RF Governmental Resolution, Regional Parliament Decree, Federal Parliament Decree, Sanitary Regulations and Standards, RF Law, RF Subject's Government Decree, Ruling Document, Ministerial Conclusive Statement, Labour Protection Rules, Ministerial Temporary Order, RF Instructional Letter, RF Code, RF Fundamentals of the Legislation, RF Constitution	14,813	58,430,223

JSSDs	30	Ruling of the RF Constitutional Court, Judgment of the RF Supreme Court, Ruling of the RF Supreme Court, Decree of the Arbitration Court of Appeal, Decree of the RF Supreme Court, Judgment of the City Arbitration Court, Decree of the RF Constitutional Court, Decree of the Federal Arbitration Court, Decree of the District Arbitration Court, Decree of the City Court, Decree of the Regional Court, Decree of the Appeal Court of general jurisdiction, Judgment of the Regional Arbitration Court, Decree of the Intellectual Property Court, Ruling of the Intellectual Property Court, Judgment of the Supreme Arbitration Court, Ruling of the RF Subject's Supreme Court, Verdict of the City Court, Verdict of the Regional Court, Decree of the RF Supreme Arbitration Court, Decree of the Regional Court, Decree of the RF Subject's Supreme Court, Prosecutor's of the RF Subject's Protest, Ruling of the Statutory Court, Conclusion of the RF Council of Judges, RF Supreme Court Protest, Ruling of the City Court, Decree of the Regional Arbitration Court, Ruling of the Regional Court, Verdict of the RF Subject's Supreme Court	26,436	50,138,771

Формат метамаркировки позволяет сравнивать документы одного жанра, выпущенные разными учреждениями, например постановления Конституционного Суда РФ и постановления Верховного Суда РФ, постановления Правительства РФ и постановления Министров РФ. Набор данных по международному праву состоит из 1617 текстов, 6400239 слов, включает международные соглашения, конвенции, указы и решения международных судов.

4.3.3 Модель оценки сложности

Модель сложности подробно описана в предыдущей главе. Модель работает в два основных этапа [4].

Первый этап состоит из прогнозирования сложности с использованием предварительно обученной модели на основе Transformer. Transformer модели показывают высокую эффективность при решении широкого спектра задач обработки языка с использованием идеи предварительного обучения — процедуры инициализации, направленной на сохранение основных особенностей языка, и тонкой настройки — процесса, направленного на адаптацию модели для решения любых заданных задач. задача. В нашем случае RuBERT был выбран в качестве базовой предварительно обученной языковой модели. Для решения задачи дообучения языковой модели был собран вспомогательный набор данных.

Этот набор данных состоит из текстовых фрагментов, случайно выбранных из 1448 учебников различной сложности: от дошкольных (используется для описания нулевого уровня сложности), школьных учебников всех классов (сложность от «1» до «11») и учебников университетского уровня (описывающий максимальный уровень сложности – «12»). Данные содержат фрагменты из книг по предметам «Юриспруденция», «Социальные науки», «Экономика», «Культурология», «История» и т. д. Предметы были выбраны на основе того, что они являются хорошими общеязыковыми дескрипторами или их отношением к нашей области исследования.

Решение обучить модель на данных учебников было продиктовано отсутствием обучающих данных, в полной мере соответствующих юридическим текстам. Поэтому в качестве ближайшей альтернативы были выбраны учебники по темам, связанным с юриспруденцией, экономикой и другими общественными

науками. Это решение может привести к созданию более обобщенной модели сложности. Эта модель способна работать с широким спектром данных с точки зрения уровней сложности, но может с трудом различать тексты высокой сложности между собой.

RuBERT был обучен как регрессионная модель с использованием стандартного процесса дообучения. Модель регрессии была выбрана в качестве средства моделирования связи между уровнями сложности и, таким образом даже неправильные прогнозы могут быть относительно близки к их реальным значениям.

Следующая часть модели — это кодировщик данных, который выводит вектор длиной 133 для каждого текста. Векторные значения представляют собой набор лингвистических признака.

Признаки разделены на 10 общих категорий:

- базовые метрики, традиционно используемые в задачах оценки читабельности;
- формулы читабельности, адаптированные для русского языка;
- слова различных классов частей речи;
- n-граммы частей речи;
- общеязыковые частотные характеристики текстовых лемм;
- словообразовательные закономерности;
- отдельные граммы;
- лексические и семантические особенности, многословные выражения;
- синтаксические особенности;
- связность.

Кодировки данных и прогнозы языковой модели затем передаются в окончательную гибридную модель. Тринадцать подходов были протестированы и сравнены с использованием различных моделей, обученных с дополнительными предсказаниями языковой модели или без них.

Было обнаружено, что во всех тестах использование прогнозов языковой модели обеспечило существенное улучшение качества прогнозов. Используя набор метрик классификации и регрессии, было обнаружено, что модель XGBoost, обученная на функциях и прогнозах, обеспечивает наилучшее качество с точностью, точностью и оценкой F1 0,78 или выше. Это справедливо даже для регрессионных показателей, таких как RMSE (с коэффициентом ошибок 0,06) и R2 (с коэффициентом детерминации 0,9479).

4.4 Результаты и обсуждение

4.4.1 Оценки сложности по подстилю и локальному/глобальному статусу

В таблицах 12, 13 и 14 ниже представлены результаты оценки языковой сложности документов национального законодательства (ASSD, LSSD и JSSD) и документов международного права. В таблице 12 показаны результаты гибридной модели, в таблице 13 показаны прогнозы ruBERT, а в таблице 14 показаны прогнозы сложности на основе метрик.

Таблица 12 — Предсказания гибридной модели

Сложность	Администр.	Законодат.	Юрисдикц.	Международн.
12	911	14002	26368	1522
11	13	516	31	46
10	12	256	37	49
9	1	5	0	0
8	1	17	0	0
7	0	2	0	0
6	0	4	0	0
4	0	5	0	0
2	0	3	0	0
0	0	3	0	0

Результаты показывают, что подавляющее большинство всех документов во всех наших больших классах оцениваются всеми моделями как максимально сложные. Например, если присмотреться к результатам гибридной модели (см. таблица 2), то класс сложности «12» включает 97,1% документов административного подстиля, 94,5% документов законодательного подстиля и 99,7% документов юрисдикционного подтипа национального законодательства. По отношению ко всем документам международного права доля документов уровня сложности «12» составляет 94,1%.

Набор LSSD самый разнообразный по сложности. Далее приводится пояснение, как работают модели на уровне сложности “0”, которого мы собственно

Таблица 13 — Предсказания RuBERT

Сложность	Администр.	Законодат.	Юрисдикц.	Международн.
12	917	14224	26385	1546
11	10	418	48	69
10	9	107	3	2
9	1	31	0	0
8	1	15	0	0
7	0	2	0	0
6	0	4	0	0
5	0	1	0	0
4	0	3	0	0
3	0	2	0	0
2	0	2	0	0
1	0	1	0	0
0	0	3	0	0

Таблица 14 — Предсказания модели на основе метрик

Сложность	Администр.	Законодат.	Юрисдикц.	Международн.
12	915	14638	26374	1607
11	2	4	0	0
10	0	71	0	0
9	0	1	0	0
8	15	18	3	2
7	0	4	0	0
6	2	3	0	0
5	0	3	0	0
4	4	66	59	8
2	0	2	0	0
0	0	3	0	0

и не ожидали увидеть в нашем наборе данных. Гибридная модель и доработанная модель ruBERT присваивают этот уровень сложности трем документам, среди которых, например, Приказ Минобрнауки РФ “О Координационном совете Министерства образования и науки Российской Федерации по модернизации региональных систем дошкольного образования”. Таким образом, уровень сложности ”0” присваивается документам, тематика которых относится к дошкольному образованию. Метричная модель присваивает уровень сложности ”0” остальным трем документам, представляющим собой длинные последовательности коротких именных словосочетаний с асиндетическим согласованием, см., например, Постановление Правительства РФ от 14 февраля 2002 г. № 103 “Об утверждении перечня жизненно необходимых и важнейших лекарственных средств и изделий медицинского назначения для бесплатного приобретения гражданами, постоянно проживающими (работающими) на территории зоны проживания с правом на отселение, в соответствии с пунктом 19 части первой статьи 18 Закона Российской Федерации “О социальной защите граждан, подвергшихся воздействию радиации вследствие катастрофы на Чернобыльской АЭС”.[5] В то же время Приказ № 103 содержит множество сверхредких слов (названий лекарственных средств), например, “Аллопуринол”, “Тригексифенидил”, “Карбоплатин”, и определяется отточенной моделью ruBERT и гибридной моделью как максимально сложный текст.

Однофакторный дисперсионный анализ (One-Way ANOVA) по сложности каждого подстиля показывает значительную разницу между средними значениями разных подстилей с F-score 278,4. На рисунке 4.1 показаны средние значения сложности для каждого конечного состояния подстиля вместе с их стандартными отклонениями; оценки сложности были получены с помощью гибридной модели.

Визуализация подтверждает, что наиболее сложными документами в изучаемом наборе данных являются JSSD.

Линейный дискриминантный анализ (LDA) был выполнен для уменьшения размерности векторов признаков со 133 языковых параметров до 3. На рисунке 4.2 показана визуализация подстилей и статусов с использованием уменьшенных векторов для каждого документа.

Рисунок 4.2, в частности, демонстрирует, что лингвистические особенности хорошо контрастируют между документами юрисдикционного и законодательного подстилей, а тексты административного подстиля смешива-

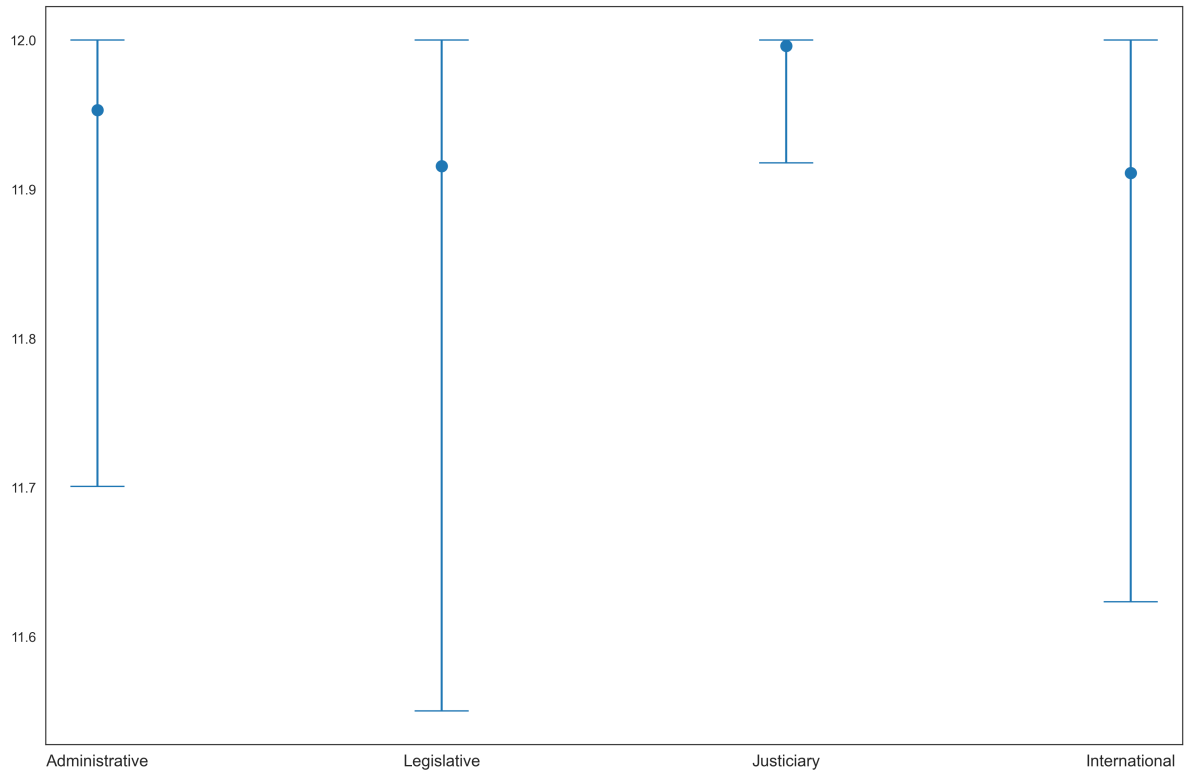


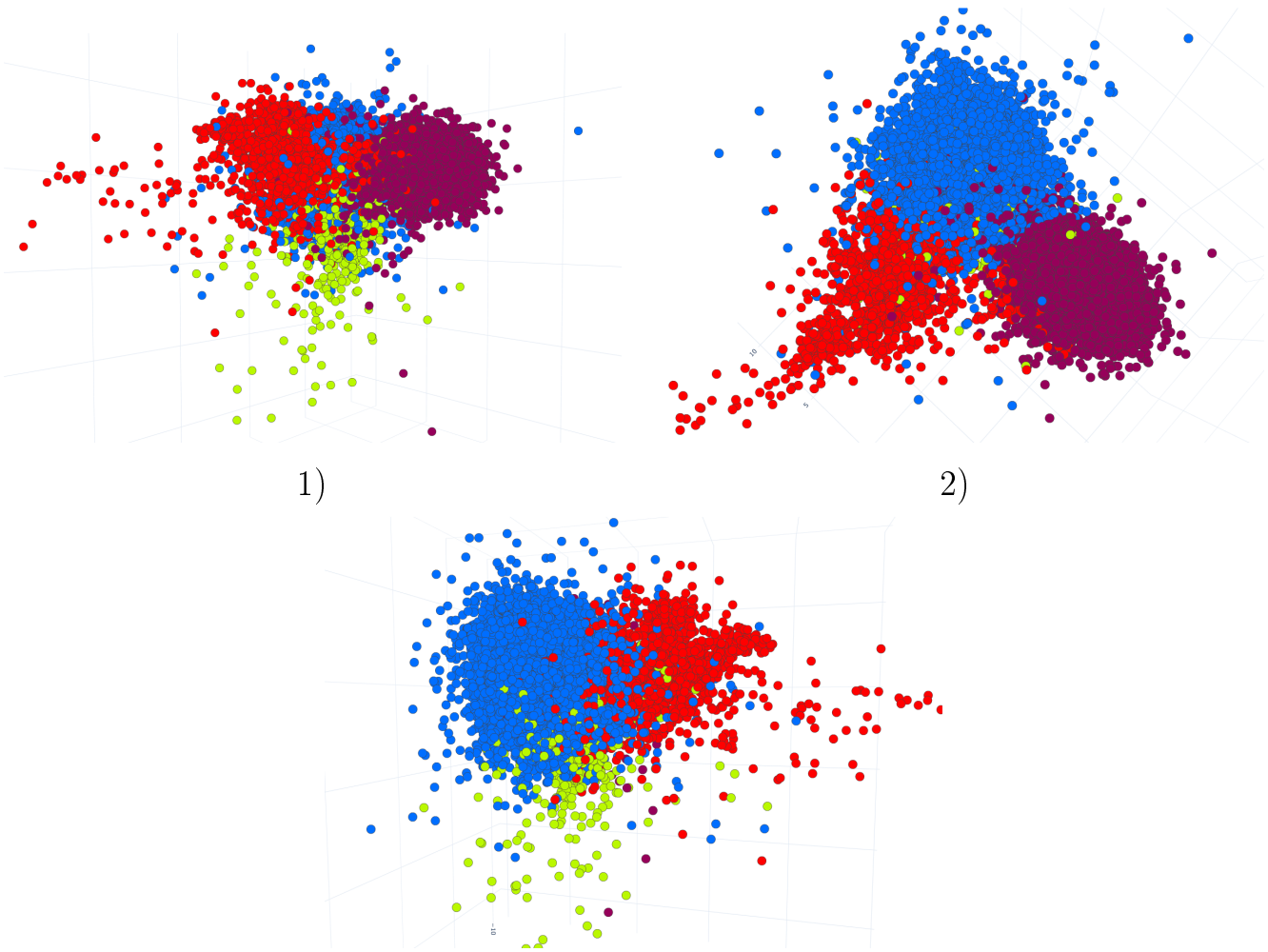
Рисунок 4.1 — Средние значения сложности (гибридные предсказания)

ются с текстами двух других классов подстилей. Кроме того, можно также утверждать, что значения лингвистических показателей успешно различают международные и отечественные правовые документы.

Для более детального сравнения документов по статусу были проанализированы средние значения лингвистических показателей. Для сравнения этих значений национальных законодательных документов и международных документов был использован t-критерий. Было обнаружено, что для скорректированных Бонферрони значений p менее 0,05 нулевая гипотеза (равные средние значения) может быть отклонена для 96 лингвистических признаков, что означает, что между средними значениями этих признаков существуют значительные различия. Для значений p менее 0,01 и менее 0,001 нулевая гипотеза отклоняется для 94 и 90 параметров соответственно.

На рисунке 4.3 показаны различия средних значений для национальных и международных документов, нормализованные и отсортированные по статистике t-критерия. Для построения графика показаны только параметры, значения критерия Стьюдента которых превышают 15.

Можно сделать некоторые наблюдения, согласно которым в отечественных документах по сравнению с международными больше производных слов,



1) 2)

3)

Рисунок 4.2 — Сравнение документов с использованием LDA для уменьшения размерности (три проекции)

последовательностей типа «существительное + существительное в родительном падеже», абстрактных слов, графических сокращений, последовательностей типа «существительное + существительное», аппозитивные конструкции, появление деепричастий. Кроме того, в отечественных документах предложения более длинные.

В международных документах по сравнению с отечественными больше глаголов будущего времени, вхождений личных местоимений, последовательностей типа «существительное + личный глагол», последовательностей типа «полное прилагательное + существительное», а также частых лемм (значение Ципфа = 7). Можно также отметить, что (согласно формуле динамического/статического) международные документы являются «более динамичными».

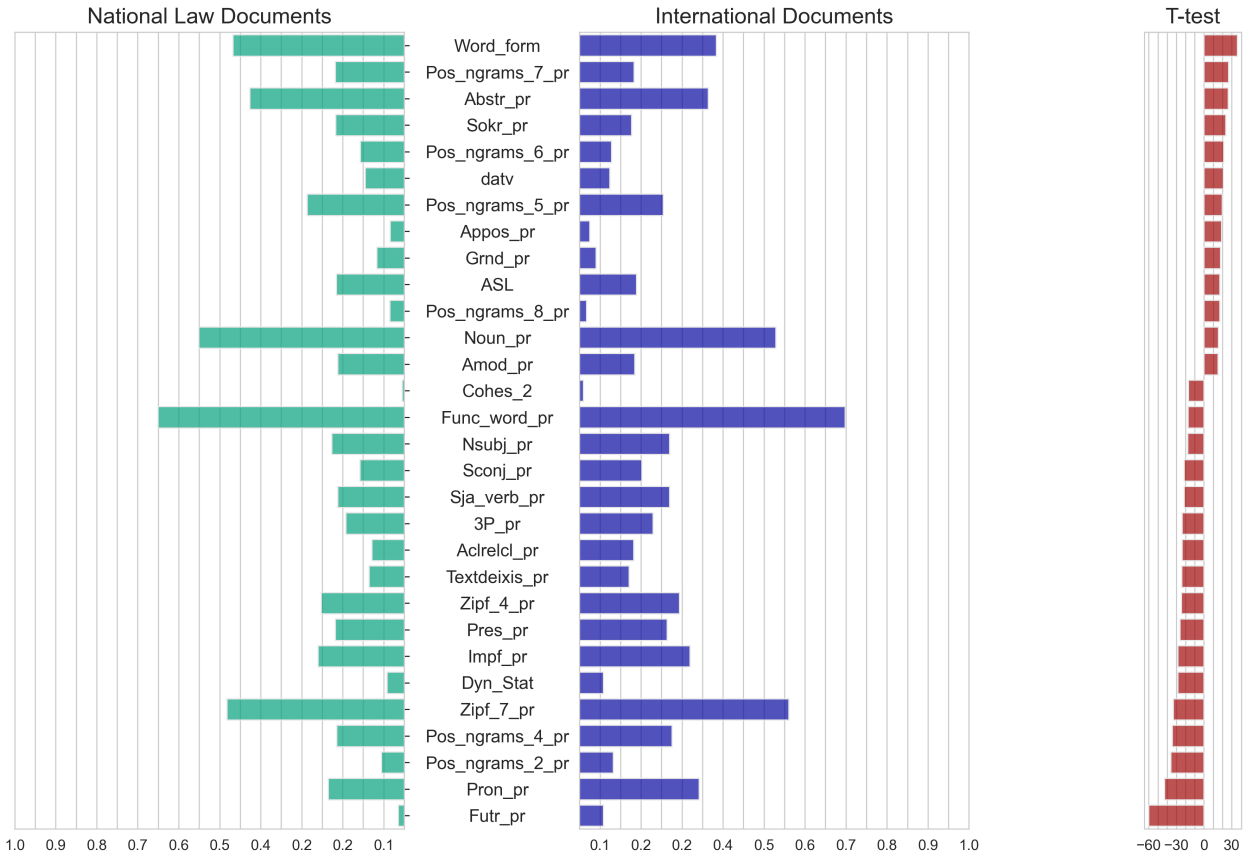


Рисунок 4.3 — Средние значения лингвистических показателей в документах по статусу

4.4.2 Оценки сложности по жанрам

Для каждого подстиля внутри группы национальных правовых документов были рассчитаны средние значения отдельных категорий признаков: «Синтаксических», «Основных» и «Частиречевых». Средние значения рассчитывались после нормализации min-max каждого признака. На рисунках 4.4, 4.5 и 4.6 представлены средние значения и соответствующие им стандартные отклонения для каждого жанра. Значения средних на визуализациях ранжированы по убыванию значений синтаксических метрик. Такое решение дает осмысленную интерпретацию полученных данных, поскольку достаточного разнообразного распределения отечественных документов по значениям сложности получено не было. Таким образом, можно делать обобщения, основанные на синтаксических особенностях, поскольку их можно считать наиболее показательными при оценке сложности текста.

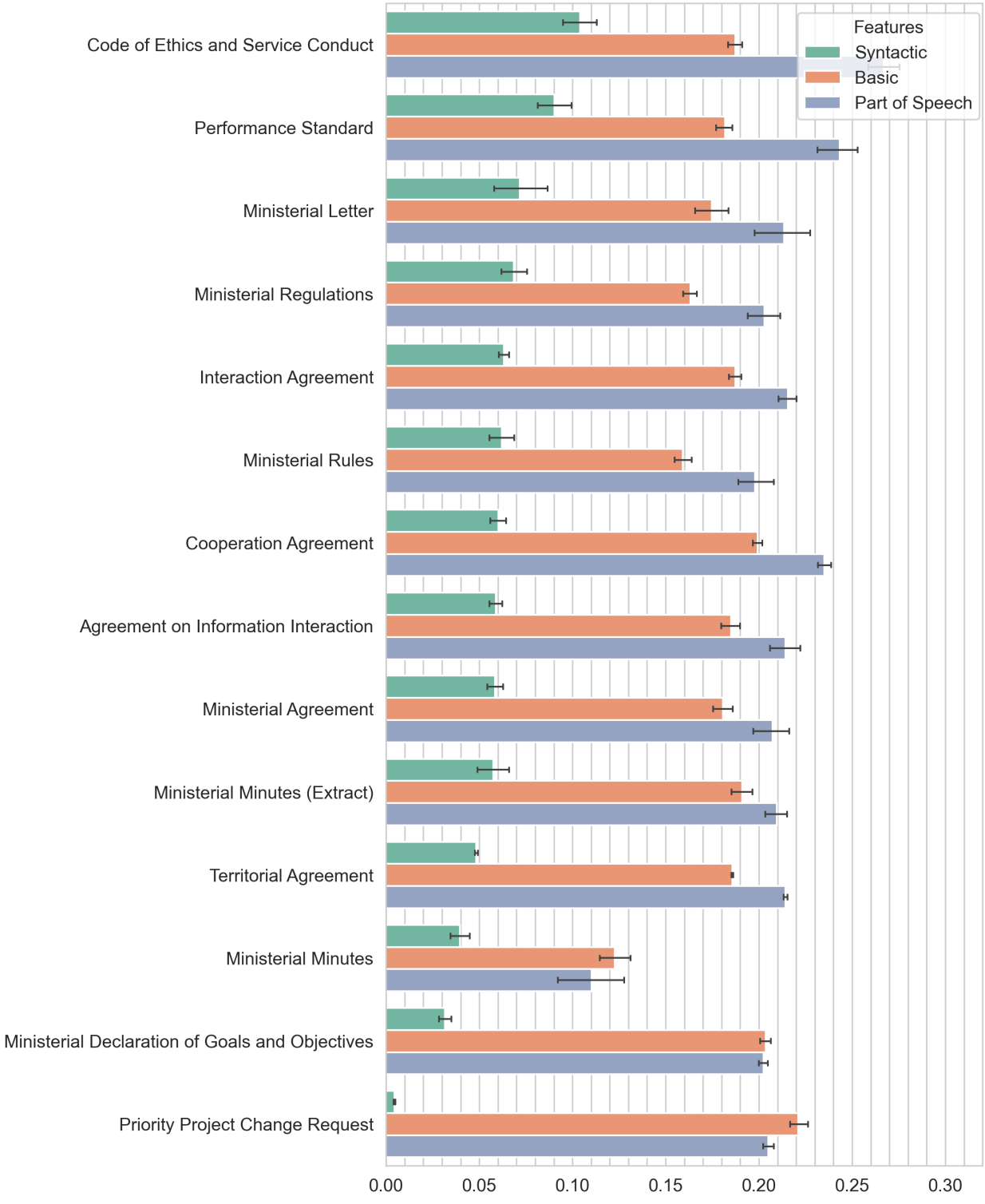


Рисунок 4.4 — Сложность жанров в рамках административного подстиля

Рисунок 4.5 — Сложность жанров в рамках законодательного подстиля

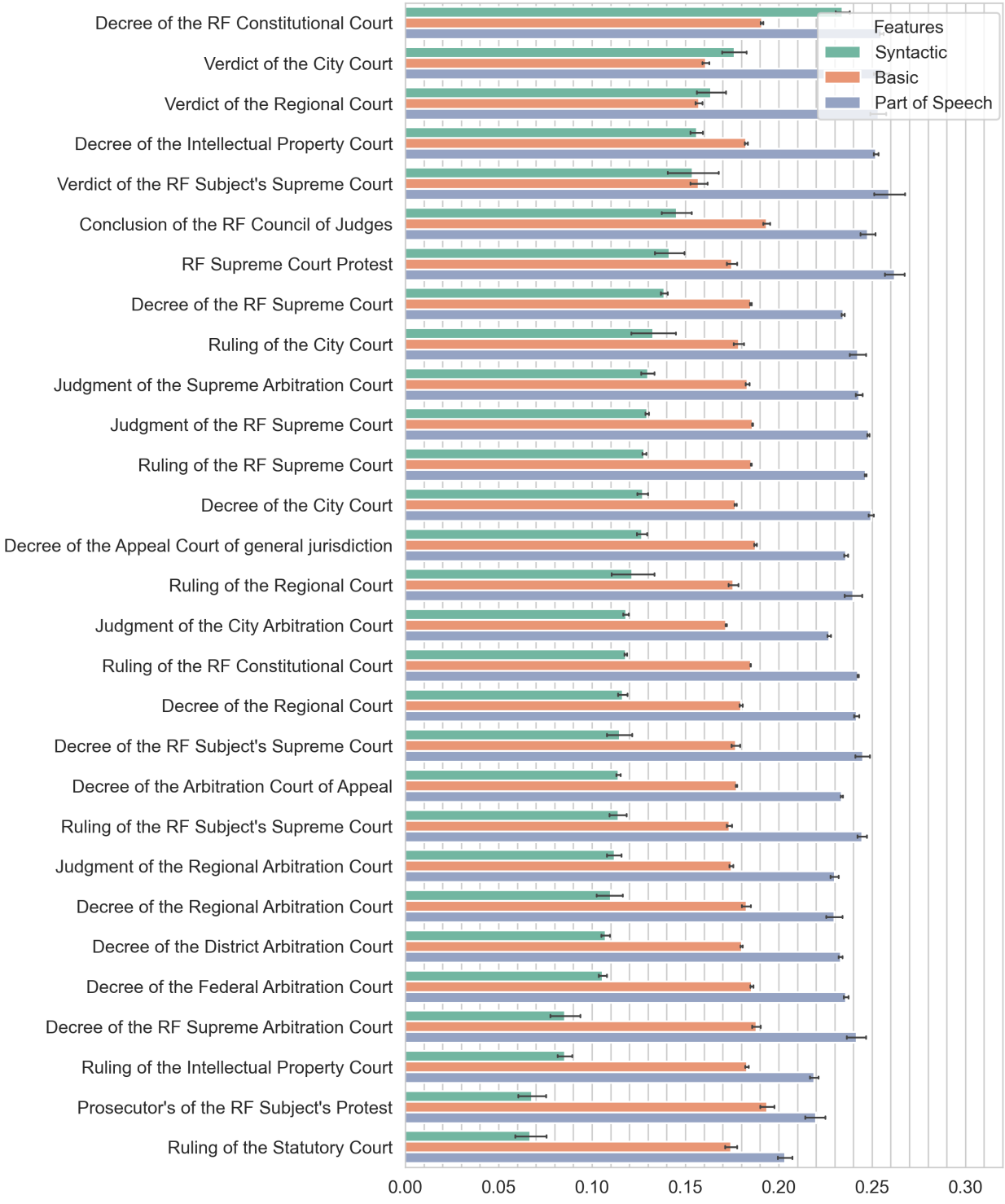


Рисунок 4.6 — Сложность жанров в рамках юрисдикционного подстиля

Далее приведены некоторые комментарии по конкретным показателям. В список синтаксических особенностей входят:

1. Признаки, показывающие структуру отдельных синтаксических словосочетаний (например, именной группы, см. метрику «Amod_p», т.е. долю прилагательных-модификаторов имени; глагольную группу, см. метрику «Advmod_pr», т.е. долю наречий сказуемого);
2. Признак, описывающий появление аппозициональных модификаторов («Appos»);
3. Признаки, указывающие на наличие сочинительного ряда (имеется в виду признак «Cc» «сочинительный союз» и признак «Conj», описывающий количество союзов);
4. Признаки, описывающие появление клаузальных определителей существительного (причастия и причастные предложения «Acl» отдельно от придаточных «Acl:relcl»), деепричастных определителей придаточного предложения, различных клаузальных дополнений («Ccomp», «Xcomp»); единицы, способные присоединять придаточные предложения, учитываются отдельно (“Mark”);
5. Признак, описывающий появление предложений с элементами, похожими на связки («Cop»);
6. Признаки, описывающие появление пассивных конструкций («Aux:pass», «Nsubj:pass», «Csubj:pass»).

Возможности анализа синтаксической сложности обусловлены и ограничены форматом синтаксического анализа. В данном случае важной составляющей модели сложности является учет особенностей на основе разметки UDPipe [62]. Кроме того, pymorphy2 использовался для разметки частей речи и морфологических аннотаций [111].

Основные выводы заключаются в следующем. Среди документов административного подтипа Кодексы этики и служебного поведения являются наиболее синтаксически сложными. Примером документа такого жанра является «Типовой кодекс этики и служебного поведения государственных и муниципальных служащих».[6] В документах законодательного подтипа прослеживается такая закономерность: самым синтаксически сложным документом неожиданно оказалась Конституция РФ. Постановления Федерального Парламента являются наименее синтаксически сложными (хотя они имеют наивысший балл сложности по базовым метрикам). Что касается документов

юрисдикционного подстиля, то наиболее синтаксически сложными (с заметным отрывом от других жанров) являются постановления Конституционного Суда РФ.

В целом, сравнение групп документов по жанрам (характеризуемых учреждениями, выпустившими конкретные тексты) показывает, что во всех трех наборах подстилей не сам жанр может иметь решающее значение для оценки сложности, но выпускающий его государственный орган или суд. Это хорошо видно на примере судебных документов, в совокупности которых постановления ВАС РФ и постановления Конституционного Суда РФ явно противопоставлены по синтаксической сложности.

4.5 Выводы главы

В этой главе был рассмотрен разнообразный по жанрам набор юридических текстов (43 804 документа, всего 118 768 028 слов). Набор данных включает документы международного права (1617 текстов, 6400239 слов) и документы национального права. Последние разделены на три подстиля, а именно административный подстиль (938 текстов, 3 798 795 слов), законодательный подстиль (14 813 текстов, 58 430 223 слова) и судебный подстиль (26 436 текстов, 50 138 771 слово). Все отечественные документы классифицируются по жанрам и учреждениям, выдавшим документ. Всего выделено 68 классов юридических жанров (14 административных, 24 законодательных и 30 судебных).

Всем документам были присвоены уровни сложности от «0» до «12». В этой главе были проанализированы прогнозы сложности точно настроенной модели ruBERT, прогнозы по 133 лингвистическим метрикам и прогнозы гибридной модели. Основные результаты анализа сложности документов по подстилям и жанрам следующие.

Подавляющее большинство всех документов всех крупных классов оценивается всеми моделями как максимально сложные. Так, гибридная модель присваивает класс сложности «12» 97,1% документов административного подстиля, 94,5% документов законодательного подстиля и 99,7% документов юрисдикционного подстиля национального законодательства. По отношению ко всем документам международного права доля документов уровня сложно-

сти «12» составляет 94,1%. Набор LSSD самый разнообразный по сложности. В среднем наиболее сложными документами в исследуемом наборе данных являются JSSD.

Языковые особенности хорошо контрастируют между документами юрисдикционного и законодательного подстилей, тогда как тексты административного подстиля смешаны с текстами двух других классов. Значения лингвистических показателей успешно различают международные и отечественные правовые документы.

Более детальное сравнение документов по внутреннему/международному статусу с помощью t-критерия показало, что существуют существенные различия между средними значениями по 110 языковым признакам. В частности, во отечественных документах по сравнению с международными больше производных слов, последовательностей типа «существительное + существительное в родительном падеже», абстрактных слов, графических сокращений, последовательностей типа «существительное + существительное + существительное», аппозитивных конструкций, появление деепричастий. Кроме того, в отечественных документах предложения более длинные. В международных документах по сравнению с отечественными больше глаголов будущего времени, вхождений личных местоимений, последовательностей типа «существительное + личный глагол», последовательностей типа «полное прилагательное + существительное», а также частых лемм (значение Ципфа = 7).

При сравнении документов по жанрам интерпретировались средние значения всех синтаксических показателей. Средние значения рассчитывались после нормализации min-max каждого признака. Среди документов административного подтипа Кодексы этики и служебного поведения являются наиболее синтаксически сложными. Самым синтаксически сложным документом законодательного подтипа неожиданно оказалась Конституция РФ. Постановления Федерального Парламента являются наименее синтаксически сложными (хотя они имеют наивысший балл сложности по базовым метрикам). Что касается документов юрисдикционного подстиля, то наиболее синтаксически сложными (с заметным отрывом от других жанров) являются постановления Конституционного Суда РФ.

В целом, сравнение групп документов по жанрам (характеризуемых учреждениями, выпустившими конкретные тексты) показывает, что во всех трех

наборах подстилей не сам жанр может иметь решающее значение для оценки сложности, но государственный орган или суд, выпускающий документ [7].

Глава 5. Доступность восприятия юридических текстов

5.1 Вступление

Группой экспертов Санкт-Петербургского государственного университета по заказу Федеральной налоговой службы России (далее — ФНС) проведения оценка уровня доступности восприятия письменных ответов налоговых органов на обращения физических лиц и организаций.

Анализ был проведен на коллекции из 2339 пар реальных вопросов и ответов, предоставленных сотрудниками ФНС для исследования и оценки по 82 регионам России (от 2 до 48 пар вопрос-ответ из каждого региона). Оценка проводилась по методике, подготовленной СПбГУ на основе результатов собственных исследований в рамках ряда научных проектов, в том числе в рамках НИИ Проблем государственного языка и посвященных изучению языка официальных (юридических) документов, и предварительно согласованной представителями ФНС. Методика предусматривала оценку на основании автоматизированного машинного анализа текста каждого ответа по 12 критериям, по каждому из которых давалась числовая оценка коммуникативного качества (доступности для восприятия) и которые были впоследствии сведены в общую оценку для каждого материала и среднюю оценку по каждому региону.

Каждый из критериев оценки имеет юридическое и лингвистическое обоснование исходя из того, какие требования к языку официальных документов, прямо предусмотрены положениями Конституции (с учетом их интерпретации Конституционным Судом РФ), действующего законодательства и подзаконных актов, а также очевидным образом вытекают из этих положений. Выполнение именно этих требований которых должно оцениваться и тщательно контролироваться.

Предложенные критерии предполагают в большей степени анализ содержания документа по его коммуникативным свойствам исходя из языковых, а не содержательных характеристик: оценка правильности ответов с точки зрения корректности разъяснения налогового законодательства не производилась. Требования к форме документа, носящие характер самостоятельных правовых требований (реквизиты, подпись уполномоченным должностным ли-

цом и проч.) не оценивалась, за исключением соблюдения правил русской орфографии и пунктуации, поскольку допущенные ошибки могут существенно сказаться на определенности и понятности ответа.

Каждый из критериев оценки предусматривает использование программных средств и инструментов, использование которых не ограничено и не требует согласия владельцев прав интеллектуальной собственности, в том числе какие-либо охраняемые объекты, интеллектуальные права на которые принадлежат СПбГУ.

Весь совокупный набор инструментов автоматического анализа текста и отдельные его элементы ранее не применялись для решения поставленных задач (оценки коммуникативного качества ответов налоговых органов на вопросы налогоплательщиков), поэтому после получения результатов автоматической обработки и автоматической оценки текстов для контроля полученных результатов проводился выборочный ручной контроль и оценка адекватности полученных данных. Кроме того, были обобщены и сопоставлены результаты автоматической оценки по разным критериям.

Некоторые критерии оказались не обеспечены эффективным инструментарием проверки именно в контексте этих конкретных задач. В частности, поиск разговорной лексики практически не дал содержательных результатов, однако анализ полученных данных дает основания считать, что причиной этого во многом стало отсутствие пригодного именно для данных задач словаря разговорной лексики. Не очень эффективные результаты дал поиск низкочастотных слов — слов, которые в частотном словаре на базе Национального корпуса русского языка (обширного представительного собрания русских текстов, призванного представить состояние современного русского языка во всём его разнообразии) не встречаются или встречаются очень редко.

В то же время по ряду критериев автоматический анализ дал показательные результаты. Прежде всего это касается общей оценки понятности текста по 19 метрикам, разработанным и применяемым в современной лингвистике. Анализ позволил разделить все проанализированные тексты на группы, сопоставив их между собой по характеристикам, влияющим на легкость и доступность восприятия, причем тексты, получившие низкие оценки по одной метрике, как правило, получали низкие оценки и по другим метрикам, что подтверждает их общую эффективность.

Эффективные результаты получены по “юридическому” характеру текстов ответов — наличие специальных юридических терминов, дословных цитат из нормативных актов и перефразированных фрагментов нормативных документов. Все эти критерии существенно осложняют восприятие текста непрофессиональными адресатами и в то же время демонстрируют низкую коммуникативную эффективность: если ответ в большей части состоит из того, что содержится в нормативных актах, то это чаще всего не то, что хотят увидеть заявители.

Можно сделать вывод, что предложенные инструменты в целом вполне успешно справляются с задачей оценки тех документов, которые были представлены для анализа, однако для масштабного системного применения требуют адаптации к поставленным задачам по результатам апробирования их применения.

5.2 Критерии оценки

5.2.1 Базовые критерии

Орфография и грамматика С юридической точки зрения, нарушение орфографических и грамматических правил создает риск наличия неопределенности содержания и влияет на восприятие полученного ответа. Кроме того, нарушение правил орфографии способно создать неопределенность относительно того, какое именно слово в действительности должно содержаться в тексте. Анализ показал, что 30% всех ответов (710) не имели ни одной орфографической или грамматической ошибки.

Пунктуация Нарушение правил пунктуации в еще большей степени создает неопределенность относительно содержания, поскольку отсутствие правильно расставленных знаков препинания позволяет по-разному интерпретировать синтаксическую структуру предложения. Анализ показал, что 16% всех ответов не имели ни одной ошибки в расстановке запятых. В остальных ответах ошибки носят единичный характер. При этом практически все проана-

лизированные ответы получили очень высокие оценки по данному критерию – от 96 до 100.

Фигуры речи и идиомы По критерию наличия в текстах ответов фигур речи и идиом проводилась оценка по общей шкале от 0 до 100, где 100 получал текст ответа, в котором не обнаружено ни одной метафоры или идиомы, а 0 – текст ответа, в котором обнаружено максимальное относительное число слов и выражений с указанными характеристиками. Каждый ответ получил оценку по данной шкале в зависимости от его положения в пределах минимального и максимального значения количества оцениваемых слов.

Для проверки наличия идиом (устойчивых неоднословных последовательностей слов, смысл которых напрямую не выводится из смысла слов, входящих в последовательность) была использована библиотека MMFLD (https://github.com/laihuiyuan/MMFLD), основанная на архитектуре T5. На вход модели подавались последовательности вида: 'Which figure of speech does this text contain? (A) Literal. (B) Idiom. | Text: Предложение текста ответа'. Таким образом, модель определяла наличие идиом в отдельных предложениях текста ответа. Исходя из данного ограничения оценка строилась относительно числа предложений в тексте.

В результате анализа в рассматриваемых текстах не нашлось ни одного примера использования метафор. Наличие фигур речи и идиом в текстах ответов носит единичный характер. Абсолютное большинство проанализированных ответов по данному критерию получили максимальную оценку.

Разговорная лексика По критерию наличия в текстах ответов разговорной лексики проводилась оценка по общей шкале от 0 до 100, где 100 получал текст ответа, в котором не обнаружено ни одного слова, относимого к разговорному стилю речи, а 0 – текст ответа, в котором обнаружено максимальное относительное число слов и выражений с указанными характеристиками. Каждый ответ получил оценку по данной шкале в зависимости от его положения в пределах минимального и максимального значения количества оцениваемых слов.

Низкочастотные слова По критерию наличия в текстах ответов низкочастотных (редких, малоупотребительных) слов проводилась оценка по общей шкале от 0 до 100, где 100 получал текст ответа, в котором не обнаружено ни одного слова, относимого к низкочастотным словам, а 0 – текст ответа, в котором обнаружено максимальное относительное число слов и выражений с указан-

ными характеристиками. Каждый ответ получил оценку по данной шкале в зависимости от его положения в пределах минимального и максимального значения количества оцениваемых слов. В качестве низкочастотных были выбраны слова из списка “Нового частотного словаря русской лексики” имеющие меру Ципфа менее 3. Мера Ципфа основывается на логарифмическом преобразовании значения частоты ipm (относительная частота, количество употреблений на миллион слов текстовой коллекции, на базе которой составлялся частотный список слов, ранжированный по убыванию частоты встречаемости) и позволяет распределить все слова, присутствующие в некотором частотном списке, по диапазонам (и таким образом отделить высокочастотные, среднечастотные и низкочастотные единицы, а затем оценить количество низкочастотных единиц).

Тональность (эмоциональная окраска текста) Оценка тональности не имеет прямого отношения к рассмотрению доступности ответа для восприятия и понимания, однако позволяет формализовать понятие “общее впечатление от текста”. Анализ тональности – это процесс анализа текста и определения соответствующего ему эмоционального тона. Его можно использовать для определения общего настроения фрагмента текста, например, является ли настроение положительным, отрицательным или нейтральным. В результате работы модели каждому тексту ответа был присвоен ранжированный индекс тональности, в котором значение 100 присваивалось – нейтральному тексту, 50 – тексту с позитивной тональностью, 0 – тексту с негативной тональностью. Таким образом, тексты с нейтральной тональностью (безоценочные и безэмоциональные, то есть соответствующие требованиям, предъявляемым к текстам официально-делового стиля) получали наивысший балл. Оценка тональности производилась с использованием дообученной модели RuBERT. Анализ показал, что все рассматриваемые тексты имеют нейтральную тональность. Это в полной мере соответствует ожиданиям экспертов, так как ответы налоговых органов и должны были иметь нейтральную тональность, а все отклонения от этого рассматривались бы как дефект.

Формальные критерии По критерию уровня соответствия ответа формальным правилам проводилась оценка по общей шкале от 0 до 100, где значение 0 присваивалось ответу, в котором нарушены требования к форме обращения и отсутствует необходимое указание на информационный характер ответа, 50 присваивалось ответу, где имелось одно из двух указанных нару-

шений, 100 присваивалось ответу, в котором не были выявлены указанные нарушения.

Критерий соблюдения формы обращения проверялся простым текстовым сравнением, критерий указания на информационный характер ответа проверялся с использованием кодирующей языковой модели. С помощью базовой предобученной модели Rubert-base-cased были получены кодировки предложений текстов и кодировка текста. указывающего на разъяснительный характер ответа. В случае нахождения значения косинусного сходства кодировок больше 0.75 ответ считался содержащим предупреждающее сообщение (значение 0.75 позволяет детектировать предупреждения в максимально свободной форме).

Для анализа были учтены два требования содержащиеся в ведомственных актах ФНС России — указание на информационный характер ответа и соблюдение формы обращения к заявителю. При необходимости в механизм оценки можно добавлять новые параметры. Полученные результаты демонстрируют то, что оба требования выполнены лишь в малой части проанализированных ответов. В абсолютном большинстве ответов отсутствовало указание на их информационный характер.

5.2.2 Юридическая терминология

По критерию наличия в текстах ответов юридической терминологии была произведена оценка с приведением к нормальному распределению со сдвигом. В текстах ответов были выявлены слова и неоднословные термины, относимые к словам юридической терминологии в обширных юридических словарях А. Б. Борисова и В. Н. Додонова. Список полученных терминов был доработан с помощью ручной оценки — были удалены слова, по мнению экспертов имеющие понятное “неюридическое” значение. Далее рассчитывалось относительное число юридических терминов, относительно всех слов текста. Из него по формуле $100e^{-1.4((x-0.2)/0.4)^2}$, где x - относительное число юридических терминов. Рисунок 5.1 показывает визуальное представление формулы.

Данная формула (частный случай нормального распределения) позволяет более плавно задать оптимальное значение числа юридических терминов. Формула с данными параметрами дает максимальную оценку — 100 для отве-

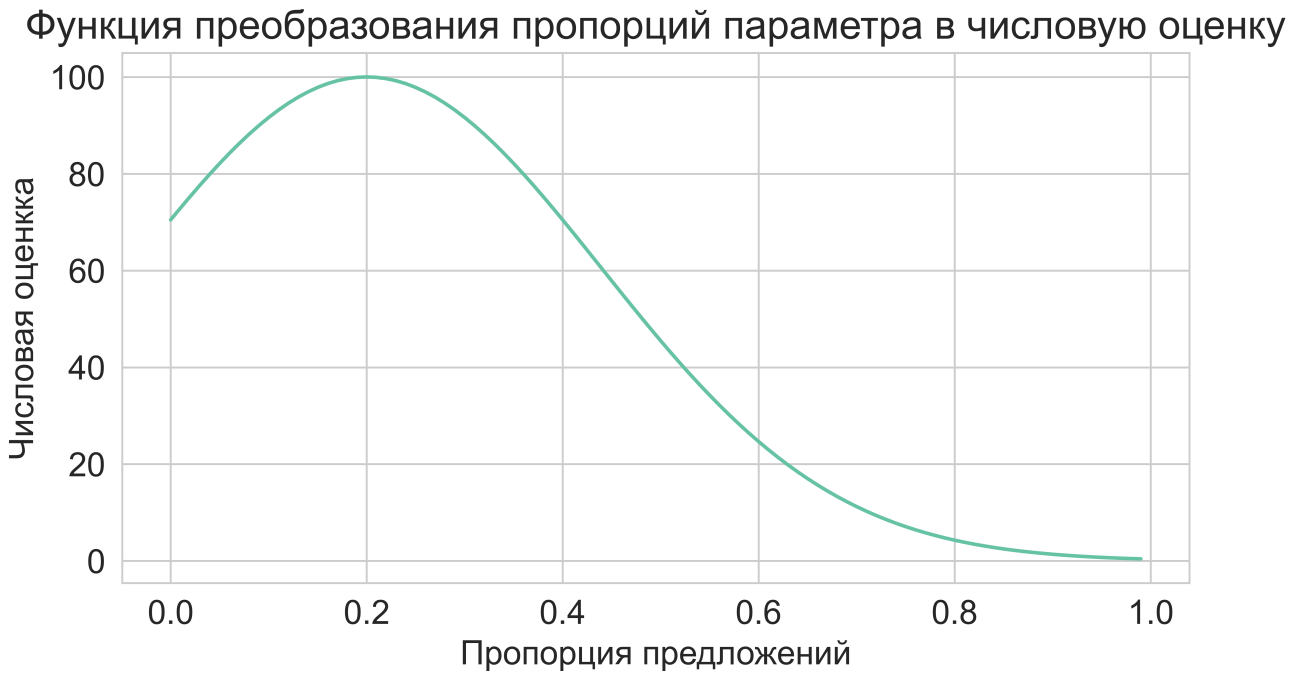


Рисунок 5.1 — Функция преобразования пропорциональных значений к системе оценивания от 0 до 100

тов, в которых доля юридических терминов составляет 20% от общего числа слов. В случае отсутствия юридических терминов, исходя из формулы, ответ получает оценку ≈ 70. С увеличением пропорции юридической лексики оценка приближается к 0 согласно нормальному распределению.

В тестовых данных - для 860 документов (37% всего набора текстов ответов) доля юридических терминов не превышает 20% от всех слов текста ответа.

Проведённый анализ показал, что в большинстве ответов налоговых органов специальная юридическая терминология употребляется в оптимальном соотношении к объёму самих ответов, что не затрудняет восприятие их содержания.

5.2.3 Соответствие вопроса и ответа

При оценке по данному критерию внимание было обращено не только на текст ответа, но и на сам вопрос, содержащийся в обращении налогоплательщика. Проверка критерия осуществлялась с использованием трех моделей — двух моделей ответов на вопросы и модели следствия предложений.

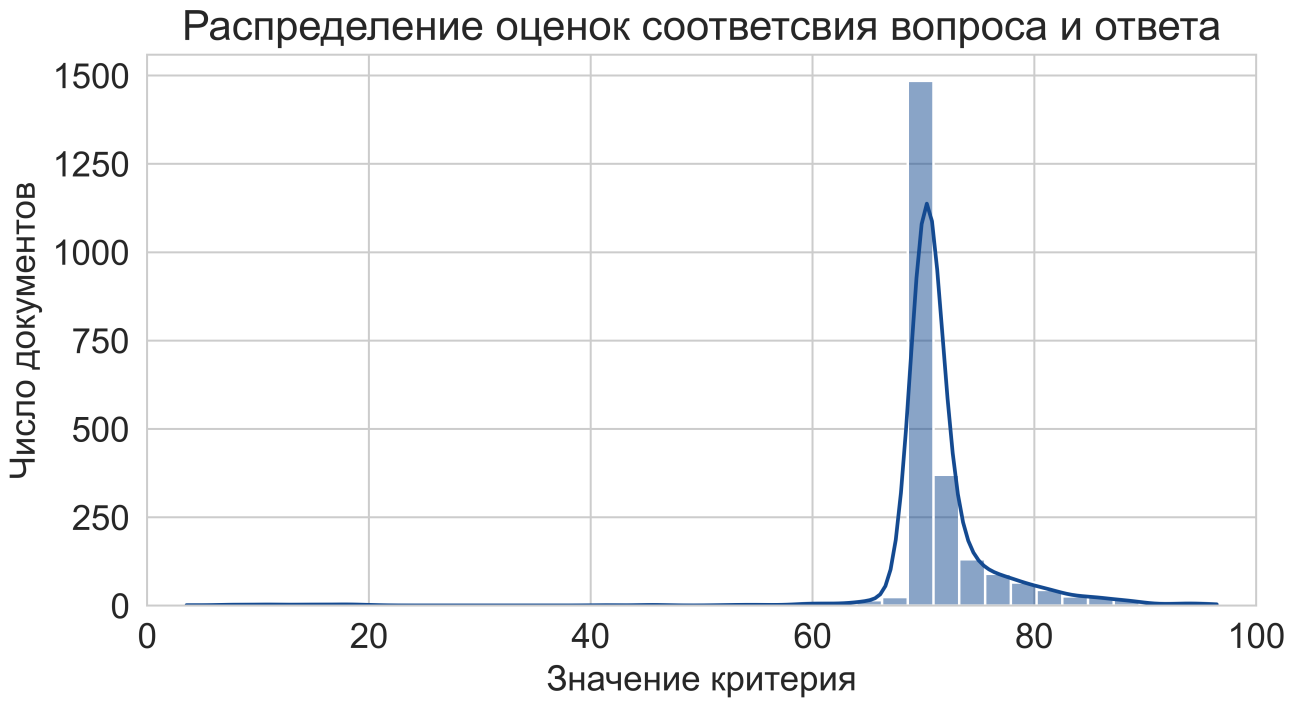


Рисунок 5.2 — Распределение оценок соответствия вопросов и ответов

Модели ответов на вопросы (mdeberta-v3, xlm-roberta) требуют на вход текст вопроса и текст контекста, из которого модель должна выбрать краткий и емкий ответ на вопрос. Представление результатов совпадает для двух моделей — модели получают ответ из контекста и степень уверенности в выборе. Последнее используется в качестве одного из параметров оценки. Для итоговой оценки берется максимальное из двух моделей значение уверенности.

Модель следствия предложений — одна из базовых моделей моделирования языка, позволяет получить вероятность следствия предложений в тексте. В данном случае полагается, что данная модель может дать наиболее простую оценку наличия общего контекста в текстах вопросов и ответов.

Подход, основанный на моделях ответов на вопросы, имеет серьезные недостатки, связанные с возможностью поиска кратких ответов на сложные вопросы. Для общеязыковых данных часто бывает достаточно указать несколько слов из текста, дающих краткий ответ на вопрос, что не всегда справедливо для ответов на сложные вопросы юридического характера. В связи с этим данному параметру критерия был присвоен пониженный вес 0.3. Модели следствия предложений был присвоен вес 0.7. Рисунок 5.2 показывает распределение оценок соответствия вопросов и ответов. Важно отметить, что низкое значение критерия потенциально может быть присвоено подробному ответу на длинный и

сложный вопрос. Данный фактор обусловлен спецификой работы нейросетевых алгоритмов, имеющих ограниченное контекстное окно и спецификой данных.

Анализ продемонстрировал, что в большинстве случаев ответы налоговых органов соответствовали вопросам заявителей. На снижение оценки оказывали влияние такие факторы, как: чрезмерная длина ответа; наличие в ответе информации, прямо не относящейся к вопросу заявителя; цитирование нормативных актов без необходимых разъяснений, отсутствие концентрированного вывода в ответе и т.д.

5.2.4 Перефразирования и цитаты

Для оценки наличия в текстах перефразирований отрывков правовых документов использовались кодировки, полученные моделью Rubert-base-cased. Каждому предложению текста ответа сополагался соответствующий ему числовой вектор. Аналогично создавались вектора кодировок предложений документов:

- Конституция Российской Федерации;
- Налоговый кодекс РФ (части первая и вторая);
- Глава 15 Кодекса Российской Федерации об административных правонарушениях.

Далее для каждого предложения текста ответа получалось максимальное значение сходства (в описываемом случае – косинусного сходства) с текстами приведенных документов. Затем считалась доля предложений, имеющих уровень сходства больше 0.85, но меньше 0.95. Наконец, с использованием подхода ранжирования, основанного на нормальном распределении со сдвигом, получались числовые оценки уровня перефразирования каждого текста ответа.

Таким образом, оценку 100 получали ответы, в которых 20% текста составляют перефразированные версии предложений из документов, представленных в списке. Полностью оригинальным ответам соответствует оценка ≈ 70, ответам, целиком состоящим из перефразирований соответствует оценка близкая 0.

Результирующее распределение оценок наличия перефразирований представлено на рисунке 5.3. Всего 17 документов из всего набора данных определены как перефразированные фрагменты нормативно-правовых актов.

Рисунок 5.3 — Распределение оценок наличия перефразирований

Проведённый анализ показал, что в ответах налоговых органов очень часто приводятся выдержки из нормативных актов. При этом, если учесть и оценку по следующему критерию — “Цитаты”, большинство таких ссылок имеют вид непрямого цитирования, а представлены в относительно перефразированном варианте, адаптированном для целей ответа на вопрос заявителя. Это существенно повышает уровень восприятия ответа. Кроме того, важное значение имеет то, насколько ответ “нагружен” подобными отсылками к нормативным актам. Представленные выше графики демонстрируют, что в большинстве случаев ответ состоит из таких перефразированных цитат на 0-20%. Значительная часть ответов состоит из перефразированных цитат на 20-40%. Обращает на себя внимание и то, что встречаются ответы, которые практически полностью состоят из подобных цитат, что очень негативно сказывается на уровне их трудности для восприятия.

Аналогично модели перефразирований определялись и прямые цитаты. Создавались вектора, представляющие тексты ответов, тесты документов, получались степени сходства предложений. Далее определялась доля предложений текста ответа, для которых максимальная степень сходства превышает 0.95.

Оценку 0 получали ответы, целиком состоящие из прямых цитат, оценку 100 получали ответы, в которых не было обнаружено ни одной прямой цитаты.

Рисунок 5.4 — Распределение оценок наличия цитат

Каждый ответ получил оценку по данной шкале в зависимости от его положения в пределах минимального и максимального значения.

Использование языковой модели вместо прямого сравнения позволяет выявлять прямые цитаты с незначительными изменениями (орфографическими ошибками, минимальными вставками и т.д.).

Результирующее распределение оценок наличия цитат представлено на рисунке 5.4. Конкретнее, 1694 ответа (72%) из проанализированного набора не содержит прямых цитат из нормативно-правовых актов. Текстов ответов, полностью состоящих из цитат, не обнаружено.

Результаты анализа показывают, что в ответах налоговых органов прямое цитирование нормативных актов используется, но делается это в допустимых объёмах, которые не оказывают серьёзного влияния на восприятие содержания ответов. В большинстве ответов прямое цитирование нормативных актов вовсе отсутствует. При этом стоит учитывать распространённость использования перефразированных цитат, описанных выше.

5.2.5 Понятность

Сама по себе понятность и доступность текста ответа для восприятия выступает требованием к языку ответов налоговых органов исходя из того, что понятность текста обеспечивает право налогоплательщика на получение полной и доступной для него информации относительно его прав и обязанностей. Некоторые другие описанные здесь критерии косвенно направлены на оценку понятности текста для восприятия, однако помимо этого следует также предусмотреть интегративный параметр оценки понятности текста ответов.

Для каждого из текста ответов было подсчитано значение каждой из следующих 19 метрик (расположены в порядке убывания важности):

- FRE_GL адаптированная формула Флеша-Кинкейда;
- SMOG адаптированная формула SMOG;
- ARI адаптированная формула подсчёта; автоматизированного индекса читабельности;
- Nouns_pr индекс именной лексики;
- Inan_pr доля неодушевлённых существительных;
- Adjif_pr доля полных прилагательных;
- ACW средняя длина словоформы в буквах;
- Gen_pr доля словоформ в родительном падеже;
- CLI индекс Колман-Лиау;
- word_long_pr доля длинных слов (4 и более слога);
- Adj_pr индекс адъективности;
- ASS средняя длина предложения в слогах;
- Prtf_pr доля полных причастий;
- DCI индекс Дейла-Чейл;
- ASW средняя длина словоформы в слогах;
- Abbr_pr доля аббревиатур;
- TTR_word простой TTR (словоформы);
- N количество числовых символов;
- Prts_pr доля кратких причастий.

“Сложность” понимается как объективный параметр, поддающийся оценке в текстах на естественном языке. “Сложность” текста в свою очередь оказывает прямое влияние на его понятность для конкретного читающего. Для

оценки сложности используются метрики сложности. Выбор метрик обоснован лингвистическим опытом оценки сложности текста.

В приведённый список метрик вошли прежде всего автоматизированные индексы читабельности (простоты/сложности текста для чтения), а именно FRE_GL, SMOG, ARI, CLI, DCI. При вычислении формул читабельности применяется нехитрая логика, согласно которой длинные (синтаксически сложные) предложения читать и интерпретировать труднее, чем короткие; длинные слова (в частном случае – слова длиннее четырёх слогов) сложнее, чем короткие и т.п. При всей простоте подхода формулы читабельности показали свою эффективность в ходе оценке читабельности текста на десятках естественных языков. Стоит добавить, что в представленной схеме оценки сложности для чтения (соответственно, понятности текста для читающего) используются только формулы читабельности, адаптированные к русским текстам.

Кроме того, в списке метрик фигурирует Inan_pr – доля неодушевлённых существительных. Эта метрика, наряду с Nouns_pr (долей именной лексики), призвана уловить введение в тексты понятий, обозначаемых существительными (этот параметр иногда называют “лексической плотностью”). Согласно общей логике, чем больше в тексте понятий, тем он сложнее (этот факт подтверждён целым рядом предшествующих исследований). Стоит добавить, что многие неодушевлённые существительные в текстах юридических документов с большой вероятностью являются абстрактными (непредметными). Экспериментальные лингвистические исследования показывают, что интерпретировать абстрактную лексику сложнее, чем предметные существительные (ср. “стол” vs “власть”). Метрика “доля полных прилагательных “Adjif_pr” также призвана оценивать как синтаксическую, так и понятийную сложность. Так, по модели “прилагательное плюс существительное” образуется ряд терминов и терминоподобных сочетаний (ср. “единый сельскохозяйственный налог”), который в силу своей редкости и специализированности пока не вошёл в общие словари юридических терминов.

Метрики “доля полных причастий (Prtf_pr)” и “доля кратких причастий (Prts_pr)” призваны описывать синтаксическую сложность текстов ответов на обращения. Причастные обороты не считаются предложениями, но лингвисты называют такие обороты “причастными клаузами”. Причастные клаузы, как и предложения, являются предикациями (как и, например, простые предложения с личными формами глаголов, ср. “за исключением случаев, предусмотрен-

ных законодательством" vs "за исключением случаев, которые предусматривает законодательство"). Наличие причастных клауз в общем случае делает синтаксическую структуру текста более сложной, что подтверждается рядом исследований.

Далее, список метрик включает долю словоформ в родительном падеже. Этот параметр косвенно описывает синтаксическую сложность текста, в том числе наличие цепочек существительных в родительном падеже (ср. "неподтверждение возможности исполнения заявки заказчика", "невозможность применения допущения непрерывности деятельности предприятия"), а также вхождения в тексты неодословных терминов и терминоподобных сочетаний, образованных по модели "существительное в именительном падеже + существительное в родительном падеже" (ср. "сумма налога", "форма декларации", "кабинет налогоплательщика").

Метрика "TTR" введена для измерения лексического разнообразия текста (проще говоря, чем выше значения метрики, тем больше в тексте разных слов, и тем меньше слова повторяются). Считается, что в общем случае читать тексты с низким лексическим разнообразием проще.

Метрика "Abbr_pr" (доля аббревиатур) используется потому, что в специальных текстах (в том числе – юридических текстах с разными тематиками) встречается значительное количество аббревиатур, привычных для профессионалов, но незнакомых "простым" носителям языка. Эти аббревиатуры в тексте часто не поясняются, однако их интерпретация очевидно затруднена (ср., например, такие аббревиатуры, как ЦБСН, ГУНР, СППФД, ИСИНПОЛ, КНП, ИНВ, РНИ, СНВ, АРНУ и мн. др.).

Метрика "N" (количество числовых символов) призвана зафиксировать в тексте наличие нумерованных списков, которые в юридических текстах имеют тенденцию быть громоздкими и обширными. Кроме того, метрика косвенно отражает наличие разнообразных ссылок на положения правовых актов (ср. "полученными из Министерства финансов Российской Федерации разъяснениями по данному вопросу (письмо Минфина России от 12. 11. 2004 № 03-03-02-02/10)") и вхождения в тексты количественных выражений.

Наконец, в списке метрик присутствуют легко интерпретируемые показатели (средняя длина словоформы в буквах, доля длинных слов, средняя длина предложения в слогах), обоснование применения которых изложена в выше в пассаже о формулах читабельности.

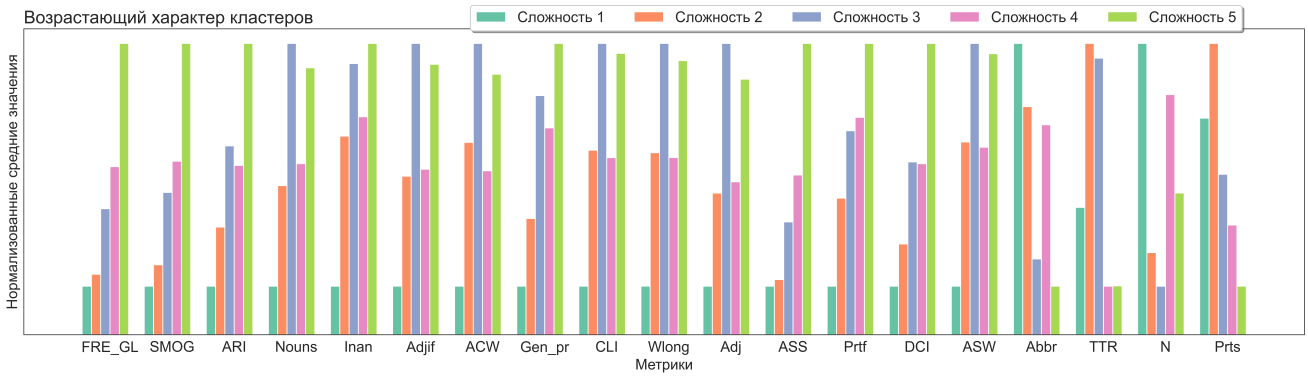


Рисунок 5.5 — Возрастающий характер сложности кластеров в выражении метрик

При оценке ответов ответов была использована кластеризация. После подсчёта значения указанных метрик была использована модель кластеризации KMeans т.к. изначально предполагаемая модель HDBSCAN показала неудовлетворительные результаты со всеми комбинациями гиперпараметров, несмотря на высокую эффективность в задачах анализа текстов [6], ответы были распределены в 5 кластеров.

Рисунок 5.5 показывает возрастающий характер сложности кластеров. Исходя из визуализации видно, что наиболее важные критерии (первые из списка) имеют прямую корреляцию с индексом сложности. Данное правило не выполняется для последних критериев, что говорит об их незначительном влиянии на оценку сложности в рассматриваемом наборе данных.

Краткая интерпретация значений трёх метрик, вносящих наибольший вклад в оценку сложности текстов ответов:

– **FRE_GL** (адаптированная формула читабельности Флеша-Кинкейда). В качестве переменных в формуле используются значения ASL (средняя длина предложения в словах) и ASW (средняя длина словоформ в слогах). Высокие значения метрики в общем случае значат, что тексты содержат много длинных (соответственно, синтаксически сложных) предложений и много длинных (4 и более слога) слов. Значение этой метрики, если смотреть на весь проанализированный набор данных, варьирует в широком диапазоне от 1,56 до 73,64;
– **SMOG** (адаптированная формула читабельности Simple Measure of Gobbledygook). Переменными в формуле являются значения “количества предложений” и “количества длинных слов”. Значение SMOG

вычисляется как отношение числа длинных (соответственно, сложных) слов к числу предложений. Таким образом, метрика способна находить предложения с большим количеством длинных слов;

– **ARI** (адаптированная формула подсчёта автоматизированного индекса читабельности). В качестве переменных в формуле используются количество знаков, количество слов и количество предложений текста. Точнее, в индексе учтено отношение количества знаков к числу слов и отношение числа слов к числу предложений. Сказанное значит, что снова рассматриваются длины слов и предложений, однако при оценке длины слов используются не слоги, а знаки (что позволяет оценить длину цифровых и буквенно-цифровых комплексов типа 03-04-07/102199).

Оценка понятности рассчитывалась обратно пропорционально индексу кластера сложности — ответы, попавшие в кластер наибольшей сложности, получали оценку 0, в кластер низкой сложности — оценку 100.

В кластер наибольшей сложности (с оценкой “0”) попало 299 текстов ответов (13%), в кластер повышенной сложности (с оценкой “25”) – 613 текстов ответов (26%), в кластер средней сложности (с оценкой “50”) – 446 текстов ответов (19%), в кластер пониженной сложности (с оценкой “75”) – 635 текстов ответов (27%), в кластер низкой сложности (с оценкой “100”) – 345 текстов ответов (15%).

Интерпретация значения метрик для конкретных текстов ответов позволила установить следующие основные факторы, влияющие на понятность текстов ответов.

1. Формирование текстов ответов из неперефразированных фрагментов нормативно-правовых актов;
2. Использование шаблонных выражений (зачастую длинных и неудобопонятных), которые без потери смысла могут быть заменены на более короткие и простые выражения (ср. “по телекоммуникационным каналам связи” vs. “по электронной почте”);
3. Использование длинных перечислительных (“сочинённых”) рядов, к тому же не оформленных в виде нумерованных списков, способных сколько-нибудь облегчить понимание;
4. Использование лексических (словесных) повторов, которые увеличивают длину предложений в словах и могут быть удалены из текста без потери смысла.

Проведенный анализ продемонстрировал высокую эффективность используемых механизмов. При этом обратило на себя внимание обобщение оценок по регионам. Данный механизм может не только выявлять отдельные “непонятные” ответы, но и на большом объёме данных определять “проблемные” участки — это могут быть любые задействованные элементы — от региональных управлений, до конкретных специалистов, подготовивших ответ (при условии наличия в представленных для анализа материалах этих данных).

5.2.6 Комбинированная оценка

Для получения итоговой оценки, сочетающей критерии, в соответствии с их важностью и качеством оценки, каждому критерию задавался соответствующий вес.

Критерий понятности получил вес 0.2, критерий наличия юридических терминов — 0.1, критерии перефразирования и цитирования получили вес 0.15, оставшиеся критерии получили вес 0.05. На рисунке 5.6 представлено итоговое

Рисунок 5.6 — Распределение итоговых комбинированных оценок

распределение комбинированных оценок ответов. Таким образом, проанализи-

рованные ответы налоговых органов имеют высокое коммуникативное качество (показатель от 80 до 100) в 58% ответов. Среднее коммуникативное качество ответов (показатель от 60% до 80%) наблюдается для 41%. Приемлемое коммуникативное качество (показатель от 40% до 60%) имеет менее 1% ответов.

Заключение

Анализ юридических текстов является актуальным, важным для страны направлением. Развитие информационно-правовых технологий LegalTech является перспективным направлением, важным как в научном контексте, так и в практическом.

На основании проведенного исследования были сделаны следующие выводы, касающиеся разработки, актуальности и применимости различных методологических подходов к обработке юридических документов.

Представленный гибридный подход, объединяющий традиционные методы анализа текстовых данных и современные нейросетевые подходы языкового моделирования, позволяет давать качественную оценку понятности. Традиционные методы статистического анализа и машинного обучения дают возможность выделить языковые характеристики, оказывающие наибольшее влияние на итоговое значение понятности. Таким образом методология объединяет точность предсказаний нейросетевых подходов и интерпретируемость классических подходов.

Приведенные в данной работе методы и программные решения нашли применение в исследовательских проектах, посвященным вопросам анализа сложности различных видов юридических документов. Отдельные методы показали высокую эффективность в практической задаче анализа текстов ответов на вопросы в юридической и экономической сфере.

Представленная методология учитывает различные особенности юридических документов, такие как специфичность языка и структуры документов, малое количество и разнообразие общедоступных данных, большое разнообразие текстов различных направленностей. Данная особенность позволяет адаптировать процесс и отдельные методы анализа для задач, имеющих схожие характеристики.

Цель работы заключалась в разработке и апробации методологических и инструментальных средств интеллектуальной обработки юридических текстов и алгоритмическое обеспечение процесса определения доступности их восприятия. Для выполнения цели были решены следующие задачи:

- Изучено современное состояние юридических, лингвистических исследований в области анализа юридических документов, выявлены актуальные проблемы и определены методы их решения.
- Разработаны методологические подходы для сбора, обработки и семантического анализа русского правового языка.
- Разработана методология статистической оценки частотных характеристик юридического языка.
- Выявлены и отобраны языковые характеристики юридических документов, наиболее полно описывающие их в контексте сложности и доступности восприятия.
- Разработан комплекс анализа сложности юридических документов на основе методов гибридного использования языковых моделей.
- Проведен сравнительный анализ сложности документов различных подстилей и жанров с использованием гибридной оценки сложности.
- Проведен практический анализ доступности восприятия юридических текстов с использованием представленных моделей и методов.

Автор выражает благодарность и большую признательность научному руководителю Блеканову Ивану Станиславовичу за поддержку, помощь в создании, обсуждении результатов и научное руководство. Также автор благодарит Блинову Ольгу Владимировну за руководство проектом “Понятность официального русского языка: юридическая и лингвистическая проблематика”, экспертную лингвистическую оценку качества методов и соавторство в ключевых статьях, ставших основой данной диссертационной работы. Автор благодарит руководителя научно-исследовательского института проблем государственного языка Белова Сергея Александровича за консультации и экспертную юридическую оценку. Бодрунову Светлану Сергеевну за совместную работу в проектах, посвященных семантическому анализу текстовых данных. Автор благодарит всех, кто сделал настоящую работу возможной.

Список литературы

1. *Свидетельство о гос. регистрации программы для ЭВМ.* Программа для выявления эхо-камер в дискуссиях социальных медиа-платформ на основе анализа поляризации пользовательских мнений (SNAOpinionPolariz) [Текст] / И. С. Блеканов, Н. А. Тарасов, С. С. Бодрунова ; Роспатент. — № 2023685490 ; заявл. 27.11.2023 (Рос. Федерация).
2. *Свидетельство о гос. регистрации программы для ЭВМ.* Программа для автоматической суммаризации пользовательских сообщений в дискуссиях социальных сетей (SNAPostSummarizer) [Текст] / И. С. Блеканов, Н. А. Тарасов ; Роспатент. — № 2021680151 ; заявл. 21.11.2021 (Рос. Федерация).
3. *Свидетельство о гос. регистрации программы для ЭВМ.* Программа для автоматического обнаружения скрытых тем в пользовательских дискуссиях социальных сетей (SNATopicDetector) [Текст] / И. С. Блеканов, Н. А. Тарасов ; Роспатент. — № 2020662702 ; опубл. 16.10.2020 (Рос. Федерация).
4. *Blinova, O.* A hybrid model of complexity estimation: Evidence from Russian legal texts [Текст] / O. Blinova, N. Tarasov // Frontiers in Artificial Intelligence. — 2022. — Т. 5. — С. 1008530.
5. *Blekanov, I. S.* Transformer-based abstractive summarization for Reddit and Twitter: single posts vs. comment pools in three languages [Текст] / I. S. Blekanov, N. Tarasov, S. S. Bodrunova // Future Internet. — 2022. — Т. 14, № 3. — С. 69.
6. Topic detection based on sentence embeddings and agglomerative clustering with Markov moment [Текст] / S. S. Bodrunova [и др.] // Future Internet. — 2020. — Т. 12, № 9. — С. 144.
7. *Блинова О. В .and Тарасов, Н. А.* Language Complexity across Sub-Styles and Genres in Legal Russian [Текст] / Н. А. Блинова О. В .and Тарасов // Research Result. Theoretical and Applied Linguistics. — 2023. — Т. 9, № 2. — С. 73—96.

8. Data Encoding for Social Media: Comparing Twitter, Reddit, and Telegram [Текст] / I. S. Blekanov [и др.] // Fifth Networks in the Global World Conference. — Springer. 2022. — С. 114—122.

9. Mapping opinion cumulation: topic modeling-based dynamic summarization of user discussions on social networks [Текст] / I. S. Blekanov [и др.] // International Conference on Human-Computer Interaction. — Springer. 2023. — С. 25—40.

10. *Блинова О. В .and Тарасов, Н. А.* Complexity of russian legal texts: assessment methods and language data [Текст] / Н. А. Блинова О. В .and Тарасов // Труды международной конференции"Корпусная лингвистика-2021". — 2021. — С. 175.

11. Modeling lemma frequency bands for lexical complexity assessment of russian texts [Текст] / O. Blinova [и др.] // Comput. Linguist. Intell. Technol. — 2020. — Т. 19. — С. 76—92.

12. *Блинова О. В .and Тарасов, Н. А.* Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности [Текст] / Н. А. Блинова О. В .and Тарасов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 21, дополнительный том. — Российская Федерация : Российский государственный гуманитарный университет, 2022. — С. 1017—1028. — (Компьютерная лингвистика и интеллектуальные технологии).

13. *Reynolds, R. J.* Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories [Текст] / R. J. Reynolds // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. — 2016. — С. 289—300.

14. *Collins-Thompson, K.* Computational assessment of text readability: a survey of current and future research [Текст] / K. Collins-Thompson. — 2014.

15. *Crossley, S. A.* Moving beyond classic readability formulas: new methods and new models [Текст] / S. A. Crossley, S. Skalicky, M. Dascalu // Journal of Research in Reading. — 2019. — Т. 42, № 3/4. — С. 541—561.

16. *Benjamin, R. G.* Reconstructing readability: recent developments and recommendations in the analysis of text difficulty [Текст] / R. G. Benjamin // Educational Psychology Review. — 2012. — Т. 24(1). — С. 63—88.

17. *Schwarm, S. E.* Reading level assessment using support vector machines and statistical language models [Текст] / S. E. Schwarm, M. Ostendorf // 05) / под ред. P. of the 43rd Annual Meeting on Association for Computational Linguistic 2005. — С. 523—530.

18. *Leroy, G.* The effect of word familiarity on actual and perceived text difficulty [Текст] / G. Leroy, D. Kauchak // Journal of the American Medical Informatics Association. — 2014. — Т. 21, e1. — e169—e172.

19. *Laposhina, A. N.* Analysis of the relevant features for automatic readability assessment for texts in Russian as a foreign language [Analiz relevantnyh priznakov dlja avtomaticheskogo opredelenija slozhnosti russkogo teksta kak inostrannogo] [Текст] / A. N. Laposhina. — 2017. — URL: http://www.dialog-21.ru/media/3993/ ; Proceedings of the International, Proceedings of the International Conference "Dialogue 2017" [Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Bekasovo.

20. *Ivanov, V. V.* Efficiency of text readability features in Russian academic texts [Текст] / V. V. Ivanov, M. I. Solnyshkina, V. D. Solovyev // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2018. — Т. 17. — С. 277—287.

21. *Sharoff, S.* Seeking needles in the web haystack: Finding texts suitable for language learners [Текст] / S. Sharoff, S. Kurella, A. Hartley // Proceedings of 8th Teaching and Language Corpora Conference (TaLC-8. — 2008.

22. *Solovyev, V.* Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics [Текст] / V. Solovyev, V. Ivanov, M. Solnyshkina // Journal of Intelligent & Fuzzy Systems. — 2018. — Т. 34. — С. 3049—3058.

23. *Collins-Thompson, K.* Predicting Reading Difficulty with Statistical Language Models [Текст] / K. Collins-Thompson, J. Callan // Journal of the American Society for Information Science and Technology. — 2005. — Т. 56, № 13. — С. 1448—1462.

24. *Chen, X.* Characterizing Text Difficulty with Word Frequencies [Текст] / X. Chen, W. D. Meurers // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. — 2016. — С. 84—94.

25. *Atkins, S.* Corpus Design Criteria [Текст] / S. Atkins, J. Clear, N. Ostler // Literary and Linguistic Computing. — 1992. — Т. 7. — С. 1—16.

26. *Biber, D.* Representativeness in Corpus Design [Текст] / D. Biber // Literary and Linguistic Computing. — 1993. — Т. 8, № 4. — С. 243—257.

27. *Brysbaert, M.* The Word Frequency Effect in Word Processing: An Updated Review, Current Directions in [Текст] / M. Brysbaert, P. Mandera, E. Keuleers // Psychological Science. — 2018. — Т. 27. — С. 45—50.

28. *Lyashevskaya, O. N.* Corpus Instruments for Russian Grammar Studies [Korpusnye instrumenty v grammaticheskih issledovanijah russkogo jazyka], Jazyki slavjanskoj kul'tury [Текст] / O. N. Lyashevskaya. — 2016. — (Moscow).

29. *Schmitt, N.* Researching vocabulary: a vocabulary research manual [Текст] / N. Schmitt. — Basingstoke, UK : Palgrave Macmillan, 2010.

30. *Zhao, Y.* The effect of lexical frequency and Lombard reflex on tone hyperarticulation [Текст] / Y. Zhao, D. Jurafsky // Journal of Phonetics. — 2009. — Т. 37. — С. 231—247.

31. Predictability effects on durations of content and function words in conversational English [Текст] / A. Bell [и др.] // Journal of Memory and Language. — 2009. — Т. 60. — С. 92—111.

32. ruTenTen11 [Текст]. — URL: https://www.sketchengine.eu/rutenten-russian-corpus/.

33. The Sketch Engine: Ten Years On [Текст] / A. Kilgarriff [и др.] // Lexicography. — 2014. — Т. 1, Iss. 1. — С. 7—36.

34. *Russicum III, A. M.* / A. M. Russicum III. — URL: http://ucts.uniba.sk/aranea%5C_about/%5C_russicum.html.

35. *Benko, V.* Aranea: Yet Another Family of (Comparable) Web Corpora [Текст] / V. Benko // Text, Speech and Dialogue. 17th International Conference, TSD 2014. Proceedings. LNCS 8655. Switzerland / под ред. P. Sojka [и др.]. — Springer International Publishing, 2014. — С. 257—264.

36. *Corpus, T.* An open-source corpus for machine learning [Текст] / T. Corpus. — URL: https://tatianashavrina.github.io/taiga%5C_site/.

37. *Corpus, R. N.* / R. N. Corpus. — URL: http://www.ruscorpora.ru/new/.

38. *Lyashevskaya, O. N.* The frequency dictionary of modern Russian language [Častotnyj slovar' sovremennogo russkogo jazyka] [Текст] / O. N. Lyashevskaya, S. A. Sharoff. — 2009. — URL: http://dict.ruslang.ru/freq.php ; csv-version.

39. *Kilgarriff, A.* Measures for corpus similarity and homogeneity [Текст] / A. Kilgarriff, T. Rose // Proceedings of the Third Conference on Empirical Methods for Natural Language Processing. — Spain : Granada, 1998. — С. 46—52.

40. *Piperski, A. C.* Corpus Size and the Robustness of Measures of Corpus Distance, Computational Linguistics and Intellectual Technologies [Текст] / A. C. Piperski // Dialogue 2018 / под ред. P. of the International Conference. — 2018. — С. 578—589.

41. *Gomaa, W. H.* A Survey of Text Similarity Approaches [Текст] / W. H. Gomaa, A. A. Fahmy // International Journal of Computer Applications. — 2013. — Т. 68. — С. 13—18.

42. The WaCky wide web: a collection of very large linguistically processed webcrawled corpora, [Текст] / M. Baroni [и др.] // Language Resources and Evaluation. — 2009. — Т. 43. — С. 209—226.

43. Corpus Linguistics: An International Handbook [Текст]. Т. 2 / под ред. A. Lüdeling, M. Kytö. — Berlin, Boston : De Gruyter Mouton, 2009.

44. *Shaikevich, A. Y.* Measures of lexical similarity between frequency dictionaries [Mery leksicheskogo shodstva chastotnyh slovarej] [Текст] / A. Y. Shaikevich // Corpus linguistics-2015 / под ред. P. of the Internationational C Saint Petersburg : Trudy mezhdunarodnoy nauchnoy konferentsii "Korpusnaya linguistica-2015"], 2015. — С. 434—442.

45. *Piperski, A.* Sum of Minimum Frequencies as a Measure of Corpus Similarity [Текст] / A. Piperski // Presented at the Corpus Linguistics 2017, Birmingham. — 2017.

46. Subtlex-UK: A new and improved word frequency database for British English [Текст] / W. J. B. Van Heuven [и др.] // Quarterly Journal of Experimental Psychology. — 2014. — Т. 67. — С. 1176—1190.

47. *Sharoff, S.* Frequency Dictionary: Russian, Quasthoff U., Fiedler S., Hallsteindóttir E. (eds.), Frequency Dictinaries 9, Leipziger Universitätsverlag [Текст] / S. Sharoff, D. Goldhahn, U. Quasthoff. — 2017.

48. *Jamieson, S.* Likert scales: how to abuse them [Текст] / S. Jamieson // Medical Education. — 2004. — Т. 38, № 12. — С. 1217—1218.

49. *Peter, M. T.* Legal Language [Текст] / M. T. Peter. — Chicago, London : The University of Chicago Press, 1999.

50. *Heikki, E. S. M.* Comparative legal linguistics: language of law, Latin and modern lingua francas [Текст] / E. S. M. Heikki // Ashgate Publishing, Ltd., Farnham, Surrey. — 2013. — Т. 2 edition.

51. *Sol, A.-A.* On drafting, interpreting, and translating legal texts across languages and cultures [Текст] / A.-A. Sol, Y. Ning // International Journal of Legal Discourse. — 2017. — Т. 2, № 1. — С. 1—12.

52. *Vijay, K. B.* Cognitive structuring in legislative provisions [Текст] / K. B. Vijay, G. John // Language and the Law. — 1994. — С. 136—155.

53. *языкознании, С. текста: этапы изучения в отечественном прикладном.* Солнышкина, М. И. and Кисельников, А. С. [Текст] / С. текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. — 2015. — Т. 6, № 38. — С. 86—99.

54. *Juhan, T.* The development of statistical stylistics (a survey) [Текст] / T. Juhan // Journal of Quantitative Linguistics. — 2017. — Т. 11, № 1/2. — С. 141—151.

55. *Голуб, И. Б.* Стилистика русского языка [Текст] / И. Б. Голуб. — Москва : Рольф, 2001.

56. *Кожина, М. Н.* Стилистика русского языка [Текст] / М. Н. Кожина, Л. Дускаева, В. А. Салимовский. — Москва : Флинта, Наука, 2011.

57. *Дружкин, К. Ю.* Метрики удобочитаемости для русского языка [Текст] / К. Ю. Дружкин. — НИУ ВШЭ, Москва : выпускная кфалификационная работа магистра, 2016.

58. *Richard, C. W.* Plain English for lawyers [Текст] / C. W. Richard, E. S. Amy. — 6-е изд. — LLC, Durham, North Carolina : Carolina Academic Press, 2019.

59. *Robert, P. C.* Making legal language understandable: A psycholinguistic study of jury instructions [Текст] / P. C. Robert, R. C. Veda // Columbia Law Review. — 1979. — Т. 79, № 7. — С. 1306—1374.

60. *Marina, S.* Readability formula for russian texts: A modified version [Текст] / S. Marina, I. Vladimir, S. Valery // Proceedings of the 17th Mexican International Conference on Artificial Intelligence. — MICAI 2018, 2018. — С. 132—145.

61. *Бегтин, И.* Plainrussian.ru [Текст] / И. Бегтин. — 2016. — URL: https://github.com/ivbeg/readability.io.

62. *Milan, S.* Universal dependencies 2.5 models for UDPipe (2019-12-06) [Текст] / S. Milan, S. Jana. — Faculty of Mathematics, Physics, Charles University : LINDAT/CLARIAH-CZ digital library at the Institute of Formal, Applied Linguistics (ÚFAL), 2019.

63. *Mikhail, K.* Morphological analyzer and generator for russian and ukrainian languages. // Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry Ignatov, and Valeri G [Текст] / K. Mikhail // of Images, Social Networks and Texts, P / под ред. A. Labunets. — Cham : Springer International Publishing, 2015. — С. 320—332.

64. *Журавлев, А. Ф.* Опыт квантитативно-типологического исследования разновидностей устной речи [Текст] / А. Ф. Журавлев // Разновидности городской устной речи. — 1988. — С. 84—150.

65. *Xiao, p. T.* Automatic genre classification via n-gr of part-of-speech tags [Текст] / p. T. Xiao, C. Jing // AMS Procedia - Social and Behavioral Sciences, 198. — 2015. — С. 474—478.

66. *Антонова, А. Ю.* Определение стилевых и жанровых характеристик коллекций текстов на основе частеречной сочетаемости. [Текст] / А. Ю. Антонова, Э. С. Клышинский, Е. В. Ягунова // Труды международной конференции «Корпусная лингвистика–2011». — Санкт-Петербург. СПбГУ, 2011. — С. 80—85.

67. *Нагель, О. В.* Словообразовательные механизмы в процессах восприятия, идентификации и использования языка [Текст] / О. В. Нагель. — Томский государственный университет, Томск : дисс. . . . докт. филол. наук, 2017. — (автореф).

68. *Tianqi, C.* Xgboost: A scalable tree boosting system [Текст] / C. Tianqi, G. Carlos // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — P, 2016. — С. 785—794.

69. Universal sentence encoder [Текст] / C. Daniel [и др.]. — 2018. — arXiv preprint.

70. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms [Текст] / J. Bergstra, D. Yamins, D. D. Cox [и др.] // Proceedings of the 12th Python in science conference. Т. 13. — Citeseer. 2013. — С. 20.

71. *Payam, R.* Lei Tang, and Huan Liu [Текст] / R. Payam // Cross-validation. Encyclopedia of database systems. — 2016. — С. 1—7.

72. *Valery, S.* Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics [Текст] / S. Valery, I. Vladimir, S. Marina // Journal of Intelligent Fuzzy Systems. — 2018. — Т. 34. — С. 3049—3058.

73. Prediction of reading difficulty in Russian academic texts [Текст] / S. Valery [и др.] // Journal of Intelligent Fuzzy Systems. — 2019. — Т. 36. — С. 4553—4563.

74. *Deutsch, T.* Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories [Текст] / T. Deutsch, M. Jasbi, S. Shieber // Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications / под ред. J. Tetreault [и др.]. — Association for Computational Linguistics, 2020. — С. 1—17.

75. *Szmrecsanyi, B.* Introduction: Linguistic complexity: Second Language Acquisition, indigenization, contact [Текст] / B. Szmrecsanyi, B. Kortmann // Linguistic Complexity: Second Language Acquisition, Indigenization, Contact / под ред. B. Kortmann, B. Szmrecsanyi. — Berlin, Boston : De Gruyter, 2012. — С. 6—34.

76. *Dahl, ö.* The growth and maintenance of linguistic complexity [Текст] / ö. Dahl. — Amsterdam : John Benjamins Publishing, 1993.

77. *Nichols, J.* Linguistic complexity : a comprehensive definition and survey [Текст] / J. Nichols // Language complexity as an evolving variable / под ред. G. Sampson, D. Gil, P. Trudgill. — Oxford : Oxford University Press, 2009. — С. 110—125.

78. *Trudgill, P.* Sociolinguistic typology: Social determinants of linguistic complexity [Текст] / P. Trudgill. — Oxford : Oxford University Press, 2011.

79. *McWhorter, J.* The worlds simplest grammars are creole grammar [Текст] / J. McWhorter // Linguistic Typology. — 2001. — Т. 5, № 2/3. — С. 125—166. — URL: https://www.degruyter.com/document/doi/10.1515/lity.2001.001/html.

80. *Frazier, L.* Syntactic complexity [Текст] / L. Frazier // Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives / под ред. D. R. Dowty, L. Karttunen, A. M. Zwicky. — Cambridge : Cambridge University Press, 1985. — С. 129—189.

81. *Collins-Thompson, K.* Computational assessment of text readability: a survey of current and future research [Текст] / K. Collins-Thompson // Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics. — 2014. — Т. 165, № 2. — С. 97—135. — URL: https://benjamins.com/catalog/itl.165.2.01col.

82. Text complexity as interdisciplinary problem [Текст] / M. Solnyshkina [и др.] // Voprosy Kognitivnoy Lingvistiki. — 2022. — № 1. — С. 18—39. — URL: http://vcl.ralk.info/issues/2022/vypusk-1-2022/slozhnost-teksta-kak-mezhdistsiplinarnaya-problema.html.

83. *Tiersma, P. M.* Legal Language [Текст] / P. M. Tiersma. — Chicago, London : The University of Chicago Press, 1999.

84. *Azuelos-Atias, S.* On drafting, interpreting, and translating legal texts across languages and cultures [Текст] / S. Azuelos-Atias, N. Ye // International Journal of Legal Discourse. — 2017. — Т. 2, № 1. — С. 1—12. — URL: https://www.degruyter.com/document/doi/10.1515/ijld-2017-1000/html.

85. *Wydick, R. C.* Plain English for lawyers [Текст] / R. C. Wydick, A. E. Sloan. — Durham, North Carolina : Carolina Academic Press, LLC, 2019.

86. *Dmitrieva, A. V.* "The art of legal writing": A quantitative analysis of Russian Constitutional Court rulings [Текст] / A. V. Dmitrieva // Sravnitel'noe konstitutsionnoe obozrenie. — 2017. — Т. 118, № 3. — С. 125—133. — URL: https://sko-journal.ru/catalog/sko-3-118-2017/iskusstvo-yuridicheskogo-pisma-kolichestvennyj-analiz-reshenij-konstitutsionnogo-suda-rossii/.

87. *Oborneva, I. V.* Automation of text perception quality assessments [Текст] / I. V. Oborneva // Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta. — 2005. — № 5. — С. 86—91. — URL: https://www.elibrary.ru/item.asp?id=12804809.

88. *Kuchakov, R.* The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note [Текст] / R. Kuchakov, D. Savel'ev. — Saint Petersburg : European University at Saint Petersburg, 2018.

89. *Savel'ev, D.* Decisions of arbitration courts of Russian Federation: lexical and syntactic quality of texts, analytic note [Текст] / D. Savel'ev, R. Kuchakov. — Saint Petersburg : European University at Saint Petersburg, 2019.

90. Complexity of Russian Laws. The Experience of Syntactic Analysis [Текст] / A. Knutov [и др.]. — Moscow : HSE University Publishing House, 2020.

91. *Collins-Thompson, K.* A language modeling approach to predicting reading difficulty [Текст] / K. Collins-Thompson, J. P. Callan // Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004. — 2004. — С. 193—200.

92. A comparison of features for automatic readability assessment [Текст] / L. Feng [и др.] // COLING'10: Proceedings of the 23rd International Conference on Computational Linguistics / под ред. С. 2. О. Committee. — International Committee on Computational Linguistics, 2010. — С. 276—284.

93. *Xia, M.* Text readability assessment for second language learners [Текст] / M. Xia, E. Kochmar, T. Briscoe // arXiv preprint arXiv:1906.07580. — 2019.

94. Automated Assessment of Language Proficiency on German Data [Текст] / E. Szügyi [и др.] // KONVENS. — 2019. — С. 41—50.

95. Automatic classification of text complexity [Текст] / V. Santucci [и др.] // Applied Sciences. — 2020. — Т. 10, № 20. — С. 7285.

96. *Lyashevskaya, O.* Automated assessment of learner text complexity [Текст] / O. Lyashevskaya, I. Panteleeva, O. Vinogradova // Assessing Writing. — 2021. — Т. 49. — С. 100529.

97. *Staudemeyer, R. C.* Understanding LSTM–a tutorial into long short-term memory recurrent neural networks [Текст] / R. C. Staudemeyer, E. R. Morris // arXiv preprint arXiv:1909.09586. — 2019.

98. *Morozov, D. A.* Text complexity and linguistic features: Their correlation in English and Russian [Текст] / D. A. Morozov, A. V. Glazkova, B. L. Iomdin // Russian Journal of Linguistics. — 2022. — Т. 26, № 2. — С. 426—448.

99. *Sharoff, S. A.* What neural networks know about linguistic complexity [Текст] / S. A. Sharoff // Russian Journal of Linguistics. — 2022. — Т. 26, № 2. — С. 371—390.

100. Efficient estimation of word representations in vector space [Текст] / T. Mikolov [и др.] // arXiv preprint arXiv:1301.3781. — 2013.

101. *Pennington, J.* Glove: Global vectors for word representation [Текст] / J. Pennington, R. Socher, C. D. Manning // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — С. 1532—1543.

102. Enriching word vectors with subword information [Текст] / P. Bojanowski [и др.] // Transactions of the association for computational linguistics. — 2017. — Т. 5. — С. 135—146.

103. *Bosco, G. L.* A neural network model for the evaluation of text complexity in Italian language: a representation point of view [Текст] / G. L. Bosco, G. Pilato, D. Schicchi // Procedia computer science. — 2018. — Т. 145. — С. 464—470.

104. A Transfer Learning Based Model for Text Readability Assessment in German [Текст] / S. Mohtaj [и др.] // arXiv preprint arXiv:2207.06265. — 2022.

105. Bert: Pre-training of deep bidirectional transformers for language understanding [Текст] / J. Devlin [и др.] // arXiv preprint arXiv:1810.04805. — 2018.

106. *Dmitrieva, A.* A Comparative Study of Educational Texts for Native, Foreign, and Bilingual Young Speakers of Russian: Are Simplified Texts Equally Simple? [Текст] / A. Dmitrieva, A. Laposhina, M. Lebedeva // Frontiers in Psychology. — 2021. — Т. 12. — URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703690.

107. Consultant Plus: Legal Reference System [Текст]. — 2022. — URL: http://www.consultant.ru ; Accessed August 30, 2022.

108. Garant: Legal information portal [Текст]. — 2022. — URL: https://www.garant.ru/ ; Accessed August 30, 2022.

109. *Ivanov, V.* Efficiency of text readability features in Russian academic texts [Текст] / V. Ivanov, M. Solnyshkina, V. Solovyev // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii 2018 (Computational Linguistics and Intellectual Technologies 2018). — 2018. — Т. 17, № 24. — С. 284—293. — URL: https://www.dialog-21.ru/media/4302/ivanovvv.pdf.

110. Using Universal Dependencies in cross-linguistic complexity research [Текст] / A. Berdicevskis [и др.] // Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). — Association for Computational Linguistics, 2018. — С. 8—17.

111. *Korobov, M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages [Текст] / M. Korobov // Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science / под ред. M. Y. Khachay [и др.]. — Springer International Publishing, 2015. — С. 320—332.

112. CoNLL 2018 Shared Task [Текст]. — 2018. — URL: https : / / universaldependencies . org / conll18 / evaluation . html ; Accessed August 30, 2022.

113. *Druzhkin, K.* Readability metrics for Russian: master's thesis [Текст] / K. Druzhkin. — Moscow : Higher School of Economics, 2016.

114. *Benjamin, R.* Reconstructing readability: recent developments and recommendations in the analysis of text difficulty [Текст] / R. Benjamin // Educational Psychology Review. — 2012. — № 24. — С. 63—88. — URL: https://link.springer.com/article/10.1007/s10648-011-9181-8.

115. *Solnyshkina, M.* Readability Formula for Russian Texts: A Modified Version [Текст] / M. Solnyshkina, V. Ivanov, V. Solovyev // Advances in Computational Intelligence. MICAI 2018. Lecture Notes in Computer Science. — 2018. — Т. 11289. — С. 132—145. — URL: https://link.springer.com/chapter/10.1007/978-3-030-04497-8%5C_11.

116. *Begtin, I.* PlainRussian [Текст] / I. Begtin. — 2016. — URL: https://github.com/ivbeg/readability.io.

117. *Straka, M.* Universal Dependencies 2.5 Models for UDPipe [Текст] / M. Straka, J. Straková. — 2016. — URL: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131.

118. *Zhuravlev, A. F.* Experience of quantitative-typological study of varieties of oral speech [Текст] / A. F. Zhuravlev // Raznovidnosti gorodskoi ustnoi rechi. Sbornik nauchnykh trudov / под ред. D. Shmelev, E. Zemskaja. — Moscow : Nauka, 1988. — С. 84—150.

119. Formation of a model of compatibility of Russian words and the study of its propertie [Текст] / J. S. Klyshinskij [и др.]. — Moscow : Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, 2013.

120. *Antonova, A. J.* Determination of stylistic and genre characteristics of text collections based on part-of-speech compatibility [Текст] / A. J. Antonova, E. S. Klyshinsky, E. V. Jagunova // Trudy mezhdunarodnoj konferencii "Korpusnaja lingvistika–2011" / под ред. V. P. Zaharov. — Saint Petersburg State University, 2011. — С. 80—85.

121. *Dobrego, A.* Processing of static and dynamic texts: an eye-tracking study of Russian [Текст] / A. Dobrego, T. Petrova // 3rd International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2016. Т. 1.1 / под ред. S. editorial board. — STEF92 Technology, 2016. — С. 991—998.

122. *Nagel', O. V.* Word-formation mechanisms in the processes of perception, identification, and use of language: author's abstract of the doctor's thesis [Текст] / O. V. Nagel'. — Tomsk : National Research Tomsk State University, 2017.

123. *Kyle, K.* Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices [Текст] / K. Kyle, S. A. Crossley // The Modern Language Journal. — 2018. — Т. 102, № 2. — С. 333—349. — URL: https://onlinelibrary.wiley.com/doi/10.1111/modl.12468.

124. *Biber, D.* Grammatical Complexity in Academic English. Linguistic Change in Writing [Текст] / D. Biber, B. Gray. — Cambridge : Cambridge University Press, 2016.

125. *Ljashevskaja, O. N.* On Determining the Complexity of Russian Texts [Текст] / O. N. Ljashevskaja // XVII Aprel'skaia mezhdunarodnaia nauchnaia konferentsiia po problemam razvitiia ekonomiki i obshchestva: v 4 kn. / под ред. E. G. Jasin. — HSE University Publishing House, 1996. — С. 408—419.

126. *Bentz, C.* Zipf's law of abbreviation as a language universal [Текст] / C. Bentz, R. Ferrer-i-Cancho // Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics / под ред. C. Bentz, G. Jäger, I. Yanovich. — University of Tübingen, 2016. — С. 1—4.

127. What happens to bert embeddings during fine-tuning? [Текст] / A. Merchant [и др.] // arXiv preprint arXiv:2004.14448. — 2020.

128. *Kuratov, Y.* Adaptation of deep bidirectional multilingual transformers for russian language [Текст] / Y. Kuratov, M. Arkhipov // arXiv preprint arXiv:1905.07213. — 2019.

129. Huggingface's transformers: State-of-the-art natural language processing [Текст] / T. Wolf [и др.] // arXiv preprint arXiv:1910.03771. — 2019.

130. *Loshchilov, I.* Decoupled weight decay regularization [Текст] / I. Loshchilov, F. Hutter // arXiv preprint arXiv:1711.05101. — 2017.

131. *Chen, T.* Xgboost: A scalable tree boosting system [Текст] / T. Chen, C. Guestrin // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — С. 785—794.

132. *Blinova, O.* Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony [Текст] / O. Blinova, S. Belov, M. Revazov // CEUR Workshop Proceedings. Vol-2813. Proceedings of the International Conference "Internet and Modern Society"(IMS-2020), St. Petersburg, Russia

17-20 June 2020 / под ред. R. Bolgov, A. V. Chugunov, A. E. Voiskounsky. — The name of the publisher, 2020. — С. 61—74.

133. *Assy, R.* Can the Law Speak Directly to its Subjects? The Limitation of Plain Language [Текст] / R. Assy // Journal of Law and Society. — 2013. — Т. 38, № 3. — С. 376—404. — URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6478.2011.00549.x.

134. *S., G.-R.* Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study // Łódź Studies in Language 22 [Текст] / G.-R. S. // Berlin, Peter Lang Verlag: — 2012. — С. 280.

135. *Orts, M. Á.* Power and Complexity in Legal Genres: Unveiling Insurance Policies and Arbitration Rules [Текст] / M. Á. Orts // International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique. — 2015. — Т. 28. — С. 485—505.

136. *Martínez, E.* Poor writing, not specialized concepts, drives processing difficulty in legal language // Cognition [Текст] / E. Martínez, F. Mollica, E. Gibson. — 224, 2022. — (Vol).

137. *Venturi, G.* Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach [Текст] / G. Venturi // IALF Porto. — 2012. — С. 138—156.

138. *McKinley, J.* Text analysis [Текст] / J. McKinley, R. (Heath // The Routledge Handbook of Research Methods in Applied Linguistics. — 2019. — С. 453—463.

139. *Swales, J. M.* English in Academic and Research Settings [Текст] / J. M. Swales. — Cambridge, Cambridge University Press, 1990.

140. *Bhatia, V. K.* Genre: Language use in Professional Settings. Applied linguistics and language study [Текст] / V. K. Bhatia. — London : Routledge, Taylor & Francis, 2013.

141. *Durant, A.* Legal Genres // Language and Law: A Resource Book for Students [Текст] / A. Durant, J. H. Leung // : Routledge / под ред. R. E. L. Introductions. — London : Taylor & Francis, 2016. — С. 11—15.

142. *Tessuto, G.* Investigating English Legal Genres in Academic and Professional Contexts [Текст] / G. Tessuto // Cambridge: Cambridge Scholars Publishing. — 2012. — Т. 315 p.

143. *Bhatia, V. K.* An applied discourse analysis of English legislative writing [Текст] / V. K. Bhatia // Birmingham: University of Aston in Birmingham. — 1983. — Т. 145 p.

144. *Kurzon, D.* How Lawyers Tell their Tales: Narrative Aspects of a Lawyer's Brief [Текст] / D. Kurzon // Poetics. — 1985. — Т. 14. — С. 467—481.

145. *Tiersma, P. M.* The Language of Offer and Acceptance: Speech Acts and the Question of Intent [Текст] / P. M. Tiersma // California Law Review. — 1986. — Т. 74. — С. 189—232.

146. *Trosborg, A.* An analysis of legal speech acts in English Contract Law. "It is hereby performed." // HERMES - Journal of Language and Communication in Business [Текст] / A. Trosborg // Vol. — 1991. — Т. 4. — С. 65—90.

147. *Trosborg, A.* Statutes and contracts: An analysis of legal speech acts in the English language of the law [Текст] / A. Trosborg // Journal of Pragmatics. — 1995. — Т. 23. — С. 31—53.

148. *Howe, P. M.* The problem of the problem question in English for academic legal purposes // English for Specific Purposes [Текст] / P. M. Howe // №. — 1990. — Т. 9. — С. 215—236.

149. *M., T. R. A.* Subject Specific Literacy and Genre Theory // Australian Review of Applied Linguistics [Текст] / T. R. A. M. // Legal English. — 1993. — Т. 16. — С. 86—122.

150. *Савельев, Д. А.* Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации [Текст] / Д. А. Савельев // Право. Журнал Высшей школы экономики. — 2020. — Т. Т. 1. С. — С. 50—74.

151. *Goźdź-Roszkowski, S.* Legal terms in context: phraseological variation across genres // Evidence-Based LSP: Translation, Text and Terminology, Linguistic Insights: Studies in Language and Communication [Текст] / S. Goźdź-Roszkowski // Bern: Peter Lang AG. — 2007. — С. 455—470.

152. *Dell'Orletta, F.* Genre-oriented Readability Assessment: a Case Study // Proceedings of the Workshop on Speech and Language Processing Tools in Education [Текст] / F. Dell'Orletta, G. Venturi, S. Montemagni // The COLING. — 2012. — Т. 2012 Organizing Committee, Mumbai. — С. 91—98.

153. Continent [Текст]. — 2023. — URL: https://continent-online.com/.

154. Techexpert [Текст]. — 2023. — URL: https://cntd.ru/about/network.

155. *Борисов, А. Б.* / А. Б. Борисов // Большой юридический словарь. М.: Книжный мир. — 2010. — С. 848.

156. *Додонов, В.* др. большой юридический словарь [Текст] / В. Додонов // М.: Научно-издательский центр ИНФРА-М. — 2001. — С. 780.

Список рисунков

Список таблиц