Вопросы безопасности
Правильная ссылка на статью:
Никитин П.В., Горохова Р.И., Бахтина Е.Ю., Долгов В.И., Коровин Д.И. — Алгоритмы извлечения информации из проблемно-ориентированных текстов на примере государственных контрактов // Вопросы безопасности. – 2023. – № 3. – С. 1 - 10. DOI: 10.25136/2409-7543.2023.3.43543 EDN: XNUXIB URL: https://nbpublish.com/library_read_article.php?id=43543

Алгоритмы извлечения информации из проблемно-ориентированных текстов на примере государственных контрактов

Никитин Петр Владимирович

ORCID: 0000-0001-8866-5610

кандидат педагогических наук

доцент, департамент анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации

125993, Россия, г. Москва, Ленинградский проспект, 49

pvnikitin@fa.ru

Горохова Римма Ивановна

кандидат педагогических наук

доцент, Департамент анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации

125167, Россия, г. Москва, Ленинградский проспект, 49

rigorokhova@fa.ru

Бахтина Елена Юрьевна

кандидат физико-математических наук

Доцент, Московский автомобильно-дорожный государственный технический университет (МАДИ)

125319, Россия, г. Москва, Ленинградский проспект, 46

elbakh@gmail.com

Долгов Виталий Игоревич

кандидат физико-математических наук

доцент, Департамент анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации

125319, Россия, г. Москва, ул. Ленинградский Проспект, 49

vidolgov@fa.ru

Коровин Дмитрий Игоревич

доктор экономических наук

профессор, Департамент анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации

125319, Россия, г. Москва, ул. Ленинградский Проспект, 49

dikorovin@fa.ru

Статья из рубрики "Информационное обеспечение национальной безопасности"

DOI:

10.25136/2409-7543.2023.3.43543

EDN:

XNUXIB

Дата направления статьи в редакцию:

09-07-2023

Дата публикации:

17-09-2023

Аннотация: Исследование направлено на решение проблемы исполнения государственных контрактов, важности использования неструктурированной информации и возможных методов анализа для улучшения контроля и управления этим процессом. Исполнение государственных контрактов имеет прямое влияние на безопасность страны, ее интересы, экономику и политическую стабильность. Правильное выполнение этих контрактов способствует защите национальных интересов и обеспечивает безопасность страны во всех смыслах. Объектом исследования являются алгоритмы, используемые для извлечения информации из текстов. Данные алгоритмы включают в себя технологии машинного обучения и обработку естественного языка. Они способны автоматически находить и структурировать различные сущности и данные из государственных контрактов. Научной новизной данного исследования является учет неструктурированной информации в анализе исполнения государственных контрактов. Авторы обратили внимание на проблемно-ориентированные тексты в документации контрактов и предложили анализировать их числовыми индикаторами для оценки текущего состояния контракта. Таким образом, был внесён вклад в развитие методов анализа государственных контрактов путем учета неструктурированной информации. Предложенные методы анализа проблемно-ориентированных текстов с использованием машинного обучения. Этот подход может значительно улучшить оценку и управление исполнением государственных контрактов. Результаты интерпретации проблемно-ориентированных текстов могут использоваться для оптимизации модели оценки риска исполнения государственного контракта, а также повышения ее точности и эффективности.

Ключевые слова:

государственные контракты, исполнение контракта, цифровизация, проблемно-ориентированный текст, числовые индикаторы, неструктурированная информация, машинное обучение, анализ текста, глубокое обучение, нейронные сети

Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финуниверситета

Введение

Система государственных контрактов создана для реализации потребностей бюджетных учреждений. Она позволяет ведомственным организациям для себя организовать поставку товаров, выполнение работ, оказание услуг за счет бюджетных средств и внебюджетных фондов. Современная проблема государственных заказов состоит в их исполнимости: не все подрядчики способны выполнить обязательства по контракту по разным причинам. Цифровизация способствовала прозрачности процесса сопровождения государственных контрактов: любое заинтересованное лицо может ознакомиться со всеми деталями сделки между бюджетным учреждением и будущим подрядчиком на государственном интернет-портале. Это позволяет контролировать целевое использование бюджетных средств, что не менее важно с точки зрения исполнения бюджета. В рамках исполнения федерального закона исполнители заполняют множество расчетной, сметной и описательной документации, которая также находится в свободном доступе на портале или сайтах для взаимодействия заказчиков (госорганов) и исполнителей (подрядчиков, поставщиков товаров или услуг). Такие документы характеризуют основные параметры любого государственного заказа, а текст в таких документах является проблемно-ориентированным. Проблемно-ориентированные тексты (ПОТ) - это особый тип текстовых данных, содержащих информацию о конкретной проблеме или вопросе. Кроме того, существуют и другие источники открытой и доступной неструктурированной информации. Анализ таких источников позволяет получить новые числовые индикаторы для оценки текущего состояния контракта.

Проблему исполнения государственных контрактов рассматривают многие эксперты и исследователи из разных научных областей. Чаще всего данную проблему берут во внимание именно экономисты, поскольку исследование влияния бюджетных средств на реальную экономику является актуальным. Как правило, для конструирования математико-экономических моделей используются уже известные количественные индикаторы, которые доступны на открытом портале государственных закупок. Так исследователи в своей работе [1] использовали все доступные численные показатели (со стороны заказчика, со стороны поставщика, а также параметры самого контракта), которые могут влиять на риски неисполнения контрактов (представлены на рис. 1). Но авторы не учитывали доступную неструктурированную информацию, которую также можно выразить в числовых индикаторах. Потому следует отдельно уделить внимание методам анализа ПОТ с использованием машинного обучения.

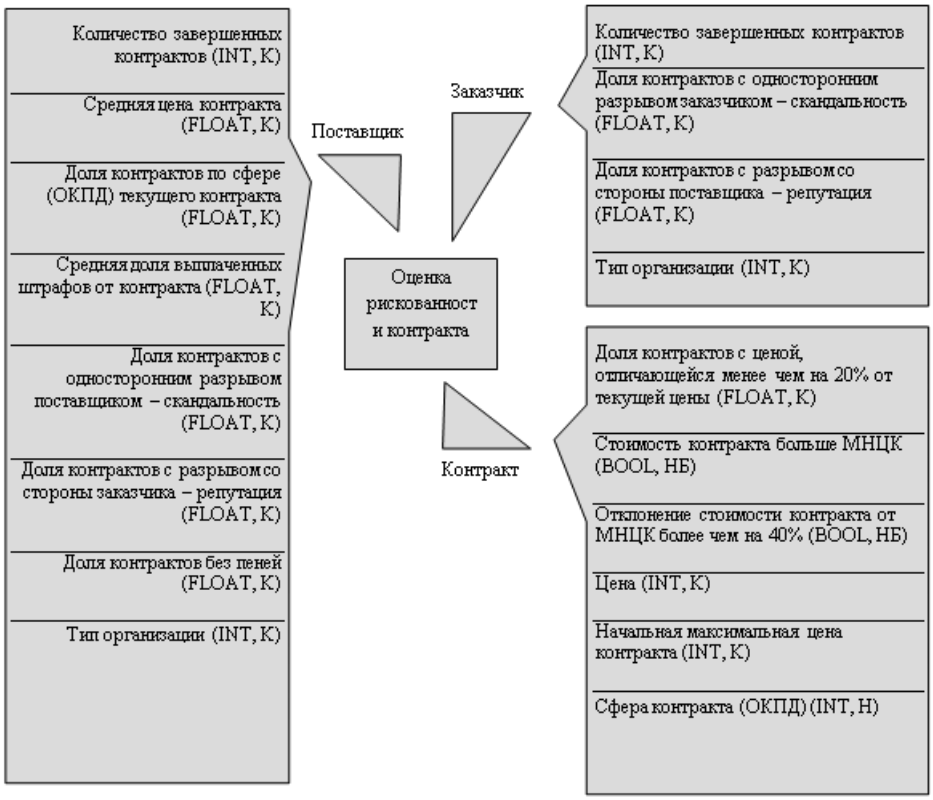

Рисунок 1. Сформированные числовые признаки в работе [1].

Существует большое количество научных работ и программных систем, посвященных использованию различных методов анализа неструктурированной информации в разнотипных сферах деятельности: в гражданской авиации, в мониторинге новостей и интернет-сайтов, медицине, в текстовом анализе документов и поиск информации в неструктурированных источниках.

Авторы статьи [2] анализируют эффективность компьютерной модели построения языка в его структуризации и формализации, что способствует конкретизации смысла устного и письменного текста в условиях работы с большими объемами неструктурированной текстовой информации. В статье описываются базовые понятия, характеризующие синтаксические связи текстовых сообщений на английском языке, а также специфика подхода к анализу синтаксических структур англоязычных текстов. На примере программного продукта, разработанного авторами, анализируется последовательность реализации его алгоритма.

В работе [3] рассматривается применение методов обработки естественного языка для анализа больших объемов клинически важной информации, содержащейся в неструктурированном виде в электронных медицинских картах. В статье представлены примеры успешного применения некоторых методов обработки естественного языка, а также перечислены основные проблемы и ограничения их более эффективного использования для анализа неструктурированных медицинских данных.

Статьи [4-6] посвящена проблеме автоматической обработки неструктурированных документов. Объем таких документов растет с каждым годом, и ручная обработка отнимает много времени и ресурсов. Для решения этой проблемы представлена общая архитектура системы извлечения информации, основанной на современных подходах обработки естественного языка и машинного обучения. Представленное решение позволяет обрабатывать большое количество неструктурированной информации без

написания кода и подготовки шаблонов синтаксического анализа.

В совокупности данные подходы анализа неструктурированного текста имеют место при исследовании возможности создания дополнительного индикатора в рамках оценки риска неисполнения государственных контрактов.

Цель исследования

Целью данного исследования ставится изучение алгоритмов глубокого машинного обучения анализа текстовой информации для формирования дополнительного индикатора оценки исполнения государственных контрактов.

Материал и методы исследования

В государственных закупках чаще всего используются стандартные типы документов, такие как договоры, счета, какие-либо сопроводительные письма и приложения. Иными словами, это юридические документы, а это еще один пример ПОТ. Они содержат юридические термины, определения и правила, используемые для описания законодательных процедур, правовых процессов, правовых аспектов сделок и т.д. Также они содержат ценную информацию о параметрах самого контракта, которая не может быть представлена численным индикатором без предварительной обработки.

В частности, у каждого контракта есть раздел проекта договора, который содержит основные параметры:

· Названия и реквизиты государственных органов и поставщиков.

· Описание работ, услуг или товаров.

· Даты и сроки выполнения.

· Суммы и условия оплаты.

· Технические характеристики и требования.

· Положения о гарантиях и ответственности.

· Санкции и штрафы.

· Реквизиты и подписи сторон контракта.

Особенно важными среди этих характеристик являются описание работ, услуг или товаров, технические характеристики и требования, положения о гарантиях и ответственности, а также санкции и штрафы. В совокупности перечисленные параметры являются целевыми для определения сложности контракта, что напрямую влияет на качество его исполнения подрядчиком. Остальные параметры являются не менее ценными, но чаще всего представлены в качестве структурированной информации на самой платформе для госзакупок.

При анализе проблемно-ориентированных текстов существует ряд сложностей: неструктурированная информация не имеет четкой структуры и может содержать большое количество шума и несущественных данных, достаточно большой объем необработанных текстов. Традиционные методы анализа данных (реляционные базы данных и статистические методы) не могут обработать и извлечь информацию из таких данных. В контексте извлечения информации из ПОТ можно рассмотреть два фундаментальных подхода, которые чаще всего используются исследователями и разработчиками при

решении данной задачи: ручное определение правил для извлечения текста и статистический метод с использованием алгоритмов машинного и глубокого обучения. Преимущество ручного составления правил заключается в контроле извлечения информации, высокой точности, но такой метод применим к небольшим наборам данных. Потому для решения проблемы анализа ПОТ эффективнее использовать методы обработки естественного языка (NLP, определение тональности текста), машинное обучение (ML, составление простых моделей кластеризации и классификации текстовых данных) и нейронные сети (NN).

В ходе реализации процессов NLP могут потребоваться как стандартные алгоритмы, так и более сложные модели обработки текстов. Среди таковых можно выделить токенизацию слов, нормализацию текста, маркировку частей речи, лемматизацию и стемминг, удаление стоп-слов, распознавание именованных сущностей. Простые методы могут стать фундаментом для реализации сложных моделей обработки текста, с помощью которых можно сформулировать числовые индикаторы и классы. Такими моделями можно назвать тональность текста, определение темы текста, выделение главного и т.д. На этапе реализации таких моделей следует рассмотреть применение моделей машинного обучения и глубокого обучения с использованием нейронных сетей. Ранее было определено, что основную часть текстовых данных для государственных контрактов составляют юридические документы, соответственно, есть смысл рассмотреть модели, которые имеют высокую эффективность при анализе ПОТ юридического характера.

Основным инструментом для извлечения сущностей государственного контракта в данном исследовании будет библиотека Natasha [7]. Библиотека Natasha является инструментом для обработки естественного языка, специально разработанным для извлечения сущностей из текстов, включая договоры по государственным контрактам.

Основной задачей Natasha является автоматическое распознавание и извлечение различных сущностей из текста, включая имена организаций, должности, даты, суммы, сроки выполнения работ и другие важные данные, которые обычно содержатся в договорах по государственным контрактам. Она работает на основе обучения моделей на больших объемах текстовых данных, что позволяет ей достичь высокой точности и полноты при извлечении информации из текста. Библиотека также предоставляет дополнительные функции, такие как поиск по категориям и классификация текстов, что делает ее полезной для разных задач обработки естественного языка связанных с государственными контрактами.

Применив данные алгоритмы и загрузив документы договоров государственных контрактов, мы можем извлекать сущности, которые в дальнейшем будем использовать как признаки в обучении моделей машинного обучения. В нашем случае существенными признаками, которых нет в системе ИЕС ГОЗ являются: положения о гарантиях и ответственности; санкции и штрафы.

Применение описанных алгоритмов позволит дополнить основные модели расчета риска исполнения государственных контрактов в целях повышения точности их прогноза и дальнейшего использования в промышленных масштабах.

Результаты исследования и их обсуждение

В рамках текущего исследования были предложены к рассмотрению основные методы извлечения полезной информации из неструктурированных текстов. В частности, было выяснено, что статистические методы работы с текстом представляются наиболее эффективными для их дальнейшего применения при создании дополнительного

индикатора. Также следует учесть, что по отдельности некоторые методы неприменимы: например, методы NLP позволяют сделать предобработку сырых текстов, а использование ML-моделей на таких данных позволит их классифицировать и определить числовые параметры для формирования дополнительных индикаторов. Таким образом можно построить формальный алгоритм работы с ПОТ для получения из него числовых характеристических индикаторов и финального индикатора сложности контракта. Пример такого алгоритма представлен на рис. 2.

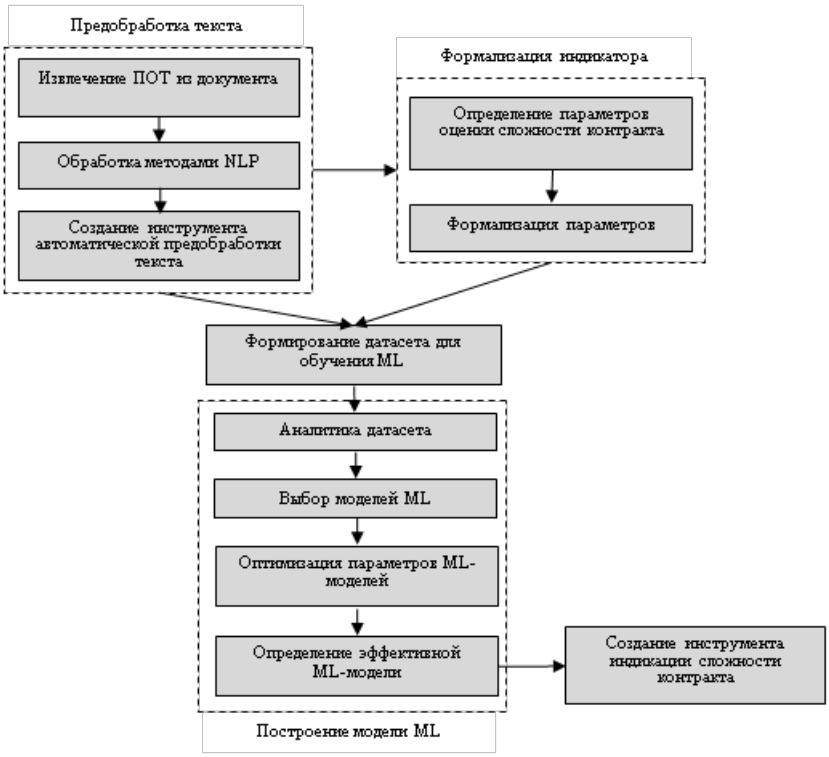

Рисунок 2. Схема алгоритма работы с ПОТ для получения дополнительных индикаторов.

Таким образом, использовав данный алгоритм мы добавляем в наш итоговой датасет новые признаки, которые позволят более детально отслеживать исполнение государственных контрактов в моделях машинного обучения. В частности, выделение таких сущностей как положения о гарантиях и ответственности, санкции и штрафы позволило повысить оценку качества обучения моделей в рамках выполнения государственных контрактов по сравнению с первыми исследованиями [8].

Заключение

В рамках исследования удалось изучить основные методы работы с проблемно-ориентированными текстами. Исследованы методы обработки естественного языка, машинного обучения и глубокого обучения в целях извлечения числовых индикаторов из текстовой информации из государственных контрактов. В дальнейшем текущее исследование будет взято за основу для реализации дополнительного инструмента индикации. Результаты интерпретации проблемно-ориентированных текстов могут использоваться для оптимизации модели оценки риска исполнения государственного контракта, а также повышения ее точности и эффективности. Для успешного выполнения государственных контрактов необходимо иметь точные и своевременные прогнозы исполнения контрактов, что позволит сократить риски и улучшить качество работы этого процесса.

Библиография

1. Елисеев Д. А., Романов Д. А. Машинное обучение: прогнозирование рисков госзакупок // Открытые системы. СУБД. 2018. № 2. С. 42-44.
2. Узких Г. Ю. Применение глубокого обучения в задачах обработки естественного языка // Вестник науки. 2023. Т. 4. №. 8 (65). С. 310-312.
3. Сердюк Ю. П., Власова Н. А., Момот С. Р. Система извлечения упоминаний симптомов из текстов на естественном языке с помощью нейронных сетей // Программные системы: теория и приложения. 2023. Т. 14. №. 1 (56). С. 95-123.
4. Курейчик В. В., Родзин С. И., Бова В. В. Методы глубокого обучения для обработки текстов на естественном языке // Известия Южного федерального университета. Технические науки. 2022. № 2 (226). С. 189-199.
5. Ежков А. А. Анализ исследований в области обработки неструктурированных текстов в медицине // Научное обозрение: актуальные вопросы теории и практики. 2022. С. 23-26.
6. Прошина М. В. Современные методы обработки естественного языка: нейронные сети // Экономика строительства. 2022. №. 5. С. 27-42.
7. Тарабрин М. А. Использование инструментов natasha api при разработке алгоритма по обезличиванию текстовых данных // Актуальные вопросы эксплуатации систем охраны и защищенных телекоммуникационных систем. 2022. С. 61-64.
8. Petr Nikitin at all. Evaluation of the execution of government contracts in the field of energy by means of artificial intelligence // E3S Web of Conf. 402 03041 (2023). DOI: 10.1051/e3sconf/202340203041

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Предметом исследования рецензируемой статьи являются алгоритмы извлечения информации из проблемно-ориентированных текстов. Выбранная магистраль изучения достаточно актуальна, нова. Базисом иллюстраций, материалом практического толка становятся государственные контракты. Как отмечает автор работы, «современная проблема государственных заказов состоит в их исполнимости: не все подрядчики способны выполнить обязательства по контракту по разным причинам. Цифровизация способствовала прозрачности процесса сопровождения государственных контрактов: любое заинтересованное лицо может ознакомиться со всеми деталями сделки между бюджетным учреждением и будущим подрядчиком на государственном интернет-портале». В целом концепция работы верифицирована, предметно-объектный пласт выверен. Стиль ориентирован на собственно научный тип, примечательна для исследования терминологическая точность, универсальность. Например, это проявляется в следующих фрагментах: «в рамках исполнения федерального закона исполнители заполняют множество расчетной, сметной и описательной документации, которая также находится в свободном доступе на портале или сайтах для взаимодействия заказчиков (госорганов) и исполнителей (подрядчиков, поставщиков товаров или услуг). Такие документы характеризуют основные параметры любого государственного заказа, а текст в таких документах является проблемно-ориентированным. Проблемно-ориентированные тексты (ПОТ) - это особый тип текстовых данных, содержащих информацию о конкретной проблеме или вопросе. Кроме того, существуют и другие источники открытой и доступной неструктурированной информации. Анализ таких источников позволяет получить новые числовые индикаторы для оценки

текущего состояния контракта», или «существует большое количество научных работ и программных систем, посвященных использованию различных методов анализа неструктурированной информации в разнотипных сферах деятельности: в гражданской авиации, в мониторинге новостей и интернет-сайтов, медицине, в текстовом анализе документов и поиск информации в неструктурированных источниках», или «при анализе проблемно-ориентированных текстов существует ряд сложностей: неструктурированная информация не имеет четкой структуры и может содержать большое количество шума и несущественных данных, достаточно большой объем необработанных текстов. Традиционные методы анализа данных (реляционные базы данных и статистические методы) не могут обработать и извлечь информацию из таких данных. В контексте извлечения информации из ПОТ можно рассмотреть два фундаментальных подхода, которые чаще всего используются исследователями и разработчиками при решении данной задачи: ручное определение правил для извлечения текста и статистический метод с использованием алгоритмов машинного и глубокого обучения» и т.д. Целевая составляющая статьи строго промаркирована, методологический уровень актуален. Отмечу, что тексту присущ синкретический характер, теоретический срез удачно совмещается с собственно практическим. Наличного текстового объема достаточно для раскрытия темы, особо расширять работу, на мой взгляд, не следует. Считаю, что материал может быть полезен как подготовленным, так и не подготовленным читателям. Комментарии по ходу работы полновесны: например, «основной задачей Natasha является автоматическое распознавание и извлечение различных сущностей из текста, включая имена организаций, должности, даты, суммы, сроки выполнения работ и другие важные данные, которые обычно содержатся в договорах по государственным контрактам. Она работает на основе обучения моделей на больших объемах текстовых данных, что позволяет ей достичь высокой точности и полноты при извлечении информации из текста. Библиотека также предоставляет дополнительные функции, такие как поиск по категориям и классификация текстов, что делает ее полезной для разных задач обработки естественного языка связанных с государственными контрактами», или «статистические методы работы с текстом представляются наиболее эффективными для их дальнейшего применения при создании дополнительного индикатора. Также следует учесть, что по отдельности некоторые методы неприменимы: например, методы NLP позволяют сделать предобработку сырых текстов, а использование ML-моделей на таких данных позволит их классифицировать и определить числовые параметры для формирования дополнительных индикаторов. Таким образом можно построить формальный алгоритм работы с ПОТ для получения из него числовых характеристических индикаторов и финального индикатора сложности контракта» и т.д. Визуальный режим систематизации наработанных данных вполне удачно коррелируется с текстовой частью; схемы, рисунки удобны для восприятия / рецепции. Выводы по тексту несколько формальны, однако они соотносятся с основной частью. Удачно пописана перспектива использования данных, это показатель научной ответственности: «в дальнейшем текущее исследование будет взято за основу для реализации дополнительного инструмента индикации. Результаты интерпретации проблемно-ориентированных текстов могут использоваться для оптимизации модели оценки риска исполнения государственного контракта, а также повышения ее точности и эффективности. Для успешного выполнения государственных контрактов необходимо иметь точные и своевременные прогнозы исполнения контрактов, что позволит сократить риски и улучшить качество работы этого процесса». Формальные требования издания учтены, список библиографических источников содержит работы разных лет, разных типов. Рекомендую статью «Алгоритмы извлечения информации из проблемно-ориентированных текстов на примере государственных контрактов» к открытой

публикации в научном журнале «Вопросы безопасности».