Litera
Правильная ссылка на статью:
Голиков А.А., Акимов Д.А., Данилова Ю.Ю. Оптимизация традиционных методов определения сходства наименований проектов и закупок с использованием больших языковых моделей // Litera. 2024. № 4. DOI: 10.25136/2409-8698.2024.4.70455 EDN: FRZANS URL: https://nbpublish.com/library_read_article.php?id=70455

Оптимизация традиционных методов определения сходства наименований проектов и закупок с использованием больших языковых моделей

Голиков Алексей Александрович

аспирант, Отделение филологии и литературы, Кафедра русского языка и литературы, Казанский (Приволжский) федеральный университет (Елабужский институт)

109316, Россия, г. Москва, ул. Волгоградский Пр., 42

ag@mastercr.ru

Акимов Дмитрий Андреевич

ORCID: 0009-0004-2800-4430

кандидат технических наук

Аналитик, ООО "Мастерская цифровых решений"

109316, Россия, г. Москва, Волгоградский пр., 42

akimovdmitry1@mail.ru

Данилова Юлия Юрьевна

ORCID: 0000-0001-5736-0590

кандидат филологических наук

доцент, кафедра русского языка и литературы, Елабужский институт (филиал) Казанского федерального университета

423604, Россия, республика Татарстан, г. Елабуга, ул. Казанская, 89

danilovaespu@mail.ru

Статья из рубрики "Автоматическая обработка языка"

DOI:

10.25136/2409-8698.2024.4.70455

EDN:

FRZANS

Дата направления статьи в редакцию:

09-04-2024

Дата публикации:

16-04-2024

Аннотация: Предметом исследования является анализ и совершенствование методов определения релевантности наименований проектов к информационному содержанию закупок с использованием больших языковых моделей. Объектом исследования служит база данных, содержащая наименования проектов и закупок в сфере электроэнергетики, собранная из открытых источников. Автор подробно рассматривает такие аспекты темы, как применение метрик TF-IDF и косинусного сходства для первичной фильтрации данных, а также детально описывает интеграцию и оценку эффективности больших языковых моделей, таких как GigaChat, GPT-3.5, и GPT-4 в задачах сопоставления текстовых данных. Особое внимание уделяется методикам уточнения сходства наименований на основе рефлексии, введенной в промпты больших языковых моделей, что позволяет повысить точность сопоставления данных. В исследовании использованы методы TF-IDF и косинусного сходства для первичного анализа данных, а также большие языковые модели GigaChat, GPT-3.5 и GPT-4 для детальной проверки релевантности наименований проектов и закупок, включая рефлексию в промптах моделей для улучшения точности результатов. Новизна исследования заключается в разработке комбинированного подхода к определению релевантности наименований проектов и закупок, сочетающего традиционные методы обработки текстовой информации (TF-IDF, косинусное сходство) с возможностями больших языковых моделей. Особым вкладом автора в исследование темы является предложенная методика повышения точности сопоставления данных за счет уточнения результатов первичного отбора с помощью моделей GPT-3.5 и GPT-4 с оптимизированными промптами, включающими рефлексию. Основными выводами проведенного исследования являются подтверждение перспективности применения разработанного подхода в задачах информационной поддержки процессов закупок и реализации проектов, а также возможность использования полученных результатов для развития систем интеллектуального анализа текстовых данных в различных отраслях экономики. Исследование показало, что использование языковых моделей позволяет улучшить значение F2-меры до 0,65, что свидетельствует о значительном повышении качества сопоставления данных по сравнению с базовыми методами.

Ключевые слова:

TF-IDF, косинусное сходство, большие языковые модели, GigaChat, GPT-4, анализ текстовых данных, рефлексия в промптах, определение релевантности, проекты и закупки, оптимизация бизнес-процессов

1. Введение

Решение задачи определения сходства документов или предложений актуально для большого количества бизнес-кейсов: в рекомендательных системах, в поисковых системах, чат-ботах, при проверках научных работ на плагиат и др. [1, 2]. В компании авторов статьи (ООО «Мастерская цифровых решений», г. Москва) данная задача решается для сопоставления проектов и закупок в электроэнергетике в России по их

наименованию. Например, закупка «Выполнение строительно-монтажных работ по модернизации систем компенсации емкостных токов на ПС 110 кВ Бурдун для нужд филиала АО «Россети Тюмень» Тюменские электрические сети», очевидно, связана с проектом «Модернизация систем компенсации емкостных токов на ПС 110 кВ Бурдун (4 шт.)» и не связана с проектом «Оборудование комплексом по обеспечению информационной безопасности энергообъектов ПАО «МОЭСК» (5 этап), в т.ч. ПИР (26 шт. (прочие))». Решение данной задачи позволяет компаниям, желающим принять участие в поставке своих материалов, комплектующих и изделий для того или иного проекта, отыскать нужные закупки и принять участие в соответствующих тендерах, а также косвенно по наличию или отсутствию закупок отслеживать ход проекта. Для заказчиков данный сервис помогает привлечь большее количество поставщиков и тем самым повысить конкуренцию и улучшить условия закупки. Кроме того, решение задачи соответствует современным трендам по интеллектуальной обработке больших объемов данных, продолжающейся цифровой трансформации предприятий и их бизнес-процессов.

Вышеупомянутая задача на первый взгляд представляется обманчиво простой, поскольку требуется эффективно попарно сопоставить между собой не большие объемы текста со сложной смысловой нагрузкой, а наименования в виде двух предложений. Однако на самом деле данная задача является довольно нетривиальной, поскольку количество проектов и закупок в электроэнергетике в России измеряется ежегодно десятками и сотнями тысяч, присутствуют специализированные термины, номинации, аббревиатуры, не представленные широко за пределами отрасли. Кроме того, зачастую в наименовании проекта или закупки не содержится полная информация, позволяющая выполнить однозначное сопоставление. Например, не представляется возможным точно определить, соотносятся ли между собой проект «Создание интеллектуальной системы учета электрической энергии в рамках исполнения Федерального закона от 27.12.2018 № 522-ФЗ в филиале ПАО «Россети Кубань» Усть-Лабинские электрические сети (5129 т.у.)» и закупка «Право заключения договора на поставку приборов учета электрической энергии (мощности) в рамках исполнения требований Федерального закона от 27.12.2018 № 522-ФЗ для обеспечения потребности ПАО «Россети Кубань» в 2023 году», поскольку невозможно, исходя из названий, однозначно ответить на вопрос, пойдут ли закупаемые для ПАО «Россети Кубань» приборы учета электрической энергии для его филиала Усть-Лабинские электрические сети, указанные в наименовании проекта.

Таким образом, решаемая задача является нечеткой, ее ответ не всегда возможно определить достоверно даже с участием эксперта, однако объем сопоставляемых данных при этом диктует необходимость создания алгоритмического решения. Вместе с тем данная задача является актуальной, обладающей большим практическим значением для заказчиков и поставщиков товаров и услуг не только в электроэнергетике, но и в иных отраслях экономики, а методологические подходы, использованные при ее успешном решении, могут быть использованы и для решения аналогичных задач (например, для создания рекомендательных систем, предлагающих схожие закупки с теми, которыми интересовался пользователь и др.). Объектом исследования является база данных, содержащая наименования проектов и закупок (для сбора эмпирического материала были использованы открытые источники, содержащие инвестиционные программы и закупки электросетевых компаний), предметом исследования – определение релевантности наименования проекта информационному содержанию закупки. Научная значимость исследования заключается в применении оперативных возможностей больших языковых моделей в качестве дополнения к классическим методам обработки текстовой информации. Целью исследования является повышение точности сопоставления проекта и закупки, а задачи включают анализ текущих методов

определения сходства двух предложений и оценку эффективности новых подходов к этой же задаче.

2. Реализация алгоритма определения сходства наименований проектов и закупок

Для решения задачи определения сходства двух предложений могут быть использованы различные методы:

1) методы, использующие метрики сходства последовательности символов (сходство Жаккара, расстояние Левенштейна, сходство Джаро-Винклера и др.) [3, 4];

2) методы, использующие преобразование текста в векторное представление (Word2vec, TF-IDF), что позволяет на следующем шаге выполнить расчет косинусного сходства [5, 6];

3) методы, также использующие преобразование текста в векторное представление, но задействующие при этом глубокие нейронные сети с большим числом параметров (BERT, Gigachat, ChatGPT и др.) [7, 8].

Поскольку, как было упомянуто выше, ежегодное количество проектов и закупок в электроэнергетике России измеряется десятками, а то и сотнями тысяч в год, то количество попарных сопоставлений наименований каждого проекта и закупки может достигать 10 миллиардов, что естественным образом подталкивает к применению не слишком вычислительно затратных решений или же, по крайней мере, к использованию для первоначального отбора потенциальных пар «проект – закупка» достаточно простых и быстрых методов.

К простым методам можно отнести методы, использующие метрики сходства последовательности символов, или же преобразование слов в векторы с использованием TF-IDF с последующим расчетом косинусного сходства. Очевидно, что в данном случае больше подходит второй вариант – вычисление TF-IDF и косинусного сходства, поскольку схожесть проекта и закупки в первую очередь проявляется в достаточно редких для рассматриваемого корпуса текста словах (адресах, наименованиях подстанций), а не в сходстве последовательностей символов полных наименований проекта и закупки.

Напомним, что TF-IDF (аббревиатура от англ. term frequency inverse document frequency, что буквально «частота термина, обратная частота документа») рассчитывается как произведение двух сомножителей – TF и IDF.

$$TF(t,d) = \frac{n_t}{\sum_k n_k},$$

где n_t – количество раз, когда термин t встречается в документе d; $\sum_k n_k$ – общее количество терминов в документе d. То есть TF – мера частоты вербального воспроизведения термина (или другой конкретной номинации) в документе.

$$IDF(t,D) = log\frac{|D|}{|\{d_i \in D \mid t \in d_i\}|},$$

где $|D|$ – общее количество документов, $|\{d_i \in D \mid t \in d_i\}|$ – количество документов, содержащих термин t. Так, IDF отражает обратную частоту встречаемости слова в документах: чем реже слово, тем больше IDF.

Соответственно TF-IDF является произведением вышеупомянутых показателей TF и IDF и принимает высокое значение для редких терминов (например, наименования подстанций, географические названия) и низкое для частотных (служебные части речи,

ключевые слова типа «подстанция», «линия» и т.д.):

$$TF\text{–}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

В данном случае, поскольку в качестве «документа» из классического определения TF-IDF выступает достаточно короткое наименование проекта или закупки, является допустимым и использование только IDF, т.к. в коротких наименованиях внутри самого наименования отдельный термин зачастую встречается один раз, однако использование классического показателя TF-IDF дает, как показали результаты анализа, аналогичный результат.

За счет вычисления TF-IDF для каждого термина в тексте возможен перевод наименований проектов и закупок в векторное представление. Например, если есть проект «установка приборов учета» и «закупка приборов учета», то возможным векторным представлением для наименования проекта является четырехмерный вектор {1, 0, 0.5, 0.5}, а для наименования проекта {0, 1, 0.5, 0.5}, т.е. общий корпус слов состоит из четырех терминов, из которых два слова встречаются один раз (более высокое значение TF-IDF) и два слова встречаются два раза каждое (более низкое значение TF-IDF). Для корпуса, содержащего тысячи слов, векторы, соответствующие отдельным наименованиям закупок и проектов, будут разреженными, т.е. почти для всех координат будут нули кроме координат, соответствующих словам, содержащимся в наименовании.

Далее после определения векторов, соответствующих наименованиям проектов и закупок, возможно провести попарный расчет косинусного сходства между каждым проектом и каждой закупкой. Косинусное сходство между двумя векторами A и B может быть рассчитано как [9]:

$$\cos(\boldsymbol{A},\boldsymbol{B}) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}},$$

где Ai, Bi – i-е элементы векторов A и B размерностью n каждый.

Таким образом, возможно получить все попарные значения сопоставлений проектов и закупок, выраженные численно: по результатам анализа было определено, что соответствующие друг другу проекты и закупки встречаются, начиная с косинусного сходства 0,44 и выше (максимальное значение 1 соответствует полностью совпадающим словам в наименовании проекта и закупки). Однако не все проекты и закупки со значением косинусного сходства больше 0,44 являются соответствующими друг другу. Можно лишь отметить общую тенденцию: чем больше косинусное сходство, тем больше вероятность того, что проект и закупка соответствуют друг другу. Поэтому отбор пар «проект – закупка» с помощью расчета косинусного сходства векторов, полученных на базе TF-IDF, можно считать первоначальным. На наш взгляд, его можно оптимизировать с помощью применения больших языковых моделей.

Было реализовано обращение по API к большим языковым моделям Gigachat, GPT‌3.5, GPT-4 со следующим промптом: «Твоя задача: точно определить, есть ли связь между проектом и закупкой. Связь существует, только если название объекта (с прописной буквы) или его номер, или географический адрес в названии проекта и закупки полностью идентичны. "Проект: 'наименование проекта'. Закупка: 'наименование закупки'. Есть ли между ними точная связь?"».

Впоследствии по результатам анализа качество работы больших языковых моделей было

улучшено за счет такого известного приема, как введение рефлексии [10], а именно в промпт было добавлено: «В случае ответа 'да', будь готов точно указать, в чем состоит связь. В случае ответа 'нет', будь готов пояснить, почему нет».

Для возможности автоматической обработки ответа модели формат ответа должен быть зафиксирован, поскольку иначе большая языковая модель может отвечать длинным предложением, непригодным для последующей автоматической обработки. Для этого в промпт было добавлено: «Ответь 'да' или 'нет' в формате JSON. Например: {'answer': 'да'}». Таким образом, финальный пайплайн обработки данных по проектам и закупкам выглядит следующим образом (рисунок 1).

Рисунок 1: Пайплайн алгоритма определения соответствующих друг другу проектов и закупок

Приведем пример, как работает данный алгоритм. Для наименования каждого проекта и каждой закупки выполняется расчет TF-IDF, на основе чего выполняется попарный расчет косинусного сходства каждого проекта и каждой закупки, что позволяет выполнить первичный отбор потенциальных пар «проект – закупка» с косинусным сходством более 0,44. Далее для всех отобранных пар выполняется запрос по API к большой языковой модели. Возьмем в качестве примера пару, состоящую из проекта «Модернизация ПС 110/35/10 кВ Пречистое с монтажом оборудования систем видеонаблюдения» и закупки «Право на заключение Договора на выполнение строительно-монтажных работ по объектам: «Модернизация ПС 110/35/10 кВ Пречистое, ПС 110/35/10 кВ Ершичи, ПС 110/35/10 кВ Каспля с монтажом оборудования систем видеонаблюдения» для нужд ПАО «Россети Центр»». Итоговый текст запроса к большой языковой модели следующий: «Твоя задача: точно определить, есть ли связь между проектом и закупкой. Связь существует, только если название объекта (с прописной буквы) или его номер, или географический адрес в названии проекта и закупки полностью идентичны. Проект: «Модернизация ПС 110/35/10 кВ Пречистое с монтажом оборудования систем видеонаблюдения». Закупка: «Право на заключение Договора на выполнение строительно-монтажных работ по объектам: "Модернизация ПС 110/35/10 кВ Пречистое, ПС 110/35/10 кВ Ершичи, ПС 110/35/10 кВ Каспля с монтажом оборудования систем видеонаблюдения" для нужд ПАО «Россети Центр»». Есть ли между ними точная связь?". В случае ответа 'да', будь готов точно указать, в чем состоит связь. В случае ответа 'нет', будь готов пояснить, почему нет. Ответь 'да' или 'нет' в формате JSON. Например: {'answer': 'да'}».

Ответ модели GPT-4 в данном случае: «{'answer': 'да'}». Таким образом, из тех первично отобранных пар «проект – закупка», для которых большая языковая модель ответила «{'answer': 'да'}», формируется окончательный перечень пар «проект – закупка».

3. Результаты и обсуждение

Для оценки результатов следует выбрать подходящую метрику для данной задачи

определения релевантных пар «проект – закупка». Очевидно, идеальным результатом была бы классификация всех релевантных пар «проект – закупка» как релевантных, а остальных как нерелевантных. Однако на практике достижение подобного идеального результата в этой задаче маловероятно, поскольку возможны 2 вида ошибок [12, 13]:

Ошибка первого рода – ложноположительный результат, False Positive, FP: алгоритм определяет пару «проект – закупка» как релевантную, в то время как на самом деле пара нерелевантна.

Ошибка второго рода – ложноотрицательный результат, False Negative, FN: алгоритм определяет пару «проект – закупка» как нерелевантную, в то время как на самом деле пара релевантна.

С точки зрения бизнес-логики, в данной задаче ложноположительный результат (вывести нерелевантную закупку для проекта) является менее грубой ошибкой, чем ложноотрицательный результат (не вывести релевантную закупку для проекта). В подобных случаях в качестве метрики зачастую выбирается полнота (recall) [14, 15]:

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)},$$

где TP – количество правильно идентифицированных релевантных пар, FN – количество релевантных пар, которые были отмечены алгоритмом как нерелевантные.

Однако при использовании подобной метрики отдается слишком большой приоритет недопущению ложноотрицательных результатов – количество ложноположительных результатов никак не влияет на итоговый результат, и в случае, если алгоритм отметит все пары «проект – закупка» как релевантные, то метрика полноты будет максимальной.

Иной часто используемой метрикой является точность (precision):

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)},$$

где TP – то же, что и в формуле полноты, FP – количество нерелевантных пар, которые были отмечены алгоритмом как релевантные. Данная метрика подходит нам еще меньше, т.к. количество ложноотрицательных результатов, которых мы стремимся избежать больше, чем ложноположительных, вообще не оценивается.

Поэтому для нас оптимальной метрикой будет метрика, сочетающая в себе метрики полноты и точности, но со смещением важности в пользу полноты (в приоритете – недопущение ложноотрицательных результатов). В этом случае зачастую используют так называемую F2-меру [Быстров и др., 2022, с. 140-141]:

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall},$$

где $\beta = 2$.

Из результатов первичного отбора потенциальных пар «проект – закупка» путем вычисления TF-IDF и попарного расчета косинусного сходства была сформирована случайная выборка из 100 пар «проект – закупка» с косинусных сходством не менее 0,44 – значение, начиная с которого встречаются релевантные пары. Среди этих 100 пар лишь 18% оказались на самом деле релевантными, что можно считать относительно приемлемым результатом для быстрого вычислительно не слишком затратного метода с учетом большого размера изначальных баз данных проектов и закупок (десятки тысяч

записей в каждой).

Применение на втором этапе больших языковых моделей по схеме, описанной в разделе 2, позволяет достичь следующих значений F2-меры в задаче классификации релевантных и нерелевантных пар «проект – закупка» (рисунок 2).

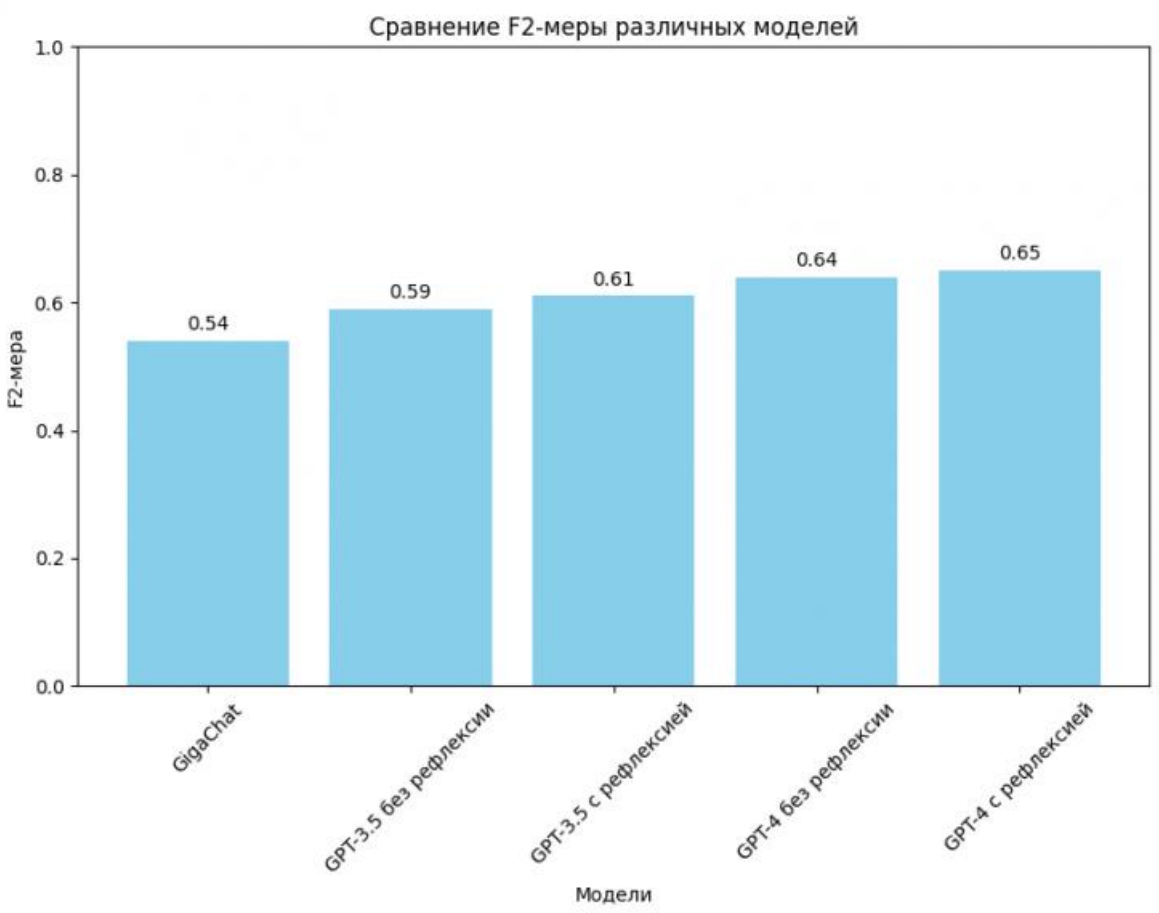

Рисунок 2: Значения F2-меры в задаче классификации пар «проект – закупка» при использовании различных больших языковых моделей и промптов

Наихудший результат продемонстрировала модель GigaChat: значение F2-меры составило всего 0,54, при этом модель практически не позволяет улучшить первичный отбор, поскольку, по мнению модели GigaChat, 96 пар из 100 являются релевантными.

Наилучшего результата (табл. 1) достигла модель GPT-4 при использовании блока рефлексии в промпте, описанного в разделе 2.

	Предсказано релевантно	**Предсказано нерелевантно**
Действительно релевантно	12	6
Действительно нерелевантно	9	73

Таблица 1: Матрица ошибок (confusion matrix) модели GPT-4 с рефлексией

Приведенные в таблице 1 результаты модели GPT-4 можно описать следующим образом:

- True Positives (TP): 12 – действительно релевантные пары, правильно классифицированные как релевантные (например: Проект «Техперевооружение КТП

10/0.4 кВ №476 ф.01 ПС 35/10 кВ Рышково с заменой силового трансформатора 100 на 100 кВА п.Рышково Железногорский р-он (трансформаторная мощность 0.1 МВА)» и закупка «Выполнение СМР по объекту: Техперевооружение КТП 10/0,4 кВ №476 ф.01 ПС 35/10 кВ Рышково с заменой силового трансформатора 100 на 100 кВА п.Рышково Железногорский р-он»);

- True Negatives (NP): 73 – действительно нерелевантные пары, правильно классифицированные как нерелевантные (например: Проект «Приобретение в ЦАР колонковых элегазовых выключателей 110кВ – 5 шт.» и закупка «Приобретение в ЦАР измерительных трансформаторов напряжения 110кВ – 9 шт.»);

- False Positives (FP): 9 – нерелевантные пары, ошибочно классифицированные как релевантные (например: Проект «Реконструкция ПС 500 кВ Усть-Кут с установкой ШР 500 кВ мощностью 180 Мвар для ВЛ 500 кВ Усть-Илимская ГЭС-Усть-Кут № 3» и закупка «Право заключения договора на выполнение работ по РД, СМР, ПНР, поставка оборудования по строительству РЭБ по титулу «Строительство ВЛ 220 кВ Усть-Кут – Ковыкта № 1 и № 2 ориентировочной протяженностью 256 км каждая»);

- False Negatives (FN): 6 – релевантные пары, ошибочно классифицированные как нерелевантные (например: Проект «Модернизация оборудования энергообъектов и РДП 12 РЭР МКС-филиала ПАО "МОЭСК, в т.ч. ПИР: г. Москва (6 шт.(прочие))» и закупка «Выполнение СМР, ПНР, материалы, оборудование по титулу: Модернизация оборудования энергообъектов и РДП 22 РЭР МКС-филиала ПАО "МОЭСК, в т.ч. ПИР: г Москва (4 шт.(прочие))»).

Таким образом, 85 пар «проект – закупка» из 100 классифицированы моделью GPT-4 верно. Например:

- Проект «Техперевооружение КТП 10/0.4 кВ №476 ф.01 ПС 35/10 кВ Рышково с заменой силового трансформатора 100 на 100 кВА п.Рышково Железногорский р-он (трансформаторная мощность 0.1 МВА)» изакупка «Выполнение СМР по объекту: Техперевооружение КТП 10/0,4 кВ №476 ф.01 ПС 35/10 кВ Рышково с заменой силового трансформатора 100 на 100 кВА п.Рышково Железногорский р-он»;

- Проект «Модернизация систем компенсации емкостных токов на ПС 110 кВ Бурдун (4 шт.)» и закупка «Выполнение строительно-монтажных работ по модернизации систем компенсации емкостных токов на ПС 110 кВ Бурдун для нужд филиала АО Росети Тюмень Тюменские электрические сети».

Такой результат можно расценивать как удовлетворительный, поскольку позволяет значительно точнее отражать для пользователей релевантные закупки. Однако следует отметить относительно высокое значение ложноотрицательных результатов – 6 пар (типа: Проект «Модернизация оборудования энергообъектов и РДП 12 РЭР МКС-филиала ПАО "МОЭСК, в т.ч. ПИР: г. Москва (6 шт.(прочие))» и закупка «Выполнение СМР, ПНР материалы, оборудование по титулу: Модернизация оборудования энергообъектов и РДП 22 РЭР МКС-филиала ПАО "МОЭСК, в т.ч. ПИР: г. Москва (4 шт.(прочие))»), т.е. трет релевантных пар неверно отмечена, как нерелевантные, что, очевидно, оставляет значительный потенциал для дальнейшего улучшения качества дифференциации пар «проект – закупка» на (не-)релевантные. Также стоит отметить относительную дороговизну обращения по API к модели GPT-4 (обработка 100 пар «проект – закупка» стоит около 0,62$), хотя и с развитием моделей имеет место тенденция к удешевлению их использования.

Таким образом, применение больших языковых моделей является перспективным инструментом, позволяющим значительно улучшать результаты более быстрых, но менее точных традиционных подходов по обработке текстовых данных. Наиболее адекватной моделью для рассматриваемой задачи оказалась модель GPT-4 с добавленным блоком рефлексии, верно классифицировавшая 85 пар «проект – закупка» из 100. Вместе с тем относительное большое количество ложноотрицательных результатов и достигнутое значение F2-меры (0,65) указывают на то, что остается широкое пространство для более четкой и точной дифференциации пар «проект – закупка» на (не-)релевантные.

4. Заключение

В работе была исследована актуальная задача определения по наименованию релевантных закупок для бизнес-процесса, в частности, для инвестиционных проектов в сфере электроэнергетики. Ее решение благоприятно отразится на рыночных условиях отрасли, позволит поставщикам материалов и оборудования эффективнее находить подходящие тендеры для участия, а заказчикам привлекать больше поставщиков. Кроме того, эффективные методы решения задачи могут быть применены и в иных отраслях экономики.

Были рассмотрены возможные подходы к определению сходства двух предложений, обосновано применение на первом этапе вычислительно быстрого метода, заключающегося в переводе наименований проектов и закупок в векторное представление с помощью расчета метрики TF-IDF каждого наименования и последующего попарного расчета косинусного сходства всех проектов и закупок с отсечением всех пар «проект – закупка» с косинусным сходством выше 0,44 (ниже 0,44 релевантные пары, как показал выборочный анализ, не встречаются). Однако из 100 отобранных таким образом пар «проект – закупка» лишь 18 в среднем являются релевантными, что диктует необходимость дальнейшей дифференциации пар «проект – закупка» на релевантные и нерелевантные.

На втором этапе были протестированы различные большие языковые модели с одним из возможных промптов и добавлением в него блока рефлексии модели. Было выявлено, что модель GigaChat практически не приводит к улучшению первичного отбора, в то время как модель GPT-4 с блоком рефлексии показывает наилучшие результаты (значение F2-меры 0,65). Было продемонстрировано, что рефлексия несколько улучшает результат работы больших языковых моделей (значение F2-меры улучшается на 0,01-0,02). Контроль формата ответа модели был осуществлен путем добавления в промпт указания отвечать в формате JSON.

Авторы планируют продолжить исследование с целью дальнейшей оптимизации алгоритма сопоставления проектов и закупок как путем уточнения промптов для обращения к большим языковым моделям (существуют и иные варианты улучшения промптов помимо введения блока рефлексии), так и использованием дополнительных алгоритмов классификации на вручную размеченных данных: например, модель градиентного бустинга с такими признаками, как длины наименований проектов и закупок, количество прописных букв в наименованиях проектов и закупок и др.

Библиография

1. Оськина К. А. Оптимизация метода классификации текстов, основанного на tf-idf, за счет введения дополнительных коэффициентов //Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2016. – №. 15 (754). – С. 175-187.

2. Murugesan M. et al. Efficient privacy-preserving similar document detection // The VLDB Journal. – 2010. – Vol. 19. – №. 4. – Pp. 457-475.

3. Знаменский С. В. Модель и аксиомы метрик сходства //Программные системы: теория и приложения. – 2017. – Т. 8. – №. 4 (35). – С. 347-357.

4. Гайдамакин Н. А. Мера сходства последовательностей одинаковой размерности // Математические структуры и моделирование. – 2016. – №. 4 (40). – С. 5–16.

5. Лыченко Н. М., Сороковая А. В. Сравнение эффективности методов векторного представления слов для определения тональности текстов //Математические структуры и моделирование. – 2019. – №. 4 (52). – С. 97-110.

6. Jurgens D. Learning about word vector representations and deep learning through implementing word2vec //Proceedings of the Fifth Workshop on Teaching NLP. – 2021. – Pp. 108-111.

7. Салып Б. Ю., Смирнов А. А. Анализ модели BERT как инструмента определения смысловой близости предложений естественного языка //StudNet. – 2022. – Т. 5. – №. 5. – С. 3509-3518.

8. Савенков П. А., Ивутин А. Н. Методы анализа естественного языка в задачах детектирования поведенческих аномалий //Известия Тульского государственного университета. Технические науки. – 2022. – №. 3. – С. 358-366.

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

10. Shinn N. et al. Reflexion: Language agents with verbal reinforcement learning // Advances in Neural Information Processing Systems. – 2024. – Vol. 36.

11. Степанов А. С., Степанов С. М. О смысле ошибок первого и второго рода // Актуальные проблемы авиации и космонавтики. – 2010. – Т. 1. – №. 6. – С. 239-241.

12. Савинов А. Н. и др. Анализ решения проблем возникновения ошибок первого и второго рода в системах распознавания клавиатурного почерка // Вестник Волжского университета им. В. Н. Татищева. – 2011. – №. 18. – С. 120-125.

13. Заикин Д. А. Подход к ранжированию результатов для терминологического поиска // Ученые записки Казанского университета. Серия Физико-математические науки. – 2014. – Т. 156. – №. 1. – С. 12-21.

14. Wang R., Li J. Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – 2019. – Pp. 4135-4145.

15. Быстров И. С., Котенко И. В. Показатели для оценки результатов машинного обучения применительно к задаче обнаружения кибер-инсайдеров // Региональная информатика (РИ-2022). – 2022. – С. 140-141.

Результаты процедуры рецензирования статьи

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

Рецензируемая статья направлена на практическое решение задачи по оптимизации традиционных методов определения сходства наименований проектов и закупок с использованием больших языковых моделей. Автор ставит цель, которая в современных условиях является актуальной и значимой. Объектом исследования является база

данных, содержащая наименования проектов и закупок (для сбора эмпирического материала были использованы открытые источники, содержащие инвестиционные программы и закупки электросетевых компаний), предметом исследования – определение релевантности наименования проекта информационному содержанию закупки. Стоит согласиться, что «научная значимость исследования заключается в применении оперативных возможностей больших языковых моделей в качестве дополнения к классическим методам обработки текстовой информации». Практический характер работы начилен: «для решения задачи определения сходства двух предложений могут быть использованы различные методы: 1) методы, использующие метрики сходства последовательности символов (сходство Жаккара, расстояние Левенштейна, сходство Джаро-Винклера и др.) [3, 4]; 2) методы, использующие преобразование текста в векторное представление (Word2vec, TF-IDF), что позволяет на следующем шаге выполнить расчет косинусного сходства [5, 6]; 3) методы, также использующие преобразование текста в векторное представление, но задействующие при этом глубокие нейронные сети с большим числом параметров (BERT, Gigachat, ChatGPT и др.) [7, 8]». Методология исследования синкретична, в статье использованы математические принципы рассмотрения и аргументации проблемы. Общий блок данных включен полновесно; считаю, что материал можно использовать далее. Суждения по ходу работы объективны, выверены: например, «было реализовано обращение по API к большим языковым моделям Gigachat, GPT 3.5, GPT-4 со следующим промптом: «Твоя задача: точно определить, есть ли связь между проектом и закупкой. Связь существует, только если название объекта (с прописной буквы) или его номер, или географический адрес в названии проекта и закупки полностью идентичны. "Проект: 'наименование проекта'. Закупка: 'наименование закупки'. Есть ли между ними точная связь?"», или «Приведем пример, как работает данный алгоритм. Для наименования каждого проекта и каждой закупки выполняется расчет TF-IDF, на основе чего выполняется попарный расчет косинусного сходства каждого проекта и каждой закупки, что позволяет выполнить первичный отбор потенциальных пар «проект – закупка» с косинусным сходством более 0,44. Далее для всех отобранных пар выполняется запрос по API к большой языковой модели. Возьмем в качестве примера пару, состоящую из проекта «Модернизация ПС 110/35/10 кВ Пречистое с монтажом оборудования систем видеонаблюдения» и закупки «Право на заключение Договора на выполнение строительно-монтажных работ по объектам: «Модернизация ПС 110/35/10 кВ Пречистое, ПС 110/35/10 кВ Ершичи, ПС 110/35/10 кВ Каспля с монтажом оборудования систем видеонаблюдения» для нужд ПАО «Россети Центр» и т.д. Примеры даются в режиме т.н. открытости данных; фактический уровень учтен. Работа дробится на смысловые / стандартные части, общая логика поддерживается на протяжении всего исследования. Обобщение полученных данных представлено в виде таблиц, схем, графиков, рисунков. Выводы промежуточного типа созвучны финальным: «применение больших языковых моделей является перспективным инструментом, позволяющим значительно улучшать результаты более быстрых, но менее точных традиционных подходов по обработке текстовых данных. Наиболее адекватной моделью для рассматриваемой задачи оказалась модель GPT-4 с добавленным блоком рефлексии, верно классифицировавшая 85 пар «проект – закупка» из 100. Вместе с тем относительное большое количество ложноотрицательных результатов и достигнутое значение F2-меры (0,65) указывают на то, что остается широкое пространство для более четкой и точной дифференциации пар «проект – закупка» на (не-)релевантные». Весьма удачен общий итог относительно того, что работа в данном русле должна быть продолжена: «авторы планируют продолжить исследование с целью дальнейшей оптимизации алгоритма сопоставления проектов и закупок как путем уточнения промптов для обращения к большим языковым моделям

(существуют и иные варианты улучшения промптов помимо введения блока рефлексии), так и использованием дополнительных алгоритмов классификации на вручную размеченных данных: например, модель градиентного бустинга с такими признаками, как длины наименований проектов и закупок, количество прописных букв в наименованиях проектов и закупок и др.». Требования издания учтены; цель работы достигнута; поставленный спектр задач решен. Рекомендую статью «Оптимизация традиционных методов определения сходства наименований проектов и закупок с использованием больших языковых моделей» к открытой публикации в журнале «Litera».