**Использование современных (лишнее – статья всегда посвящена современному) языковых моделей (основанных на использовании методов моделей ……) для распознавания информации в нумизматических описаниях.**

**Аннотация.** Статья затрагивает актуальную тему применения алгоритмов глубокого обучения и больших языковых моделей для анализа текстовых описаний в нумизматике. Особое внимание уделено неопределенности и вариативности нумизматического рынка, что делает его сложным для прогностических моделей. Исследование направлено на разработку метода обработки текстовых данных с использованием предварительно обученных языковых моделей для эффективного анализа нумизматических описаний.

Целью исследования является разработка методологии использования PLM в нумизматике для извлечения и классификации информации из текстов. Задача включает элементы распознавания именованных сущностей и требует адаптации моделей к специфике нумизматических описаний, которые часто бывают неструктурированными и представлены на разных языках.

Проведенная оценка эффективности моделей GPT 3.5 Turbo и GPT 4.5 Turbo показала высокую точность в идентификации и классификации нумизматического материала. Особое внимание уделено анализу необходимых атрибутов для различных категорий монет и медалей. Выявлено, что предобученные языковые модели эффективны при работе с заранее заданными атрибутами, но могут иметь ограничения в самостоятельном выделении комплексных атрибутов.

Использование GPT моделей оценивается как экономически выгодный подход, особенно в контекстах с ограниченными ресурсами. Однако стоимость API может ограничивать применение в проектах с большим объемом запросов. Планируется анализ альтернативных LLM и дальнейшая модификация алгоритмов для повышения эффективности и снижения затрат.

**Введение.** В современном контексте экономической информатики одной из заметных задач является прогнозирование ценовых колебаний в специфических рыночных сегментах. Достаточно интересным в этом отношении сегментом является нумизматический рынок. Характеризуясь высокой степенью неопределенности и вариативности, он представляет собой сложную область для прогностических моделей. Ключевым элементом таких моделей является способность точно интерпретировать и анализировать текстовые описания объектов, что ставит перед исследователями задачу эффективной обработки естественного языка (Natural Language Processing, NLP).

Текущее исследование фокусируется на разработке и апробации метода обработки текстовых данных, основанного на применении больших языковых моделей (LLM). В отличии от большинства подобных исследований, здесь внимание сосредоточено на использовании предварительно обученных языковых моделей (Pre-trained language models, PLM), как инструмента для извлечения информации из текстовых описаний монет. Основываясь на принципах глубокого обучения и адаптивности к контексту, PLM представляет собой передовую технологию в сфере NLP, что делает ее особенно подходящей для анализа сложных и неструктурированных данных, характерных для нумизматических описаний.

Цель данной статьи - представить методологический подход к использованию PLM для решения задачи NLP в контексте нумизматики и обозначить первоначальные результаты его применения. Особое внимание уделено оценке точности модели, анализу ее применимости для данной задачи, а также рассмотрению экономической эффективности в сравнении с другими методами и моделями. Исследование рассматривается как часть более обширного проекта по прогнозированию цен на нумизматический материал, предполагающего последующую разработку и сравнение различных подходов и моделей LLM.

**Описание Задачи.** В рамках данного исследования решается задача извлечения информации (Information Extraction), которая является подразделом области обработки естественного языка. Специфика задачи заключается в анализе и обработке текстовых данных для целей идентификации и структурирования определенных, релевантных данных из неструктурированных текстовых источников. Конкретно, исследование фокусируется на выделении и классификации ключевых атрибутов нумизматического материала, таких как период выпуска, географическое происхождение, используемый материал и состояние монет, из текстовых описаний.

Задача включает в себя элементы распознавания именованных сущностей (Named Entity Recognition, NER), требуя определения и категоризации специфических сущностей, присутствующих в тексте. Это необходимо для преобразования объемных массивов неструктурированных текстов в структурированную, систематизированную форму, подлежащую последующему аналитическому изучению и использованию в рамках нумизматической дисциплины.

Одна из проблем при сборе информации о ценообразовании нумизматического материала заключается в том, что информация берется не из строго систематизированных каталогов, а из пользовательских описаний, составленных в произвольной форме. Обычно это данные, вносимые пользователями на интернет-аукционах и торговых площадках. Эти данные зачастую являются неполными и характеризуются неформальными правилами описания, что увеличивает сложность задачи извлечения информации. Также следует отметить, что описания могут быть представлены на различных языках (иногда даже на латыни), что вносит дополнительную вариативность в анализируемые тексты.

Задача предполагает реализацию комплексного подхода к извлечению информации и её последующей классификации с целью установления однозначной идентификации каждого элемента нумизматического материала. В частности, ключевыми аспектами являются разработка и валидация алгоритмов для обработки и анализа текстовых данных, способных учитывать как явные, так и неявные лингвистические характеристики, присущие нумизматическим описаниям.

Для решения этой задачи предусмотрено исследование возможности использования предварительно обученных языковых моделей, которые обладают потенциалом глубокого понимания естественного языка и способны обрабатывать широкий спектр лингвистических нюансов. Важным аспектом исследования является анализ способности LLM извлекать значения ключевых атрибутов из текста. Это исследование стремится не только к созданию эффективных инструментов для классификации нумизматического материала, но и к расширению понимания возможностей современных технологий NLP в применении к различным прикладным задачам.

**Обзор литературы.**

Мало ссылок – нужно 10-15. Описываются достоинства и недостатки. И затем из недостатков вытекает постановка задачи.

На основе обзора обосновывается что не так (чего не хватает) и делается вывод что мы хотим улучшить или разработать и почему (что нас не устраивает в существующих методах). Т. е. ставится задача – что дано, что находим, какие методы и модели предполагается использовать…

Использование больших языковых моделей (LLM) для идентификации и структурирования информации из текстовых описаний является темой, которая в последнее время активно обсуждается в научной литературе. Большинство публикаций сосредотачивается на создании, переобучении (fine-tuning) или дообучении (retraining) этих моделей.

Подобные методы могут быть широко использованы, но имеют явные ограничения. Во-первых, большинство известных больших языковых моделей не являются открытыми или платными. Это затрудняет проведение серийных экспериментов и серийное использование. Кроме того, имеющееся небольшое количество моделей с открытым исходным кодом требуют много ресурсов GPU при использовании и обучении, что является проблемой для большинства исследователей [1].

Отдельно обращает на себя внимание исследование Alexander Brinkmann «Product Information Extraction using ChatGPT» [2]. В нём используется схожий подход для более общей предметной области. Акцент делается на характеристиках производительности модели.

Что касается идентификации информации непосредственно в нумизматике, то автору не удалось найти работ, посвященных именно этой предметной области. Тем не менее, обращают на себя внимание работы, посвященные распознаванию монет по изображениям. В частности, Understanding Ancient Coin Images [3]. Они позволяют оценить возможность принципиально другого подхода к решению задачи (хотя такое сопоставление вероятно не совсем корректно). С другой стороны, распознавание нумизматического материала по изображению может дополнять данные, полученные на основе текстовых описаний. Например, для объективной оценки состояния монет.

**Выбор инструментария.** На начальных стадиях данного исследования были выбраны GPT-3.5 Turbo и GPT-4.5 Turbo, разработанные компанией OpenAI, в качестве основного инструмента обработки текста. Такой выбор обусловлен несколькими ключевыми факторами. Во-первых, передовая способность этих алгоритмов к генерации и пониманию естественного языка позволяет эффективно обрабатывать неструктурированные текстовые описания монет. Во-вторых, благодаря обширной обучающей базе и мощным алгоритмам обучения, модели GPT способны выявлять сложные паттерны и извлекать ключевую информацию из текста, что критически важно для идентификации атрибутов нумизматического материала. Наконец, гибкость API GPT обеспечивает удобную интеграцию в существующую систему обработки данных.

**Процесс интеграции LLM.** Процесс интеграции GPT с системой обработки данных включает несколько ключевых этапов. Во-первых, описания лотов на естественном языке собираются и систематизируются в реляционной базе данных. К сожалению, большинство интернет-площадок, не предоставляют API для сбора данных о продажах. Но и не запрещают собирать данные в автоматическом режиме, а в большинстве случаев предоставляют специальные теги для «пауков».

Описание каждого лота передается в небольшой Python-скрипт, который функционирует как интерфейс между базой данных и GPT. Python-скрипт формирует и отправляет запросы к GPT через его API, где входным параметром является текстовое описание монеты. Запросы разработаны таким образом, чтобы модель GPT возвращала значения атрибутов в формате JSON. Эти атрибуты включают в себя заранее определенные характеристики для каждой категории монеты, что позволяет однозначно идентифицировать и классифицировать нумизматический материал.

После получения ответа от GPT, скрипт анализирует возвращенные значения атрибутов и проводит соответствующее сопоставление с записями в справочнике базы данных. В случае обнаружения совпадения запись идентифицируется; если совпадение отсутствует, создается новая запись. Важно отметить, что GPT способен заполнять недостающие атрибуты на основе контекста описания, что значительно повышает точность и эффективность процесса идентификации.

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

Figure 1. Диаграмма последовательности, иллюстрирующая процесс интеграции GPT с системой обработки данных.

Нужно более подробно расписать Диаграмму и связать с **Процесс интеграции LLM** котрый описан выше.

Эта методология ?? позволяет использовать потенциал предобученных LLM для извлечения и классификации информации из текстовых данных в сфере нумизматики, обеспечивая высокую степень точности и эффективности обработки запросов.

Главное, что не хватает --- эксперимента – экспериментальной проверки предложенного подхода (это в статье обязательный раздел!!!). На его основе оцениваются результаты и делается вывод, что цель исследования достигнута получены новы, важные научные результаты—например, сделана модификация существующего метода, разработан новый метод, который чем-то лучше, эффективнее, существующих.

Поэтому нужно сравнение результатов, полученных на основе разработанного метода с существующими аналогичными (аналогами) и констатация факта, что в чем-то он превосходит эти аналоги (скорость, точность). Обычно приводят графики, таблицы сравнения и т.д.

**Результаты.** В ходе исследования была проведена оценка эффективности моделей GPT 3.5 Turbo и GPT 4.5 Turbo в задаче распознавания и классификации нумизматического материала на основе текстовых описаний. Опираясь на выборку из примерно 100 записей, было установлено, что GPT 4.5 Turbo демонстрирует достаточно высокую точность: в анализированной выборке ошибок не обнаружено. Младшая модель, GPT 3.5 Turbo, выдала более 80 процентов корректных результатов. Этот результат подчеркивает способность LLM эффективно обрабатывать и интерпретировать неструктурированные текстовые данные, точно выделяя ключевые атрибуты и характеристики нумизматического материала.

В рамках проведенного исследования особое внимание было уделено анализу наборов атрибутов, необходимых для классификации различных категорий нумизматического материала. Один из ключевых выводов заключается в том, что набор атрибутов значительно варьируется в зависимости от категории монет и медалей. Эта особенность предъявляет уникальные требования к процессу идентификации и классификации, так как каждая категория требует индивидуального подхода к определению атрибутов.

В процессе работы выяснилось, что определение и настройка этих атрибутов для каждой категории требовали ручного вмешательства. Попытки использовать предварительно обученные языковые модели (PLM), такие как GPT 4.5 Turbo, для автоматизации процесса выделения и настройки атрибутов из текстовых описаний не принесли ожидаемых результатов.

Наличие описаний на различных языках в данном случае не стало проблемой. Были использованы модели, обученные на нескольких языках. Исследователи полагают, что многоязычные языковые модели превосходят модели, обученные только на одном языке. [4]

Таким образом, PLM показали высокую эффективность в идентификации и классификации нумизматического материала на основе заранее заданных атрибутов, однако их способность к самостоятельному выделению и определению комплексных наборов атрибутов из текста оказалась ограниченной.

**Экономический аспект.** С точки зрения экономической эффективности, использование моделей GPT представляет собой доступный и удобный подход, не требующий значительных вычислительных ресурсов или глубоких знаний в области программирования для развертывания и обучения собственных моделей. Основным недостатком является стоимость использования API GPT 4.5 Turbo, которая на момент проведения исследования составляла около 3 центов за запрос. Помимо этого, GPT 4.5 Turbo имеет ограничения на количество запросов в единицу времени. Младшая модель GPT избавлена от этих ограничений и стоит почти на порядок дешевле. Но вместе с тем, при ее использовании необходима проверка результатов. Несмотря на это, простота реализации и отсутствие необходимости в сложном оборудовании делают этот подход экономически привлекательным, особенно для задач, где не требуется частая обработка больших объемов данных.

В целом, результаты исследования демонстрируют, что использование LLM GPT в качестве инструмента для классификации и идентификации нумизматического материала является эффективным с точки зрения точности и имеет потенциал для экономически выгодной реализации, особенно в контекстах с ограниченными ресурсами. Однако стоимость использования API может ограничивать его применение в проектах с большим объемом запросов.

Отдельный интерес представляет задача определение качества монеты, которая упоминалась выше. Несмотря на то, что существуют чёткие градации качества нумизматического материала, большинство выставляемых на аукцион монет не проходит оценку качеству профессионалами.

**Оценка и анализ результатов.** Использование GPT-4.5 Turbo в настоящем исследовании демонстрирует способность просто и эффективно решать задачу классификации нумизматического материала на основе текстовых описаний. Этот подход обеспечивает точное извлечение ключевых атрибутов из разнообразных описаний, демонстрируя значительные перспективы для текущего и будущего использования в аналогичных задачах.

Предварительно обученные языковые модели, опираясь на сложные алгоритмы обработки естественного языка, демонстрируют значительный потенциал в автоматизации задач, связанных с извлечением и интерпретацией текстовой информации практически в любых дисциплинарных контекстах.

**Планы по продолжению исследования. *Проверка альтернативных моделей LLM.*** В рамках текущего исследования предусматривается анализ и сравнение ряда доступных предварительно обученных языковых моделей, обладающих API, для оценки их эффективности в задачах аналогичного характера. Рассмотрение разнообразных моделей позволит получить более глубокое представление о потенциале PLM в контексте извлечения информации из текстов. Перечень моделей, потенциально подходящих для изучения, включает следующие LLM:

1. **Bard**: Модель, ориентированная на творческое письмо и создание историй, идеально подходит для генерации увлекательного контента.
2. **GooseAI**: Сфокусирована на создании качественного, привлекательного контента, особенно ценна для маркетологов и создателей контента благодаря своей способности понимать человеческие эмоции.
3. **Cohere**: Многофункциональная модель для различных задач NLP, включая классификацию текста, суммаризацию и анализ сентиментов, с возможностью настройки под конкретные задачи.
4. **Claude**: Новая модель на рынке, привлекает внимание своей способностью генерировать оригинальный и увлекательный контент, идеальна для маркетологов.
5. **PaLM**: Разработана для понимания языка и может использоваться в широком спектре приложений NLP, включая чат-боты, переводчики и поисковые системы. [5]

Стоит подчеркнуть, что, несмотря на то что на момент написания данной статьи модель GPT 4.5 Turbo превосходит большинство аналогов по основным характеристикам, многие эксперты указывают на то, что для разных задач различные другие модели могут стать наиболее подходящими моделями языковых нейросетей. [6]

***Модификация алгоритма для снижения количества запросов.*** С целью оптимизации процесса и снижения количества запросов к LLM, предполагается модификация алгоритма. Планируется создание словарей в процессе распознавания текстовых описаний для заполнения значений атрибутов и использование LLM только для объектов, не распознанных с помощью этих словарей. Это повысит эффективность системы и сократит затраты на обработку запросов.

***Дополнение системы распознаванием изображений.*** Для расширения функциональности системы планируется ее дополнение модулем распознавания изображений. Это позволит решать более широкий спектр задач, связанных не только с текстовым анализом, но и с визуальной идентификацией нумизматического материала. Модели ИИ, подобные OpenAI CLIP (Contrastive Language–Image Pretraining), выступают в качестве базовых моделей или моделей, обученных на большом количестве немаркированных данных, которые могут быть адаптированы ко многим приложениям. [7]   
В рамках исследования уже проводились опыты использования CLIP по распознаванию изображений монет, где использовался аналогичный подход. Это подтверждает потенциал мультимодальных моделей искусственного интеллекта в решении комплексных задач, сочетающих анализ текста и изображений, и открывает новые перспективы для их применения в нумизматике.

Литература: По хорошему должно быть около 20 ссылок

1. Xuanfan Ni1, Piji Li1, Huayang Li (2023). Unified Text Structuralization with Instruction-tuned Language Models. Получено с <https://arxiv.org/abs/2303.14956>
2. Alexander Brinkmann, Roee Shraga, Reng Chiz Der, Christian Bizer (2023). Product Information Extraction using ChatGPT. Получено с <https://doi.org/10.48550/arXiv.2306.14921>
3. Jessica Cooper and Ognjen Arandjelovi´c (2019). Unified Text Structuralization with Instruction-tuned Language Models. Получено с <https://arxiv.org/abs/1903.02665v2>
4. [Gabriel Nicholas](https://arxiv.org/search/cs?searchtype=author&query=Nicholas,+G), [Aliya Bhatia](https://arxiv.org/search/cs?searchtype=author&query=Bhatia,+A). (2023). Lost in Translation: Large Language Models in Non-English Content Analysis Получено с <https://arxiv.org/abs/2306.07377>
5. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLAMA: Open and efficient foundation language models. Получено с arXiv:2306.07377.
6. Jiang, D., Ren, X., & Lin, B. Y. (2023). LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. Получено из <https://doi.org/10.48550/arXiv.2306.02561>
7. Nwatu, J., Ignat, O., & Mihalcea, R. (2023). Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models. arXiv. DOI: 10.48550/arxiv.2311.05746. Получено с <https://techxplore.com/news/2023-12-biases-large-image-text-ai-favor.html>