*Серія: ТЕХНІКА*

**УДК: 004.93 (477)**

*БОЖКО ОЛЕКСАНДР ЮРІЙОВИЧ***,**

аспірант факультету комп’ютерних наук,

Харківський національний університет радіоелектроніки,

Україна, 61166, м. Харків, пр. Науки, 14

ORCID: <https://orcid.org/0009-0004-6820-1228>

e-mail: [oleksandr.bozhko@nure.ua](mailto:oleksandr.bozhko@nure.ua)

**ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ РОЗПІЗНАВАННЯ ІНФОРМАЦІЇ В НУМІЗМАТИЧНИХ ОПИСАХ**

**Анотація.** Стаття зачіпає актуальну тему застосування алгоритмів глибокого навчання і великих мовних моделей для аналізу текстових описів у нумізматиці. Особливу увагу приділено невизначеності та варіативності нумізматичного ринку, що робить його складним для прогностичних моделей. Дослідження спрямоване на розробку методу опрацювання текстових даних з використанням попередньо навчених мовних моделей для ефективного аналізу нумізматичних описів.

Метою дослідження є розробка методології використання попередньо навчених мовних моделей у нумізматиці для отримання та класифікації інформації з текстів. Завдання включає елементи розпізнавання іменованих сутностей і вимагає адаптації моделей до специфіки нумізматичних описів, що часто бувають неструктурованими та подані різними мовами.

Проведена оцінка ефективності моделей GPT 3.5 Turbo і GPT 4 Turbo показала високу точність в ідентифікації та класифікації нумізматичного матеріалу. Особливу увагу приділено аналізу необхідних атрибутів для різних категорій монет і медалей. Виявлено, що попередньо навчені мовні моделі ефективні під час роботи із заздалегідь заданими атрибутами, але можуть мати обмеження в самостійному виділенні комплексних атрибутів.

Використання GPT моделей оцінюється як економічно вигідний підхід, особливо в контекстах з обмеженими ресурсами. Однак вартість API може обмежувати застосування в проектах із великим обсягом запитів. Планується аналіз альтернативних великих мовних моделей і подальша модифікація алгоритмів для підвищення ефективності та зниження витрат.

**Ключові слова:** *глибоке навчання, великі мовні моделі, попередньо навчені мовні моделі, класифікація інформації, аналіз текстових описів, ідентифікація атрибутів, нумізматика.*

**Постановка проблеми.** У сучасному контексті економічної інформатики одним із помітних завдань є прогнозування цінових коливань у специфічних ринкових сегментах. Досить цікавим у цьому плані сегментом є нумізматичний ринок. Характеризуючись високим ступенем невизначеності та варіативності, він являє собою складну область для прогностичних моделей. Ключовим елементом таких моделей є здатність точно інтерпретувати й аналізувати текстові описи об'єктів, що ставить перед дослідниками завдання ефективної обробки природної мови (Natural Language Processing, NLP) [1].

Одна з проблем під час збирання інформації про ціноутворення нумізматичного матеріалу полягає в тому, що інформацію беруть не з суворо систематизованих каталогів, а з користувацьких описів, складених у довільній формі. Зазвичай це дані, що вносяться користувачами на інтернет-аукціонах і торгових майданчиках. Ці дані часто є неповними і характеризуються неформальними правилами опису, що збільшує складність завдання вилучення інформації. Також слід зазначити, що описи можуть бути подані різними мовами (іноді навіть латиною), що вносить додаткову варіативність в аналізовані тексти.

**Аналіз останніх досліджень і публікацій.** Використання великих мовних моделей (LLM) [2] для ідентифікації та структурування інформації з текстових описів є темою, яка останнім часом активно обговорюється в науковій літературі в основному в працях зарубіжних вчених, зокрема, таких як Бріхман О., Ізакард Г., Крістофер Д., Лавріл Т., Туврон Х., Рен X., Цзян Д. Втім, більшість публікацій зосереджується на створенні, навчанні (fine-tuning) [3] або перенавчанні (retraining) [4] цих моделей.

Подібні методи можуть бути широко використані, але мають явні обмеження. По-перше, більшість відомих великих мовних моделей не є відкритими або платними. Це ускладнює проведення серійних експериментів і серійне використання. Крім того, наявна невелика кількість моделей з відкритим вихідним кодом вимагають багато ресурсів GPU під час використання та навчання, що є проблемою для більшості дослідників [5].

Окремо звертає на себе увагу дослідження О. Бріхмана на тему «Product Information Extraction using ChatGPT [6]», в якому використовується схожий підхід для більш загальної предметної області та робиться акцент на характеристиках продуктивності моделі. Щодо ідентифікації інформації безпосередньо в нумізматиці, то, нажаль, не має жодної наукової роботи, присвяченої саме цій предметній сфері. Проте звертають на себе увагу роботи, присвячені питанням розпізнавання монет за зображеннями. У цьому випадку стане в нагоді наукова робота Дж. Купера «Understanding Ancient Coin Images» [7], яка надає змогу оцінити можливість принципово іншого підходу до розв'язання поставленої задачі (хоча, на нашу думку, й не зовсім коректне). З іншого боку, розпізнавання нумізматичного матеріалу за зображенням може доповнювати дані, отримані на основі текстових описів. Наприклад, для об'єктивної оцінки стану монет. Саме тому є необхідність в глибокому дослідженні проблеми використання великих мовних моделей для розпізнавання інформації в нумізматичних описах. Дана стаття може розглядатися як частина ширшого проекту з прогнозування цін на нумізматичний матеріал, який передбачає подальше розроблення і порівняння різних підходів і моделей LLM.

**Метою статті є**  розв'язання задачі NLP за допомогою методологічного підходу до використання попередньо навчених мовних моделей (Pre-trained language models, PLM) та у контексті нумізматики окреслити початкові результати його застосування. Особливу увагу приділено оцінці точності моделі, аналізу її застосовності для цієї задачі, а також розгляду економічної ефективності порівняно з іншими методами і моделями.

**Виклад основного матеріалу.** Специфікарозв'язання задачі вилучення інформації (Information Extraction) [8], що є підрозділом галузі обробки природної мови, полягає в аналізі та обробці текстових даних для цілей ідентифікації і структуруванні певних, релевантних даних із неструктурованих текстових джерел. Конкретно, дослідження фокусується на виділенні та класифікації ключових атрибутів нумізматичного матеріалу, таких як період випуску, географічне походження, матеріал, що використовується, і стан монет, з текстових описів.

Так, завдання містить елементи розпізнавання іменованих сутностей (Named Entity Recognition, NER) [9], вимагаючи визначення і категоризації специфічних сутностей, присутніх у тексті. Це необхідно для перетворення об'ємних масивів неструктурованих текстів на структуровану, систематизовану форму, яка підлягає подальшому аналітичному вивченню і використанню в рамках нумізматичної дисципліни. Також задача включає в себе елементи семантичної нормалізації, зокрема:

- уніфікація синонімів: перетворення різних термінів, що позначають одне й те саме поняття, на єдину стандартну форму;

- розв'язання амбігвітності: визначення та усунення неоднозначності в тексті, наприклад, коли одне й те саме слово може мати кілька значень залежно від контексту;

- стандартизація форматів: приведення даних до однакового формату, наприклад, стандартизація дат, часу, валютних значень тощо [10].

Поточне дослідження фокусується на розробці та апробації методу опрацювання текстових даних, заснованого на застосуванні великих мовних моделей. На відміну від більшості подібних досліджень, тут увага зосереджена на використанні попередньо навчених мовних моделей, як інструменту для вилучення інформації з текстових описів монет. Ґрунтуючись на принципах глибокого навчання й адаптивності до контексту, PLM є передовою технологією у сфері NLP, що робить її особливо придатною для аналізу складних і неструктурованих даних, характерних для нумізматичних описів.

Завдання передбачає реалізацію комплексного підходу до вилучення інформації та її подальшої класифікації з метою встановлення однозначної ідентифікації кожного елемента нумізматичного матеріалу. Зокрема, ключовими аспектами є розробка і валідація алгоритмів для оброблення та аналізу текстових даних, здатних враховувати як явні, так і неявні лінгвістичні характеристики, властиві нумізматичним описам.

Для розв'язання цього завдання досліджували можливість використання попередньо навчених мовних моделей, які володіють потенціалом глибокого розуміння природної мови і здатні обробляти широкий спектр лінгвістичних нюансів. Важливим аспектом дослідження є аналіз здатності LLM витягувати значення ключових атрибутів із тексту. Це дослідження прагне не тільки до створення ефективних інструментів для класифікації нумізматичного матеріалу, а й до розширення розуміння можливостей сучасних технологій NLP у застосуванні до різних прикладних завдань.

На початкових стадіях цього дослідження було обрано GPT-3.5 Turbo і GPT-4 Turbo (gpt-4-1106-preview)[11], розроблені компанією OpenAI, як основний інструмент обробки тексту. Такий вибір зумовлений кількома ключовими факторами. По-перше, передова здатність цих алгоритмів до генерації та розуміння природної мови дозволяє ефективно обробляти неструктуровані текстові описи монет. По-друге, завдяки великій навчальній базі і потужним алгоритмам навчання, моделі GPT здатні виявляти складні патерни і витягувати ключову інформацію з тексту, що критично важливо для ідентифікації атрибутів нумізматичного матеріалу. Нарешті, гнучкість API GPT забезпечує зручну інтеграцію в наявну систему обробки даних.

У свою чергу, процес інтеграції GPT із системою обробки даних включає кілька ключових етапів. По-перше, описи лотів природною мовою збираються і систематизуються в реляційній базі даних. На жаль, більшість інтернет-майданчиків не надають API для збору даних про продажі. Але й не забороняють збирати дані в автоматичному режимі, а в більшості випадків надають спеціальні теги для «павуків».

Опис кожного лота передається в невеликий Python-скрипт, який функціонує як інтерфейс між базою даних і GPT. Python-скрипт формує і надсилає запити до GPT через його API, де вхідним параметром є текстовий опис монети. Запити розроблені таким чином, щоб модель GPT повертала значення атрибутів у форматі JSON. Ці атрибути містять заздалегідь визначені характеристики для кожної категорії монети, що дає змогу однозначно ідентифікувати та класифікувати нумізматичний матеріал.

Після отримання відповіді від GPT, скрипт аналізує повернуті значення атрибутів і проводить відповідне зіставлення із записами в довіднику бази даних. У разі виявлення збігу запис ідентифікується; якщо збіг відсутній, створюється новий запис. Важливо зазначити, що GPT здатний заповнювати атрибути, яких бракує, на основі контексту опису, що значно підвищує точність і ефективність процесу ідентифікації (Див. малюнок 1).

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

***Малюнок 1****. Діаграма послідовності, що ілюструє процес інтеграції GPT із системою обробки даних.*

Такий підхід дає змогу використовувати потенціал навчання LLM для вилучення і класифікації інформації з текстових даних у сфері нумізматики, забезпечуючи високий ступінь точності та ефективності опрацювання запитів.

Під час дослідження було проведено оцінку ефективності моделей GPT 3.5 Turbo і GPT 4 Turbo у завданні розпізнавання та класифікації нумізматичного матеріалу на основі текстових описів. Спираючись на вибірку з приблизно 1000 записів, було встановлено, що обидві моделі демонструють досить високу якість виявлення значень атрибутів.

Попереднє оцінювання за допомогою стандартних метрик, які зазвичай використовують для оцінювання алгоритмів отримання інформації, визначило:

1. ***Точність (Precision)****:* частка правильно ідентифікованих елементів серед усіх елементів, які алгоритм класифікував як позитивні. Precision = TP / (TP + FP), де TP - істинно позитивні результати, FP - хибно позитивні результати.
2. ***Повнота (Recall)****:* частка правильно ідентифікованих позитивних елементів з усіх реальних позитивних елементів. Recall = TP / (TP + FN), де TP - істинно позитивні результати, FN - хибно негативні результати.
3. ***F1-міра***: гармонійне середнє між точністю і повнотою, що забезпечує баланс між ними. F1 = 2 \* (Precision \* Recall) / (Precision + Recall).

Результати оцінки, наведені в табл. 1, підкреслюють здатність обох LLM ефективно опрацьовувати та інтерпретувати неструктуровані текстові дані, точно виокремлюючи ключові атрибути та характеристики нумізматичного матеріалу.

*Таблиця 1*

**Результати оцінки алгоритмів отримання інформації**

|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Точність** | **Повнота** | **F1-міра** |
| ChatGPT 4,5 | 0,997 | 0,996 | 0,997 |
| ChatGPT 3,5 Turbo | 0,988 | 0,987 | 0,987 |

Також у рамках проведеного в статті дослідження особливу увагу було приділено аналізу наборів атрибутів, необхідних для класифікації різних категорій нумізматичного матеріалу. Один із ключових висновків полягає в тому, що набір атрибутів значно варіюється залежно від категорії монет і медалей. Ця особливість висуває унікальні вимоги до процесу ідентифікації та класифікації, оскільки кожна категорія вимагає індивідуального підходу до визначення атрибутів.

У процесі роботи з'ясувалося, що визначення та налаштування цих атрибутів для кожної категорії вимагали ручного втручання. Спроби використовувати попередньо навчені мовні моделі (PLM), як-от GPT 4 Turbo, для автоматизації процесу виділення та налаштування атрибутів із текстових описів не принесли очікуваних результатів. Наявність описів різними мовами в цьому випадку не стала проблемою. Були використані моделі, навчені кількома мовами. Дослідники вважають, що багатомовні мовні моделі перевершують моделі, навчені тільки однією мовою [12].

Таким чином, PLM показали високу ефективність в ідентифікації та класифікації нумізматичного матеріалу на основі заздалегідь заданих атрибутів, проте їхня здатність до самостійного виділення та визначення комплексних наборів атрибутів із тексту виявилася обмеженою.

На жаль, автору не вдалося знайти у відкритих джерелах описів інших досліджень, що використовують моделі машинного навчання в галузі нумізматики. Однак, ці результати можна порівняти з результатами подібних досліджень у галузі медицини [13].

Хоча, слід зазначити, що результати роботи однієї й тієї самої переднавченої моделі можуть істотно відрізнятися при її застосуванні в різних предметних областях. Це може бути зумовлено кількома чинниками. Зокрема, специфіка мови та термінології, структурою та форматом даних, контекстуальними особливостями, різноманітністю та складністю даних.

Перевагою використання моделей GPT є доступний і зручний підхід, який не потребує значних обчислювальних ресурсів або глибоких знань у сфері програмування для розгортання та навчання власних моделей.

Основним недоліком є вартість використання API GPT 4 Turbo. GPT 4 Turbo, що наведено на малюнку 2. Для розпізнавання опису однієї монети використовувалося в середньому 4 тис. вихідних і 4 тис. вхідних токенів.

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

***Малюнок 2****. Вартість вхідних і вихідних токенів для різних моделей [14].*

Потрібно зазначити, що GPT 4 Turbo має обмеження на кількість запитів за одиницю часу. Молодша модель GPT позбавлена цих обмежень. Але разом з тим, при її використанні необхідна перевірка результатів.

GPT 3.5 Turbo має найвищу середню швидкість відгуку (малюнок 3).

Изображение выглядит как текст, снимок экрана, диаграмма

Автоматически созданное описание

***Малюнок 3****. Середня швидкість відгуку для різних моделей [14].*

Незважаючи на це, простота реалізації та відсутність необхідності в складному обладнанні роблять цей підхід економічно привабливим, особливо для завдань, де не потрібне часте опрацювання великих обсягів даних.

Загалом, результати дослідження демонструють, що використання LLM GPT як інструменту для класифікації та ідентифікації нумізматичного матеріалу є ефективним з погляду точності та має потенціал для економічно вигідної реалізації, особливо в контекстах з обмеженими ресурсами. Однак вартість використання API може обмежувати його застосування в проєктах із великим обсягом запитів.

Окремий інтерес представляє завдання визначення якості монети, яка згадувалася вище. Незважаючи на те, що існують чіткі градації якості нумізматичного матеріалу, більшість монет, що виставляються на аукціон, не проходять оцінку якості професіоналами.

У межах поточного дослідження передбачається аналіз і порівняння низки доступних попередньо навчених мовних моделей, що володіють API, для оцінювання їхньої ефективності в завданнях аналогічного характеру. Розгляд різноманітних моделей дасть змогу отримати глибше уявлення про потенціал PLM у контексті вилучення інформації з текстів. Варто підкреслити, що, незважаючи на те, що на момент написання цієї статті модель GPT 4 Turbo перевершує більшість аналогів за основними характеристиками, багато експертів вказують на те, що для різних завдань різні інші моделі можуть стати найбільш підходящими моделями мовних нейромереж [15].

З метою оптимізації процесу і зниження кількості запитів до LLM, передбачається модифікація алгоритму. Планується створення словників у процесі розпізнавання текстових описів для заповнення значень атрибутів і використання LLM тільки для об'єктів, не розпізнаних за допомогою цих словників. Це підвищить ефективність системи і скоротить витрати на обробку запитів. Для розширення функціональності системи планується її доповнення модулем розпізнавання зображень. Це дасть змогу розв'язувати ширший спектр завдань, пов'язаних не тільки з текстовим аналізом, а й з візуальною ідентифікацією нумізматичного матеріалу. Моделі ШІ, подібні до OpenAI CLIP (Contrastive Language-Image Pretraining), виступають базовими моделями або моделями, навченими на великій кількості немаркованих даних, які можуть бути адаптовані до багатьох додатків. Зазначимо, що у рамках дослідження вже проводилися аналізи використання CLIP з розпізнавання зображень монет, де використовувався аналогічний підхід. Це підтверджує потенціал мультимодальних моделей штучного інтелекту в розв'язанні комплексних завдань, що поєднують аналіз тексту та зображень, і відкриває нові перспективи для їхнього застосування в нумізматиці.

**Висновки.** Підбиваючи підсумки усього вище вказано, слід зазначити, що дослідження використання великих мовних моделей OpenAI демонструє здатність просто і ефективно вирішувати завдання класифікації нумізматичного матеріалу на основі текстових описів. Цей підхід забезпечує точне вилучення ключових атрибутів з різноманітних описів, демонструючи значні перспективи для поточного і майбутнього використання в аналогічних завданнях. При цьому попередньо навчені мовні моделі, спираючись на складні алгоритми опрацювання природної мови, демонструють значний потенціал в автоматизації завдань, пов'язаних із витяганням та інтерпретацією текстової інформації практично в будь-яких дисциплінарних контекстах.

***Література:***

1. Christopher D. Manning. Human Language Understanding & Reasoning. Daedalus. Cambridge, UK. Spring 2022. URL: <https://www.amacad.org/publication/human-language-understanding-reasoning>.
2. Wei Jason. Emergent Abilities of Large Language Models. Jun 2022. URL: <https://doi.org/10.48550/arXiv.2206.07682>
3. Aston Zhang, Zachary C. Lipton, Mu Li, Alexander J. Smola. Dive into Deep Learning. ‎ Cambridge. UK: Cambridge University Press. 2023. ‎ 574 p.
4. Cem Dilmegan. Model Retraining: Why & How to Retrain ML Models?. January 11, 2024. URL: <https://research.aimultiple.com/model-retraining>.
5. Xuanfan Ni1, Piji Li1, Huayang Li. Unified Text Structuralization with Instruction-tuned Language Models. 2023. URL: <https://arxiv.org/abs/2303.14956>
6. Alexander Brinkmann, Roee Shraga, Reng Chiz Der, Christian Bizer. Product Information Extraction using ChatGPT. 2023. URL: <https://doi.org/10.48550/arXiv.2306.14921>
7. Jessica Cooper and Ognjen Arandjelovi´c. Understanding Ancient Coin Images. 2019. URL: <https://doi.org/10.48550/arXiv.1903.02665>
8. 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.
9. 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>
10. Kim, Y., Lee, J. H., Choi, S., Lee, J. M., Kim, J.‑H., Seok, J., & Joo, H. J. (2020). Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Получено с <https://doi.org/10.1038/s41598-020-77258-w>
11. Open AI Official Models Documentation. URL: <https://platform.openai.com/docs/models/>
12. Gabriel Nicholas, Aliya Bhatia. Lost in Translation: Large Language Models in Non-English Content Analysis. 2023. URL: <https://arxiv.org/abs/2306.07377>
13. 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>
14. LLM comparison highlights. URL: <https://artificialanalysis.ai/>
15. Dastan Hussen Maulud, Subhi R. M. Zeebaree. A State of Art for Semantic Analysis of Natural Language Processing. 2021. URL: <https://journal.qubahan.com/index.php/qaj/article/view/44>

***References:***

1. Manning, C. D. (2022). Human Language Understanding & Reasoning. Daedalus. Cambridge, UK. Retrieved from <https://www.amacad.org/publication/human-language-understanding-reasoning>.
2. Wei, J. (2022, June). Emergent Abilities of Large Language Models. Retrieved from <https://doi.org/10.48550/arXiv.2206.07682>.
3. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). Dive into Deep Learning. Cambridge, UK: Cambridge University Press.
4. Dilmegan, C. (2024, January 11). Model Retraining: Why & How to Retrain ML Models? Retrieved from <https://research.aimultiple.com/model-retraining>.
5. Ni, X., Li, P., & Li, H. (2023). Unified Text Structuralization with Instruction-tuned Language Models. Retrieved from <https://arxiv.org/abs/2303.14956>.
6. Brinkmann, A., Shraga, R., Der, R. C., & Bizer, C. (2023). Product Information Extraction using ChatGPT. Retrieved from <https://doi.org/10.48550/arXiv.2306.14921>.
7. Cooper, J., & Arandjelovi´c, O. (2019). Understanding Ancient Coin Images. Retrieved from <https://doi.org/10.48550/arXiv.1903.02665>.
8. 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. Retrieved from arXiv:2306.07377.
9. Jiang, D., Ren, X., & Lin, B. Y. (2023). LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. Retrieved from <https://doi.org/10.48550/arXiv.2306.02561>.
10. Kim, Y., Lee, J. H., Choi, S., Lee, J. M., Kim, J.‑H., Seok, J., & Joo, H. J. (2020). Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Retrieved from <https://doi.org/10.1038/s41598-020-77258-w>.
11. OpenAI. (n.d.). Official Models Documentation. Retrieved from <https://platform.openai.com/docs/models/>.
12. Nicholas, G., & Bhatia, A. (2023). Lost in Translation: Large Language Models in Non-English Content Analysis. Retrieved from <https://arxiv.org/abs/2306.07377>.
13. Nwatu, J., Ignat, O., & Mihalcea, R. (2023). Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models. Retrieved from <https://techxplore.com/news/2023-12-biases-large-image-text-ai-favor.html>.
14. (n.d.). LLM comparison highlights. Retrieved from <https://artificialanalysis.ai/>.
15. Maulud, D. H., & Zeebaree, S. R. M. (2021). A State of Art for Semantic Analysis of Natural Language Processing. Retrieved from <https://journal.qubahan.com/index.php/qaj/article/view/44>.

**Bozhko Oleksandr,** graduate student of the faculty of computer sciences (specialty 122 − Computer Sciences), Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine, e-mail: oleksandr.bozhko@nure.ua, <https://orcid.org/0009-0004-6820-1228>

ABSTRACT

BOZHKO OLEKSANDR. USING LARGE LANGUAGE MODELS TO RECOGNIZE INFORMATION IN NUMISMATIC DESCRIPTIONS

**Abstract.** The article touches upon the relevant topic of applying deep learning algorithms and large language models to analyze textual descriptions in numismatics. Particular attention is paid to the uncertainty and variability of the numismatic market, which makes it challenging for predictive models. The research is aimed at developing a method for processing textual data using pre-trained language models for the efficient analysis of numismatic descriptions.

The aim of the study is to develop a methodology for using pre-trained language models in numismatics to extract and classify information from texts. The task involves elements of named entity recognition and requires the adaptation of models to the specifics of numismatic descriptions, which are often unstructured and presented in different languages.

The evaluation of the effectiveness of the GPT 3.5 Turbo and GPT 4 Turbo models showed high accuracy in the identification and classification of numismatic material. Particular attention is paid to the analysis of the necessary attributes for different categories of coins and medals. It was found that the pre-trained language models are effective when working with predefined attributes but may have limitations in the independent selection of complex attributes.

The use of GPT models is a cost-effective approach, especially in resource-constrained contexts. However, the cost of the API may limit its use in projects with a large volume of requests. We plan to analyze alternative large language model and further modify the algorithms to improve efficiency and reduce costs.

**Keywords:** deep learning, large language models, pre-trained language models, information classification, analysis of textual descriptions, identification of attributes, numismatics.