

УДК 004.89:81'32:82-1

DOI <https://doi.org/10.32689/maup.philol.2025.1.11>**Володимир ПАСІЧНИК**

доктор технічних наук, професор, професор кафедри інформаційних систем та мереж,

Національний університет «Львівська політехніка», vpasichnyk@gmail.com

ORCID: 0000-0002-5231-6395

Максим ЯРОМИЧ

аспірант кафедри прикладної лінгвістики, Національний університет «Львівська політехніка»,

yamax0312@gmail.com

ORCID: 0009-0005-3299-6695

**ЖАНРОВА КЛАСИФІКАЦІЯ ЛІТЕРАТУРИ ЗА МЕТРИКАМИ
ЗА ДОПОМОГОЮ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ**

У статті проаналізовано можливість використання великих мовних моделей для жанрової класифікації літературних текстів із застосуванням метрик параметризації та жанрових онтологій. Розглянуто теоретичні основи жанрової класифікації, зокрема традиційні підходи до визначення жанрів та сучасні алгоритмічні методи, що використовують великі мовні моделі. Значну увагу приділено вибору метрик для параметризації текстів, серед яких: рівень формальності, глибина технічного аналізу, методологічний підхід, цільова аудиторія, область застосування, тип даних у дослідженні, наявність емпіричних результатів і спосіб візуалізації інформації. Запропоновано використання багаторівневої класифікації, що дозволяє більш точно ієрархічно структурувати жанрові ознаки.

Метою роботи є розглянути основні метрики параметризації жанрової класифікації літератури, а також провести практичний експеримент із класифікації наукових робіт предметної галузі "штучний інтелект". Наукова новизна статті полягає у розробці та застосуванні комплексної параметризації жанрів літератури на основі чітко визначених метрик, що дозволяє використовувати великі мовні моделі для автоматизованої жанрової класифікації.

У рамках дослідження було проведено практичний експеримент із жанровою класифікацією 10 наукових робіт у галузі штучного інтелекту. Аналіз здійснювався за допомогою GPT-4o та супутніх алгоритмів машинного навчання. Отримані результати підтвердили ефективність параметризації текстів за визначеними метриками та їх використання для автоматизованої класифікації. Було виявлено, що великі мовні моделі демонструють високу точність у визначенні ключових характеристик текстів, проте мають труднощі з розпізнаванням змішаних жанрів та пояснюваністю класифікаційних рішень.

Основними викликами автоматизованої жанрової класифікації є розмитість жанрових меж, вплив навчального корпусу на результати класифікації, необхідність підвищення пояснюваності класифікаційних рішень та адаптація моделей до специфіки різних жанрів. Запропоновано напрямки подальших досліджень, зокрема інтеграцію жанрових онтологій, покращення параметризації текстів і розробку алгоритмів, здатних працювати з багаторівневою жанровою структурою.

Таким чином, дослідження підтвердило перспективність використання великих мовних моделей для автоматизованої класифікації літератури за метриками, проте потребує подальшого вдосконалення алгоритмів та підходів до параметризації текстів.

Ключові слова: жанрова класифікація, великі мовні моделі, параметризація текстів, жанрові онтології, обробка природної мови, автоматизований аналіз текстів.

**Volodymyr Pasichnyk, Maksym Yaromych. GENRE CLASSIFICATION OF LITERATURE
BASED ON METRICS USING LARGE LANGUAGE MODELS**

The paper explores the use of large language models (LLMs) for literary genre classification by employing metric-based parameterization and genre ontologies.

The study examines the theoretical foundations of genre classification, including traditional approaches to defining genres and modern algorithmic methods that leverage large language models. Particular attention is paid to the selection of metrics for text parameterization, including the level of formality, depth of technical analysis, methodological approach, target audience, application domain, type of research data, presence of empirical results, and methods of information visualization. The paper proposes a multi-level classification system that allows for a more precise hierarchical structuring of genre features.

The aim of this study is to examine the main metrics for parametrizing literary genre classification and to conduct a practical experiment on the classification of scientific papers in the field of "artificial intelligence." The scientific novelty of the article lies in the development and application of a comprehensive parametrization of literary genres based on clearly defined metrics, enabling the use of large language models for automated genre classification.

As part of the study, a practical experiment was conducted on genre classification using 10 academic papers in the field of artificial intelligence. The analysis was performed using GPT-4o and associated machine learning algorithms. The results confirmed the effectiveness of text parameterization based on predefined metrics and their application for automated classification. It was found that large language models exhibit high accuracy in identifying

key textual characteristics but struggle with recognizing hybrid genres and providing explainable classification decisions.

The primary challenges of automated genre classification include blurred genre boundaries, the influence of training data on classification outcomes, the need to enhance the explainability of classification decisions, and the adaptation of models to the specifics of different genres. The paper suggests directions for further research, such as the integration of genre ontologies, improvements in text parameterization, and the development of algorithms capable of handling multi-level genre structures.

Thus, this study confirms the potential of large language models for automated literary genre classification based on text metrics. However, further refinement of classification algorithms and approaches to text parameterization is required to achieve higher accuracy and reliability.

Key words: genre classification, large language models, text parameterization, genre ontologies, natural language processing, automated text analysis.

Постановка проблеми. Класифікація літературних жанрів є одним із основних завдань як для класичного літературознавства, так і для сучасних методів обробки природної мови та штучного інтелекту. Раніше цей процес базувався на експертному аналізі, що враховував стилістичні, лексичні та структурні особливості текстів. Однак із розширенням цифрових текстових архівів та необхідністю автоматизованої обробки виникла потреба у розробці нових підходів, здатних ефективно визначати жанрову приналежність творів без втручання людини.

Сучасні технології, зокрема великі мовні моделі, відкривають нові можливості для автоматизованої класифікації текстів за жанрами, так як вони дозволяють аналізувати великі корпуси текстів, враховуючи лексичні, синтаксичні та семантичні характеристики. Проте існують також і значні виклики у застосуванні великих мовних моделей до жанрової класифікації. Основними проблемами є розмитість жанрових меж, недостатня пояснюваність класифікаційних рішень, а також вплив упередженості навчальних даних на точність результатів. Крім того, існуючі підходи не завжди враховують складні ієрархічні структури жанрів та їхні перехідні форми, що ускладнює класифікацію творів з гібридними жанровими характеристиками.

Ще одним важливим аспектом є вибір метрик для параметризації жанрів. Традиційні методи класифікації текстів часто зосереджуються лише на поверхневих характеристиках, таких як ключові слова чи стилістичні ознаки. Водночас сучасні підходи передбачають використання багатофакторних моделей, що враховують формальні, тематичні та структурні параметри. Запровадження жанрових онтологій дозволяє створювати більш точні та структуровані моделі класифікації, які можуть інтегруватися з великими мовними моделями для підвищення ефективності аналізу.

Практична потреба у розробці автоматизованих систем жанрової класифікації також обумовлена необхідністю оптимізації цифрових бібліотек, рекомендаційних сервісів і

наукових репозиторіїв. Вдосконалення методів параметризації та впровадження адаптивних алгоритмів дозволить значно покращити точність класифікації, забезпечити зручний пошук та систематизацію текстів, а також сприяти розвитку міждисциплінарних досліджень у сфері літературознавства, лінгвістики та штучного інтелекту.

Таким чином, проблема автоматизованої жанрової класифікації потребує комплексного підходу, що поєднує сучасні методи обробки природної мови, параметризацію жанрів на основі чітких метрик та інтеграцію жанрових онтологій для забезпечення високої точності аналізу. Дослідження у цій сфері має на меті розробку ефективних алгоритмів, що дозволять вирішити поточні проблеми та покращити якість класифікації літературних текстів.

Мета статті: розглянути основні метрики параметризації жанрової класифікації літератури, а також провести практичний експеримент із класифікації наукових робіт предметної галузі “штучний інтелект”.

Аналіз стану досліджень. Жанрова класифікація літератури є предметом багаторічних досліджень у літературознавстві, лінгвістиці та інформаційних науках. Традиційно жанри визначалися за тематичними, стилістичними та структурними характеристиками, як це розглядається в підручнику [12], однак сучасні підходи дедалі більше схиляються до кількісної параметризації жанрів. Автоматизовані методи, включаючи використання великих мовних моделей, відкривають нові перспективи для точного визначення жанрової належності текстів, що особливо важливо в умовах цифрової інформаційної ери. Важливим є також враховувати сучасні міждисциплінарні підходи. У статті [13] розглянуто взаємозв'язок категорій жанру та тексту, а також трансформація жанрових одиниць.

Одним із важливих аспектів сучасних досліджень є визначення метрик, які можуть бути використані для жанрової класифікації. Лексичні та синтаксичні характеристики тексту відіграють ключову роль у цьому процесі.

Наприклад, у дослідженні [2] розглядається структурно-типологічна параметризація англомовного фентезі, де аналізуються такі показники, як частотність використання певних синтаксичних конструкцій та тематичних лексем. Аналогічні підходи застосовуються і для інших жанрів, включаючи наукові тексти, художню прозу та технічну літературу.

Розвиток методів автоматизованої класифікації текстів значно активізувався завдяки впровадженню алгоритмів машинного навчання та штучного інтелекту. Дослідження [1] показує, що алгоритми кластеризації жанрів можуть ефективно групувати літературні твори за схожими стилістичними та тематичними ознаками, що наближує автоматичну жанрову класифікацію до традиційного експертного аналізу. Важливу роль в автоматизації цієї задачі відіграють великі мовні моделі. У роботі [3] було досліджено можливості LLM у визначенні тематичних структур текстів, що може бути застосовано для жанрової класифікації. Автори показали, що великі мовні моделі здатні ефективно замінити традиційні методи тематичного моделювання, забезпечуючи більш гнучкий підхід до аналізу текстів.

Значну роль у дослідженнях жанрової класифікації відіграють онтології жанрів. Автори статті [8] у своїй роботі розглядають побудову формальних жанрових онтологій, які дозволяють структурувати знання про жанри на формальному рівні та забезпечують інтероперабельність між різними системами жанрової класифікації. У цьому контексті перспективним напрямом є інтеграція LLM із жанровими онтологіями, що дозволяє підвищити точність автоматизованої класифікації та розширити можливості адаптації моделей до різних контекстів. Крім того, у дослідженні [11] продемонстровано можливість автоматизації аналізу літературних оглядів із використанням великих мовних моделей. Це підкреслює їх потенціал у систематизації великих обсягів текстової інформації, що є ключовим для жанрової класифікації.

Окремим напрямом досліджень є оцінка ефективності великих мовних моделей у жанровій класифікації. У роботі [6] проаналізовано точність прогнозів окремих класифікаторів та ансамблевих моделей у завданнях жанрової класифікації та наголошують на необхідності вдосконалення механізмів інтерпретації результатів. Важливим аспектом є також боротьба з упередженнями моделей, що можуть виникати внаслідок нерівномірного представлення жанрів у навчальних корпусах.

Таким чином, сучасні дослідження у сфері жанрової класифікації літератури демон-

струють значний прогрес у напрямку параметризації жанрів та автоматизації процесів класифікації. Однак низка проблем залишається відкритою: точність і пояснюваність рішень великих мовних моделей, адаптація методів до різних мов та жанрових традицій, а також вдосконалення жанрових онтологій. Подальші дослідження у цій сфері можуть зосереджуватися на розробці комбінованих підходів, що поєднують методи обробки природної мови, семантичного аналізу та експертної оцінки для підвищення ефективності жанрової класифікації літературних творів.

Виклад основного матеріалу. Для точного визначення жанру літературного твору використовують різні метрики. Лексичні метрики дозволяють аналізувати частоту використання певних слів або фраз, характерних для конкретного жанру. Наприклад, у науковій фантастиці часто зустрічаються терміни, пов'язані з технологіями та майбутнім, тоді як у детективній літературі домінують слова, що відображають розслідування, загадковість і кримінальні аспекти. Синтаксичні метрики допомагають оцінити складність речень, використання певних граматичних конструкцій, що особливо важливо для аналізу наукових і публіцистичних текстів. Дослідження показують, що наукові статті часто містять складні підрядні речення та термінологічно насичені конструкції, тоді як художня література допускає більшу варіативність у структурі речень [1].

Окрім лексичних і синтаксичних особливостей, важливу роль відіграють стилістичні та структурні метрики. Аналіз художніх засобів, таких як метафори, епітети, символи, допомагає виокремити жанрові особливості тексту. Наприклад, лірична поезія відзначається високою концентрацією метафоричних конструкцій і емоційно насиченою лексикою, тоді як публіцистика тяжіє до раціональності та логічності викладу. Крім того, у різних жанрах спостерігається специфічна структура тексту: наприклад, драматичні твори чітко поділяються на акти та сцени, тоді як класичний роман містить розділи з послідовним розвитком сюжету [2].

Великі мовні моделі, такі як GPT-4 або BERT, навчаються на великих корпусах текстів, що дозволяє їм розпізнавати складні мовні структури та патерни. Для виділення стилістичних прийомів моделі аналізують різні рівні мовних особливостей. На лексичному рівні вони визначають частоту та вибір слів, специфічні терміни або діалектизми. На синтаксичному рівні досліджують структуру речень, наприклад, використання інверсій, повторів

чи інших синтаксичних конструкцій. Семантичний рівень дозволяє виявляти метафори, іронію, сарказм та інші тропи. Для досягнення цього великі мовні моделі використовують механізми самоуваги (self-attention), які дозволяють моделі визначати важливість різних частин тексту в контексті всього документа. Це забезпечує глибоке розуміння тексту та його стилістичних особливостей.

Стилістичні прийоми є ключовими маркерами, що характеризують певний жанр. Наукові статті, наприклад, часто містять формальний стиль та специфічну термінологію, тоді як художня література може включати багатий набір метафор та емоційно забарвлених виразів. Автоматичне виявлення цих прийомів за допомогою великих мовних моделей дозволяє точно класифікувати тексти за жанрами. Дослідження [4] демонструє, як аналіз стилістичних особливостей може бути використаний для ідентифікації жанру медіа-контенту.

Наукові дослідження підтверджують ефективність таких підходів. У роботі [5] розглядається застосування великих мовних моделей для створення текстів, що відповідають певному стилю або жанру, підкреслюючи їх здатність адаптуватися до різних стилістичних вимог. Аналіз різних стилістичних прийомів та їх ролі у формуванні жанрової специфіки текстів також бути корисним для налаштування LLM на розпізнавання таких елементів.

Оцінка точності та адекватності великих мовних моделей у жанровій класифікації. Попри те, що використання великих мовних моделей для автоматичної класифікації текстів за жанрами є перспективним напрямом у галузі обробки природної мови, для забезпечення надійності та точності таких систем необхідно застосовувати відповідні методи оцінки їхньої ефективності, оскільки різні жанри можуть мати подібні характеристики, що ускладнює їх диференціацію.

Загальна точність (Accuracy) є базовою метрикою, що визначає відсоток правильно класифікованих текстів від загальної кількості. Вона дозволяє оцінити загальну продуктивність моделі, однак у випадках нерівномірного розподілу жанрів ця метрика може бути недостатньо інформативною. Наприклад, якщо певний жанр значно переважає в наборі даних, модель може демонструвати високу точність просто завдяки домінуванню одного класу, не обов'язково добре розрізняючи менш поширені жанри.

Окрім загальної точності, важливими є метрики повноти (Recall) та прогностичної точності (Precision). Повнота визначає, наскільки добре модель ідентифікує всі тексти

певного жанру, тоді як прогностична точність показує, який відсоток текстів, класифікованих як певний жанр, дійсно до нього належать. Наприклад, якщо модель має високу повноту, але низьку прогностичну точність, вона може часто відносити тексти до жанру, навіть якщо це хибна класифікація. Натомість висока прогностична точність за низької повноти означає, що модель рідко помиляється, але при цьому пропускає значну частину відповідних текстів. Для балансування цих характеристик використовується оцінка F1 – середнє гармонійне між точністю та повнотою, яке дає узагальнену оцінку ефективності моделі.

Матриця невідповідностей (Confusion Matrix) є важливим інструментом аналізу, що дозволяє побачити, які жанри модель найчастіше плутає між собою. Це особливо корисно для розуміння помилок у класифікації. Наприклад, якщо модель часто плутає наукову фантастику з фентезі, це може свідчити про необхідність додаткового навчання на специфічних мовних конструкціях, що відрізняють ці жанри. Використання матриці невідповідностей дозволяє знаходити закономірності в помилках та коригувати модель відповідно.

Оцінка впевненості (Confidence Estimation) також є критичним аспектом аналізу ефективності жанрової класифікації. Великим мовним моделям властива неоднорідна впевненість у своїх прогнозах: вони можуть бути впевнені в одних жанрах, але припускати значних помилок у визначенні інших. Визначення порогових значень впевненості дозволяє автоматично відсівати менш надійні класифікації або передавати їх на додатковий аналіз. Дослідження [6] підкреслює значення оцінки впевненості для підвищення надійності класифікації жанрів.

Ще одним методом оцінки є перехресна валідація (Cross-validation), яка дозволяє перевірити, наскільки модель стабільна в різних вибірках. Цей підхід розбиває навчальний набір даних на кілька підмножин, використовуючи кожен з них по черзі для тестування та навчання. Це допомагає виявити проблеми з перенавчанням та оцінити, як добре модель здатна узагальнювати закономірності на нових даних.

У практичних дослідженнях застосування цих метрик дозволяє глибше зрозуміти поведінку моделі та її слабкі сторони. Наприклад, аналіз матриці плутанини може виявити, що модель часто плутає наукові статті з науково-популярними, що може свідчити про необхідність додаткового навчання на специфічних особливостях цих жанрів. Дослідження [7] аналізує ефективність різних архітектур

нейромереж для класифікації текстів, підкреслюючи важливість правильної вибірки навчальних даних та метрик оцінки.

Оцінка точності та адекватності моделей є важливою складовою для їхнього подальшого вдосконалення. Використання комплексного підходу, що включає загальну та прогностичну точність, повноту, оцінку F1, матрицю невідповідності, оцінку впевненості та перехресну валідацію, дозволяє досягти більш надійних результатів у класифікації текстів за жанрами. Таким чином, комплексне використання зазначених методів оцінки забезпечує високу точність та адекватність великих мовних моделей у завданнях жанрової класифікації, сприяючи більш ефективній організації та доступу до інформації.

Побудова жанрових онтологій із використанням метрик параметризації. Для оптимізації задачі жанрової класифікації за допомогою LLM доцільно використовувати онтологічний підхід. У сучасному літературознавстві та інформаційних науках онтологія визначається як формалізоване представлення знань про певну предметну область, що є придатним для автоматизованої обробки. Створення жанрових онтологій передбачає систематизацію та структурування літературних жанрів на основі їхніх характеристик, що дозволяє покращити організацію, пошук та аналіз літературних творів. У роботі [8] автори досліджують, як формальна онтологія може бути застосована для класифікації та аналізу різних жанрів документів, що дає змогу розширити її використання на художню та наукову літературу.

Для побудови жанрової онтології необхідно виконати параметризацію за метриками, які можуть включати тематичні елементи, стилістичні прийоми, структурні особливості, цільову аудиторію та інші характеристики. Наприклад, для наукової фантастики характерними метриками можуть бути наявність технологічних інновацій, футуристичних сценаріїв та наукових концепцій, тоді як для романтичної літератури – фокус на емоційних стосунках між персонажами та відповідна лексика. У дослідженні [9] аналізується, як жанрові онтології можуть бути побудовані автоматично на основі великих корпусів текстів і краудсорсингових жанрових міток.

Процес побудови жанрової онтології з використанням метрик параметризації включає декілька етапів. Спочатку дослідники визначають та відбирають релевантні метрики, які найкраще описують особливості кожного жанру. Це може бути досягнуто через аналіз літературних творів, консультації

з експертами та вивчення існуючих досліджень. Далі відібрані атрибути організовуються в ієрархічну структуру, що відображає взаємозв'язки між різними жанрами та піджанрами, дозволяючи створити багаторівневу модель, де загальні категорії розбиваються на більш специфічні. Після цього структуровані дані перетворюються у формальний формат, придатний для автоматизованої обробки, наприклад, використовуючи мови опису онтологій, такі як OWL (Web Ontology Language). Прикладом інтеграції жанрових онтологій у великі цифрові архіви, що полегшує класифікацію літературних творів і автоматизацію їх аналізу, може слугувати онтологія CWRC Genre Ontology [10].

Використання метрик параметризації при побудові жанрових онтологій забезпечує декілька переваг. По-перше, це дозволяє створити більш точні та детальні моделі жанрів, що враховують різноманітність літературних творів. По-друге, така онтологія сприяє покращенню пошукових систем та рекомендаційних сервісів, оскільки забезпечує глибше розуміння змісту та особливостей творів. Нарешті, це відкриває можливості для автоматизованого аналізу та класифікації великих обсягів текстових даних, що є актуальним у сучасному цифровому середовищі.

Таким чином, інтеграція метрик параметризації у процес побудови жанрових онтологій є важливим кроком для розвитку інструментів обробки та аналізу літературних даних, що сприяє більш глибокому розумінню та ефективному використанню літературної спадщини.

Приклад параметризації та визначення жанрів. Спробуємо за допомогою великої мовної моделі GPT-4o провести параметризацію наукової літератури із предметної області «штучний інтелект» та класифікувати її за жанрами. Проаналізуємо 10 наукових робіт:

1) Emmanuel Gillain. *Demystifying Artificial Intelligence*.

2) Peter Norvig. *Paradigms of Artificial Intelligence Programming: CASE STUDIES IN COMMON LISP*.

3) Kyle Gallatin & Chris Albon. *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*.

4) Charlie Gerard. *Practical Machine Learning in JavaScript*.

5) Vijaya Kumar Suda. *Data Labeling in Machine Learning with Python*.

6) Jakub Langr, Vladimir Bok. *Deep learning with Generative Adversarial Networks*.

7) Mario Alemi. *The Amazing Journey of Reason from DNA to Artificial Intelligence*.

Наукова робота	Середня довжина речення	К-сть наукових термінів	К-сть схем	К-сть таблиць	К-сть вставок коду	Рівень формальності	Рівень технічного аналізу
1	3.299	0	1	0	0	неформальний	оглядовий
2	18.577	3	0	0	5	неформальний	оглядовий
3	9.723	28	0	8	11	неформальний	помірний
4	10.844	33	3	4	3	неформальний	помірний
5	33.036	31	0	8	13	високоформальний	глибокий
6	16.882	16	1	1	12	середньоформальний	помірний
7	17.519	13	0	1	0	неформальний	оглядовий
8	10.009	12	0	0	0	неформальний	оглядовий
9	5.743	2	1	0	0	неформальний	оглядовий
10	6.324	6	2	0	1	неформальний	оглядовий

Наукова робота	Методологічний підхід	Цільова аудиторія	Область застосування	Тип даних у дослідженні	Спосіб візуалізації	Наявність емпіричних результатів
1	Оглядова стаття	Широка публіка	Комп'ютерний зір	Невідомий	Мінімальна	Немає
2	Оглядова стаття	Широка публіка	Обробка природної мови	Невідомий	Мінімальна	Немає
3	Експериментальне дослідження	Широка публіка	Глибоке навчання	Невідомий	Кодові блоки	Є
4	Оглядова стаття	Широка публіка	Комп'ютерний зір	Невідомий	Мінімальна	Немає
5	Експериментальне дослідження	Дослідники	Комп'ютерний зір	Невідомий	Кодові блоки	Є
6	Прикладне програмне дослідження	Інженери та студенти	Комп'ютерний зір	Невідомий	Кодові блоки	Немає
7	Оглядова стаття	Широка публіка	Глибоке навчання	Невідомий	Мінімальна	Немає
8	Оглядова стаття	Широка публіка	Комп'ютерний зір	Невідомий	Мінімальна	Немає
9	Оглядова стаття	Широка публіка	Глибоке навчання	Невідомий	Мінімальна	Немає
10	Оглядова стаття	Широка публіка	Комп'ютерний зір	Невідомий	Мінімальна	Немає

Таблиця 1

Результати жанрової класифікації за допомогою GPT-4o

Наукова робота	Жанри
1	Науково-популярні роботи та есе, Оглядові статті
2	Науково-популярні роботи та есе, Оглядові статті
3	Прикладні дослідження, Науково-популярні роботи та есе, Практичні керівництва та документація
4	Науково-популярні роботи та есе, Оглядові статті
5	Прикладні дослідження, Фундаментальні дослідження, Практичні керівництва та документація
6	Практичні керівництва та документація
7	Науково-популярні роботи та есе, Оглядові статті
8	Науково-популярні роботи та есе, Оглядові статті
9	Науково-популярні роботи та есе, Оглядові статті
10	Науково-популярні роботи та есе, Оглядові статті

8) Thomas D. Grant, Damon J. Wischik. On the path to AI.

9) Bernd Carsten Stahl. Artificial Intelligence for a Better Future An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies.

10) Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Christoph Busch. Handbook of Digital Face Manipulation and Detection. Для оцінювання робіт оберемо такі метрики:

1) Рівень формальності (високоформальний, середньоформальний, неформальний) – визначається через аналіз частоти вживання наукових термінів, пасивного стану, складних синтаксичних конструкцій.

2) Рівень технічного аналізу (оглядовий, помірний, глибокий) – визначається за кількістю математичних формул, алгоритмічних описів, кодових блоків у тексті.

3) Методологічний підхід (математичне моделювання, прикладне дослідження, експериментальне дослідження, оглядові статті) – визначається за ключовими словами.

4) Цільова аудиторія (дослідники, інженери, студенти, широка публіка) – визначається через аналіз складності термінів та стилістичних особливостей.

5) Область застосування (комп'ютерний зір, обробка природної мови, глибоке навчання) – визначається за ключовими словами.

6) Тип даних у дослідженні (синтетичні дані, реальні експерименти, теоретичні розрахунки) – визначається через аналіз опису датасетів у тексті.

7) Наявність емпіричних результатів (кількість графіків і таблиць, частота вживання числових даних) – визначається через виявлення розділів, що містять експериментальні результати.

8) Спосіб візуалізації інформації (відсоток графіків/таблиць у тексті) – визначається через аналіз типів візуальних матеріалів.

На основі цих метрик наукові роботи з предметної області «штучний інтелект» будуть віднесені до таких жанрів:

1) Фундаментальні дослідження – високоформальні роботи, орієнтовані на математичне моделювання алгоритмів, оптимізаційні методи, теоретичну фізику обчислення.

2) Прикладні дослідження – праці, присвячені реальним експериментам у ШІ: створення мовних моделей, комп'ютерного зору, застосування нейромереж.

3) Технічні звіти та опис алгоритмів – статті, що детально розглядають реалізацію конкретних методів, архітектур та алгоритмів.

4) Оглядові статті – метааналіз досліджень у певній галузі ШІ, порівняння різних підходів, висновки про тенденції.

5) Науково-популярні роботи та есе – роботи, що пояснюють принципи ШІ для ширшої аудиторії, можуть містити філософські міркування, прогнози майбутнього розвитку.

6) Підручники та навчальні посібники – література, що структурує знання про ШІ в освітньому контексті.

7) Практичні керівництва та документація – опис реалізації алгоритмів, інструкції для роботи з бібліотеками, програмні гайди.

8) Етичні та правові дослідження – роботи, що аналізують вплив ШІ на суспільство, моральні аспекти, питання приватності та регуляції.

Інструментарій. Для автоматичного визначення метрик LLM застосує такі алгоритми, як TF-IDF + LDA (Latent Dirichlet Allocation) (для визначення основних тем у статтях), NER (Named Entity Recognition) (для виявлення ключових термінів для області дослідження), Syntax Tree Parsing (для оцінки рівня формальності та стилістики). Для аналізу графіків та формул будуть застосовані бібліотеки PyMuPDF та pdfplumber. Для групування робіт за жанрами – алгоритм BERT/GPT Embeddings + K-Means.

Після аналізу робіт за допомогою GPT-4o було отримано такі результати (табл. 1).

Із результатів бачимо, що GPT-4o здатний визначити більшість метрик за підсумками аналізу наукових робіт та визначити найбільш відповідні для кожної роботи жанри. Велика мовна модель досить точно порахувала кількість схем, таблиць та блоків коду в роботах, із високою ймовірністю визначила рівні формальності та технічного аналізу, а також методологічний підхід, цільову аудиторію та область застосування досліджень, наведених у роботах. Певні проблеми виникли із обчисленням середньої довжини речення, також виявилось неможливим визначити тип даних у дослідженні, адже більшість робіт є оглядовими статтями, які не обов'язково містять дані, про що свідчать і результати метрики «наявність емпіричних результатів».

У цілому можна констатувати, що жанрова приналежність визначена із високою точністю. Необхідно також врахувати постійне вдосконалення LLM, внаслідок чого точність подібного аналізу буде підвищуватись із новими моделями.

Висновки. Жанрова класифікація літератури є складним багатофакторним процесом, що вимагає як чіткої параметризації текстів, так і ефективних методів їх аналізу. Розвиток вели-

ких мовних моделей відкриває нові можливості для автоматизованої класифікації, дозволяючи застосовувати алгоритмічний підхід до визначення жанрової приналежності текстів. Проте, як показало проведене дослідження, ефективність таких моделей значною мірою залежить від параметризації жанрів, використання метрик та наявності жанрових онтологій.

У межах роботи було розроблено систему параметризації наукових текстів предметної галузі «штучний інтелект», що включала аналіз формальності тексту, глибини технічного дослідження, методологічного підходу, цільової аудиторії, області застосування, типу даних у дослідженні, а також способів візуалізації інформації. На основі цих параметрів було виконано класифікацію текстів за жанрами, що дозволило виявити сильні та слабкі сторони застосування великих мовних моделей у цьому процесі.

Результати експерименту підтвердили, що великі мовні моделі добре розпізнають ключові характеристики текстів та можуть ефективно визначати загальні жанрові категорії. Розгляд методології параметризації текстів засвідчив, що ефективна класифікація потребує багаторівневого підходу, в якому тексти аналізуються за сукупністю формальних, стилістичних і структурних характеристик. Визначені в межах дослідження метрики – рівень формальності, глибина технічного аналізу, методологічний підхід, цільова аудиторія, область застосування, тип дослідницьких даних, наявність емпіричних результатів і спосіб візуалізації інформації – дозволили створити більш об'єктивну базу для класифікації жанрів. Це підтвердило, що навіть для таких складних текстів, як наукові статті, можна розробити ефективну систему жанрового аналізу.

Однак при цьому було виявлено також певні труднощі у класифікації змішаних жанрів, де роботи поєднували риси оглядових досліджень та прикладних експериментів. Це свідчить про необхідність інтеграції багаторівневих жанрових онтологій, які б дозволяли більш точно ієрархічно структурувати жанрові ознаки та враховувати їхню взаємодію. Крім цього, очевидним є вплив навчального корпусу моделі на результати класифікації. Моделі, навчені на даних з переважанням певних жанрів, демонструють зміщення у своїх прогнозах, що підкреслює необхідність розробки збалансованих навчальних корпусів, що містять репрезентативне представлення різних жанрових категорій.

Також, дослідження виявило проблему пояснюваності результатів класифікації. Попри високу точність розпізнавання жанрів, моделі не завжди можуть обґрунтувати свої рішення, що ускладнює їх застосування в науковій та літературознавчій сфері. Покращити розуміння механізмів роботи великих мовних моделей у жанровій класифікації може використання у майбутніх дослідженнях методів Explainable AI (XAI).

Загалом, результати експерименту демонструють, що параметризація жанрів є критично важливим етапом автоматизованої класифікації літературних текстів. Поєднання жанрових онтологій, збалансованих навчальних даних і адаптивних алгоритмів може значно покращити ефективність великих мовних моделей у розпізнаванні жанрових структур. Подальші дослідження в цій сфері можуть зосередитися на оптимізації жанрових метрик, використанні гібридних підходів до класифікації та розробці алгоритмів, здатних працювати з багаторівневою жанровою структурою літературних текстів.

Література:

1. Бехта І. А., Марчук О. В. Структурно-типологічна параметризація художнього текстопростору англomовного фентезі. *Науковий вісник Міжнародного гуманітарного університету*. Сер.: Філологія. 2021. Вип. 47. №. 1. С. 17–21.
2. Бовсунівська Т. В. Теорія літературних жанрів : Жанрова парадигма сучасного зарубіжного роману : Підручник / Т.В. Бовсунівська. К.: Видавничополіграфічний центр «Київський університет». 2009. 519 с.
3. Ворочек О. Г., Соловей І. В. Використання мовних моделей штучного інтелекту для генерації публікацій у соціальних мережах. *Технічна інженерія*. 2024. Вип. 1. №. 93. С. 128–134. DOI: [https://doi.org/10.26642/ten-2024-1\(93\)-128-134](https://doi.org/10.26642/ten-2024-1(93)-128-134).
4. Драненко Г. Теорія літературних жанрів у світлі сучасних міждисциплінарних учень: емпіричний та онтологічний дискурси. *Науковий вісник Східноєвропейського національного університету імені Лесі Українки*. 2015. Вип. 8. С. 49–56.
5. Doulaty M., Saz O., Raymond W. M. Automatic Genre and Show Identification of Broadcast Media. 17th Annual Conference of the International Speech Communication Association, Interspeech. San Francisco, USA. Duration: September 8-12, 2016. P. 2115–2119. <https://doi.org/10.48550/arXiv.1606.03333>
6. Garbacz P. An Outline of a Formal Ontology of Genres. Conference: Knowledge Science, Engineering and Management, First International Conference. Guilin, China. Duration: August 5-8, 2006. P. 151–163. DOI: 10.1007/11811220_14.

7. Lepekhin M., Sharoff S. Estimating Confidence of Predictions of Individual Classifiers and Their Ensembles for the Genre Classification Task. 2022. <https://doi.org/10.48550/arXiv.2206.07427>.
8. Martin C., Hood D. The Use of Natural Language Processing in Literature Reviews. 2024. URL: <https://insights.axtria.com/hubfs/thought-leadership-whitepapers/Axtria-Insights-White-Paper-The-Use-of-Natural-Language-Processing-in-Literature-Reviews.pdf>
9. Mu Y., Dong C., Bontcheva K., Song X. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. 2024. <https://doi.org/10.48550/arXiv.2403.16248>.
10. Peng X. A Comparative Study of Neural Network for Text Classification. 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS). Shenyang, China. Duration: December 11-13, 2020. P. 2115–2119. DOI: <https://doi.org/10.1109/TOCS50858.2020.9339702>.
11. Schreiber H. Genre Ontology Learning: Comparing Curated with Crowd-Sourced Folksonomies. Proceedings of the 17th International Society for Music Information Retrieval Conference. New York City, USA. Duration: August 7-11, 2016. P. 400–406. URL: <https://archives.ismir.net/ismir2016/paper/000074.pdf>.
12. Sobchuk O., Sela A. Computational thematics: Comparing algorithms for clustering the genres of literary fiction. 2023. URL: <https://doi.org/10.48550/arXiv.2305.11251>.
13. The CWRC Genre Ontology Specification 0.7: вебсайт. URL: <https://sparql.cwrc.ca/ontologies/genre.html> (дата звернення: 24.02.2025).