Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

РОЗРОБКА МОДЕЛІ ДЛЯ ВИКОНАННЯ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ В УКРАЇНСЬКОМУ ТЕКСТІ НА ОСНОВІ ПРЕТРЕНОВАНОГО ТРАНСФОРМЕРА GPT-3.5

Кваліфікаційна робота бакалавра
студентки 4 курсу
освітньої програми
«Прикладна (комп'ютерна) лінгвістика та англійська мова»
спеціальності – 035.10 Філологія (прикладна лінгвістика)
галузі знань – 03 гуманітарні науки
Андріани Володимирівни Страп
Науковий керівник:
Микола КОСТІКОВ

«Допущено до захисту»
Протокол засідання
кафедри української мови та прикладної лінгвістики
протокол № 15 **від «** 06 **»** 06 **2024 року**
завідувач кафедри _____(підпис)
к.філол.н., доц. Сергій РІЗНИК

КИЇВ – 2024

АНОТАЦІЯ

У роботі розкрито значення розпізнавання іменованих сутностей (NER) та роль великих мовних моделей для автоматичного тегування частин мови в українському тексті.

Об'єктом дослідження є методи та інструменти розпізнавання іменованих сутностей для автоматизації обробки текстових даних.

Предметом є порівняння ефективності різних інструментів для розпізнавання іменованих сутностей з натренованою моделлю GPT-3.5.

Метою дослідження є розробка моделі для автоматичного тегування частин мови в українському тексті на основі GPT-3.5 та оцінка її ефективності. Завдання включають аналіз наявних інструментів розпізнавання іменованих сутностей, тренування моделі GPT-3.5 на українських текстах та порівняння результатів.

Методологічні підходи базуються на сучасних досягненнях глибокого навчання та трансформерних моделей, що дозволило досягти нових результатів у підвищенні точності та ефективності розпізнавання іменованих сутностей. Новизна дослідження полягає у впровадженні GPT-3.5 для обробки українського тексту, що покращило якість автоматичного тегування.

У першому розділі було розглянуто основні поняття та застосування розпізнавання іменованих сутностей, а також проведено огляд наявних інструментів для розпізнавання іменованих сутностей. Детально проаналізовано поняття великих мовних моделей та їх значення для обробки природної мови. Окрему увагу приділено значенню великих мовних моделей для розпізнавання іменованих сутностей, а також наведеним прикладом використання великих мовних моделей для цієї мети. Розділ завершується аналітичним оглядом напрацювань у сфері розпізнавання іменованих сутностей на основі великих мовних моделей.

У другому розділі розглядається розробка моделі для виконання розпізнавання іменованих сутностей в українському тексті на основі

претренованого трансформера GPT-3. Описано підпункти, які охоплюють етапи розробки моделі, а також метрики, що використовувалися під час тренування моделі.

У третьому розділі розглядається процес тестування розробленої моделі. Описано збір даних для тестування та вивчено проблематику. Розроблено застосунок для порівняльного аналізу, а також детально описано процедуру порівняльного аналізу та оцінки. Представлено результати тестування та експертизи моделі.

Результати дослідження підтверджують, що модель GPT-3.5 демонструє високу точність та ефективність порівняно з іншими інструментами для NER.

Ключові слова: розпізнавання іменованих сутностей, GPT-3.5, Обробка природної мови (NLP), Глибоке навчання, Великі мовні моделі (LLM), SpaCy, Stanza, Flair, XLM-RoBERTa.

ABSTRACT

The paper reveals the importance of named entity recognition (NER) and the role of large language models for automatic named entity recognition in Ukrainian text.

The object of the study is methods and tools for named entity recognition to automate text data processing.

The subject is to compare the efficiency of different named entity recognition tools with the trained GPT-3.5 model.

The purpose of the study is to develop a model for automatic named entity recognition in Ukrainian text based on GPT-3.5 and to evaluate its effectiveness.

The tasks include analyzing existing named entity recognition tools, training the GPT-3.5 model on Ukrainian texts, and comparing the results.

The methodological approaches are based on modern advances in deep learning and transformational models, which has led to new results in improving the accuracy and efficiency of named entity recognition . The novelty of the study lies in the implementation of GPT-3.5 for processing Ukrainian text, which improved the quality of automatic tagging.

The first section discusses the basic concepts and applications of NER, as well as an overview of existing tools for named entity recognition . The concept of large language models and their importance for natural language processing are analyzed in detail. Special attention is paid to the importance of large language models for named entity recognition, as well as examples of how large language models are used for this purpose. The section ends with an analytical review of developments in the field of large language models-based named entity recognition.

The second section discusses the development of a model for performing named entity recognition in Ukrainian text based on the trained GPT-3 transformer. The sub-sections covering the stages of model development are described, as well as the metrics used during model training.

The third section discusses the process of testing the developed model. The data collection for testing is described and the problems are studied. The benchmarking application is developed and the benchmarking and evaluation procedure is described in detail. The results of testing and evaluation of the model are presented.

The results of the study confirm that the GPT-3.5 model demonstrates high accuracy and efficiency compared to other named entity recognition tools.

Keywords: named entity recognition, GPT-3.5, Natural language processing (NLP), Deep learning, Large Language Models (LLM), SpaCy, Stanza, Flair, XLM-RoBERTa.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

LLM- Large Language Models- Великі Мовні Моделі

POS- Part-of-speech tagging - Тегування частин мови

NER- Named Entity Recognition- Розпізнавання іменованих сутностей

API- Application programming interface- Прикладний програмний інтерфейс

NLP- Natural language processing - Обробка природної мови

ВСТУП

Із появою великих мовних моделей (LLM), таких як GPT-3, BERT та їхніх наступників, у галузі обробки природної мови (NLP) відбулася зміна парадигми. Ці моделі, навчені на великих і різноманітних корпусах, продемонстрували значні можливості у вирішенні цілої низки лінгвістичних завдань, перевершуючи ефективність традиційних моделей, розроблених для конкретних застосувань. Одним із таких фундаментальних завдань у NLP є розпізнавання іменованих сутностей (Named Entity Recognition, NER), що передбачає ідентифікацію та класифікацію сутностей у тексті, таких як імена людей, організації, місця тощо.

Історично NER спиралося на статистичні моделі та системи, що базуються на правилах, але поява LLM дає нагоду переоцінити, як ці моделі справляються з цим завданням. У цьому проєкті досліджується ефективність LLM у NER та з'ясовується, чи дають їхні складні архітектури та великі обсяги навчальних даних перевагу над традиційними моделями. На відміну від традиційних NER систем, які часто обмежуються попередньо визначеними правилами і шаблонами, LLM володіють багатими фоновими знаннями і контекстним розумінням, отриманими в результаті навчання на великих наборах текстових даних. Ця багата база знань дозволяє LLM потенційно розуміти нюанси лінгвістичних патернів і контекстуальні тонкощі, які простіші моделі можуть не врахувати.

Крім того, адаптивність LLM до різних лінгвістичних контекстів викликає питання щодо глибини їхнього розуміння мови. Оцінюючи LLM на завданні NER, ми можемо отримати уявлення про їхнє семантичне розуміння та здатність до узагальнення лінгвістичних знань, а також про те, чи їхнє виконання цього завдання свідчить про глибшу, більш узагальнену лінгвістичну компетенцію.

У цьому дослідженні ми прагнемо порівняти продуктивність LLM з традиційними моделями NER, використовуючи стандартні набори даних і

метрики оцінювання. Припускається , що LLM з їхніми вдосконаленими архітектурами та попередньо навченими знаннями перевершать традиційні моделі, демонструючи свій потенціал як більш інтелектуальні та універсальні інструменти для лінгвістичного аналізу. Це дослідження сприятиме нашому розумінню можливостей і обмежень LLM, надаючи більш чітку картину їхньої ролі в майбутньому NLP.

Вивчаючи продуктивність LLM у NER, це дослідження намагається відповісти на ширші питання про їхнє розуміння мови та здібності до узагальнення. Воно також має на меті висвітлити трансформаційний потенціал LLM в комп'ютерній лінгвістиці.

РОЗДІЛ 1. СУЧАСНИЙ СТАН ДОСЛІДЖЕННЯ

1.1. Поняття розпізнавання іменованих сутностей

Розпізнавання іменованих сутностей (NER) – це підзадача обробки природної мови (НЛП), яка ідентифікує та класифікує іменовані об’єкти в тексті за попередньо визначеними категоріями. Ці категорії можуть включати імена осіб, організації, місцезнаходження, час, грошову вартість та інші [3, 2].

NER має широкий спектр застосувань. Це може бути, наприклад, інформаційний пошук. Його задачею є виділення ключових слів з тексту для покращення точності пошуку.Також одним із застосувань може бути аналіз даних. Під час цього процесу стягується інформації з великих обсягів тексту, таких як новинні статті або звіти. Машинний переклад такою є одним із областей застосування NER. Завдяки цьому покращується точність перекладу шляхом збереження ідентичності іменованих сутностей.Ще одна область застосування це чати та віртуальні асистенти. За допомогою NER з’являється можливість розуміння намірів користувачів та надання більш релевантних відповідей[26].

І наостанок видобуток даних. Під час цього процесу автоматично виявляється та класифікується інформації з тексту [29].

Існує два основних підходи до NER.Підхід на основі правил використовує набір вручну визначених правил для ідентифікації та класифікації іменованих сутностей. Статистичний підхід використовує машинне навчання для тренування моделі на наборі даних анотованого тексту[4].

Статистичні підходи, як правило, більш точні, ніж підходи на основі правил, але вони потребують більших наборів даних для тренування.

NER є важливою підзадачею NLP [38] з широким спектром застосувань. Завдяки постійному розвитку технологій NLP точність та ефективність NER-систем постійно зростають.

1.2. Застосування розпізнавання іменованих сутностей

NER має дуже широкий спектр застосувань. Найперше слід згадати пошукові системи. Використання NER дає покращення релевантності результатів пошуку за рахунок кращого розуміння контексту запиту користувача. Також це корисно для виявлення та класифікація інформації про сутності, що може бути використано для створення фрагментів та розширених результатів пошуку. І наостанок, звісно ж персоналізація результатів пошуку на основі історії пошуку та інтересів користувача.

NER також використовується у сфері маркетингу. Першим прикладом є виявлення та сегментація потенційних клієнтів на основі їхньої демографічної інформації, інтересів та поведінки. Також за допомогою NER відбувається створення персоналізованих маркетингових кампаній, які з більшою ймовірністю зацікавлять цільову аудиторію.

Не менш важливим є аналіз відгуків клієнтів для виявлення проблем та покращення продуктів або послуг.

Також є й інші сфери у яких NER відіграє важливу роль.Найперше це обробка медичних текстів.За допомогою цього можна виявити ліки, захворювання та медичні процедури.

У сфері фінансів це в більшій мірі стосується фінансових звітів. Відбувається швидке виявлення компаній, людей та фінансових інструментів. Також NER може класифікувати цілі документи (наприклад, рахунки-фактури, квитанції, паспорти) на різні типи, адаптуючи розпізнавання сутності на основі конкретних характеристик документа.

В сфері юриспруденції використання NER найбільш пов'язане з юридичними документами, особливо це стосується виявлення людей(фізичних чи юридичних собі), організацій, дат та місць.

В сфері науки слід згадати про наукові статті, оскільки NER- тегування допомагає з виявленням дослідників, установ та галузей досліджень [6].

1.3. Огляд наявних інструментів для розпізнавання іменованих сутностей

Stanza – це набір точних та ефективних інструментів для лінгвістичного аналізу багатьох мов. Він починає з сирих текстів, розбиває їх на речення та слова, а потім визначає частини мови та іменовані сутності, виконує синтаксичний аналіз та багато іншого. Stanza пропонує сучасні моделі обробки природної мови для обраної мови. Це офіційна бібліотека Python від групи NLP Стенфордського університету, яка підтримує більше 60 мов та надає доступ до Java-пакету Stanford CoreNLP [32].

NER виконується за допомогою NERProcessor, який викликається за іменем ner. Також Stanza підтримує 23 мови [27] і визначає згадки певних типів сутностей (наприклад, особа, організація) в реченнях.Після виконання пайплайну Stanza результатом є Document, який містить список об’єктів Sentence.Кожене Sentence містить список об’єктів Token.Іменовані сутності можна отримати через властивості entities або ents Document або Sentence.Також теги NER на рівні токенів доступні через поле ner кожного Token[32].

SpaCy – це відкрита бібліотека для розширеного аналізу природної мови (NLP), написана на мовах програмування Python та Cython.SpaCy відзначається високою швидкістю та точністю обробки тексту.Вона розроблена для використання в реальних продуктах . SpaCy справляється з попередньо навченими моделями мов та векторами слів для понад 60 мов.Це дозволяє використовувати готові моделі без необхідності навчання з нуля [11].

Одним з компонентів spaCy є EntityRecognizer (NER), який використовується для розпізнавання та класифікації іменованих сутностей в тексті. Він може ідентифікувати особи, організації, місця та інші сутності [15].

Spacy виконує NER за допомогою Компоненту **EntityRecognizer**, який визначає іменовані сутності в тексті.Він розпізнає мітки ,які не перетинаються на токенах.Результати NER зберігаються в об’єкті **Doc.ents**[36].

Flair – це бібліотека для обробки природної мови (NLP), розроблена Гумбольдтським університетом Берліна та співавторами [16]. Flair дозволяє вам застосовувати сучасні моделі обробки природної мови (NLP) до вашого тексту.Це включає розпізнавання іменованих сутностей (NER), аналіз настроїв, визначення частин мови (PoS), підтримку біомедичних текстів, розрізнення смислів та класифікацію[40].

XLM-RoBERTa – це модель, яка була запропонована в дослідженні “Unsupervised Cross-lingual Representation Learning at Scale”. Вона є розширеною версією моделі RoBERTa, але здатна працювати з багатьма мовами.XLM-RoBERTa навчена на 100 різних мовах [39]. Вона не потребує спеціальних індикаторів мови (наприклад, lang tensors) для розпізнавання мови в тексті. XLM-RoBERTa показує хороші результати на крос-мовних завданнях.Наприклад, в порівнянні з мультилінгвальною моделлю BERT (mBERT), вона досягає +13,8% середньої точності на XNLI, +12,3% середнього F1-показника на MLQA та +2,1% середнього F1-показника на NER.Також XLM-RoBERTa показує високу точність на низько ресурсних мовах[9].

1.4. Поняття великих мовних моделей та їх значення для обробки природної мови

Великі мовні моделі (LLM) [5] відіграють значну роль у сфері обробки природної мови (NLP), зокрема у таких завданнях, як розпізнавання іменованих сутностей (NER), частиномовний розбір (POS) [25, 35] та аналіз сентименту. LLM використовують нейронні мережі з великою кількістю параметрів для розширеного аналізу мови, що дозволяє їм ефективно виявляти складні відносини між сутностями в тексті та генерувати текст, використовуючи семантичні та синтаксичні особливості конкретної мови.

LLM можуть бути особливо корисними для завдань NER, оскільки вони здатні точно класифікувати іменовані сутності в тексті, такі як імена осіб, організацій, локацій тощо. У POS-аналізі [5] LLM допомагають визначати

частини мови для кожного слова в реченні, що є важливим для розуміння структури мови. Щодо аналізу сентименту, LLM можуть виявляти емоційний тон тексту, визначаючи, чи є висловлювання позитивним, негативним чи нейтральним, та навіть виявляти більш тонкі емоції, такі як радість, гнів чи смуток [19].

Застосування LLM у цих областях може значно покращити точність та ефективність обробки мови, забезпечуючи більш глибоке розуміння тексту та його контексту. Це відкриває нові можливості для різноманітних застосувань, від автоматизації відповідей у чат-ботах до аналізу великих обсягів текстових даних для досліджень ринку чи моніторингу бренду.

1.5. Значення великих мовних моделей для розпізнавання іменованих сутностей

Сучасні інструменти для тегування частин мови (POS) та іменованих сутностей (NER) дійсно часто базуються на великих мовних моделях (LLM), таких як BERT. BERT (Bidirectional Encoder Representations from Transformers) є однією з найвідоміших LLM, яка використовується для різноманітних завдань NLP, включаючи POS та NER тегування [19].

BERT та інші подібні моделі, такі як GPT-3 та GPT-3.5, показали високу ефективність у завданнях NER, навіть з обмеженою кількістю тренувальних даних, завдяки своїй здатності до глибокого розуміння контексту та семантики мови. Наприклад, у дослідженні, опублікованому в Journal of the American Medical Informatics Association, було показано, що за допомогою інженерії запитів можна значно покращити продуктивність LLM для клінічного NER[22].

Також, існує підхід GPT-NER, який перетворює завдання послідовного маркування на завдання генерації тексту, що дозволяє LLM адаптуватися до завдань NER [10, 18, 20]. Це демонструє, що LLM можуть бути ефективно адаптовані до завдань, які традиційно вимагали більш спеціалізованих підходів.Використання LLM для POS тегування також є ефективним, оскільки

ці моделі можуть точно визначати частини мови для кожного слова в тексті. Наприклад, проект на GitHub [19] демонструє, як можна налаштувати модель BERT для завдання POS тегування.

Отже, LLM, такі як BERT, відіграють ключову роль у сучасних системах NER та POS тегування, забезпечуючи високу точність та гнучкість для обробки природної мови.

1.6. Приклади використання великих мовних моделей для розпізнавання іменованих сутностей

Великі мовні моделі (LLM) можуть бути використані для завдань розпізнавання іменованих сутностей (NER). Хоча LLM може бути досить ефективною для загальних завдань, таких як створення резюме статей, вона може не мати достатньо спеціалізованих знань для деяких доменно-специфічних завдань[28]. Наприклад, LLM може не мати достатньо медичних знань для точного резюмування складних документів про хірургічні процедури, які містять складні технічні деталі та термінологію. Однак, подальше навчання LLM на медичних даних може навчити її спеціалізованих знань та словникового запасу, необхідних для якісних медичних резюме[28].

Наступним прикладом є те,що LLM може бути використана для тегування NER. LLM може бути додатково налаштована на конкретні завдання, такі як NER, за допомогою техніки Файн тюнінгу. Це може включати навчання моделі на спеціалізованих наборах даних, що допомагає моделі краще розуміти та виконувати конкретні завдання[28].

LLM, такі як GPT, BERT та T5, базуються на трансформаторній архітектурі, яка використовує механізми уваги для розуміння контексту мови 1. Це дозволяє моделям краще зосереджуватися на важливих словах у вхідному реченні та використовувати цю інформацію для генерації виходу [33].

1.7. Аналітичний огляд напрацювання в сфері розпізнавання іменованих сутностей на основі великих мовних моделей

У статті, опублікованій в Journal of the American Medical Informatics Association (JAMIA) [23], автори досліджують потенціал великих мовних моделей (LLMs), зокрема GPT-3.5 та GPT-4, для задач розпізнавання іменованих сутностей (NER) у клінічних текстах. Дослідження спрямоване на покращення точності розпізнавання завдяки інноваційному підходу, який вони називають GPT-NER.

Основною новацією в GPT-NER є використання конструкції підказок (prompts), які включають опис задачі, демонстрації few-shot та вхідне речення. Це дозволяє моделям краще адаптуватися до задачі NER. Демонстрації few-shot слугують прикладами речень з маркованими сутностями, що допомагає моделі краще розпізнавати нові сутності.

У процесі дослідження автори використовували різні стратегії вибору демонстрацій, включаючи випадковий вибір та вибір на основі найближчих сусідів (NN-based retrieval). Останній метод показав кращі результати, оскільки забезпечував семантично близькі приклади.

Результати дослідження показали, що модель GPT-3.5 продемонструвала середнє покращення загальних F1 балів на 0.09, з діапазоном від 0.04 до 0.14. У той час як GPT-4 показала середнє покращення на 0.06, з діапазоном від 0.01 до 0.10. Це свідчить про те, що використання підказок значно підвищує результати NER у порівнянні зі стандартними методами.

Автори також зазначають, що великим мовним моделям притаманні певні обмеження, зокрема потреба у великій кількості пам'яті для зберігання демонстрацій та обмеження на довжину підказок через апаратні обмеження. Незважаючи на це, GPT-NER показує конкурентоспроможні результати у порівнянні з іншими сучасними методами NER, що підтверджує ефективність великих мовних моделей для специфічних завдань обробки природної мови.

Висновок статті підкреслює значний потенціал великих мовних моделей, таких як GPT-3.5 і GPT-4, у вирішенні задач NER у клінічних текстах. Використання підходу GPT-NER дозволяє досягти високої точності завдяки контекстному навчанню та адаптації до специфічних завдань. Саме тому було вирішено тренувати моделі одну з цих моделей, а саме GPT-3.5 для задач NER, оскільки їх здатність до контекстного навчання та адаптації дозволяє досягти високої точності у розпізнаванні іменованих сутностей.

Висновки розділу 1

У цьому розділі було здійснено комплексний аналіз основних аспектів розпізнавання іменованих сутностей (NER) та його ролі у сучасних системах обробки природної мови (NLP). Визначення концепції NER, огляд його застосувань та аналіз наявних інструментів продемонстрували важливість цієї технології для ефективної автоматизації процесу обробки текстової інформації. Зокрема, NER є ключовою технологією, що дозволяє виділяти та класифікувати іменовані сутності, такі як імена, дати, та місця, у текстових даних, що значно підвищує можливості автоматичного структурування неструктурованої інформації.

Великі мовні моделі (LLM), такі як GPT та BERT, стали значним проривом у галузі NLP, забезпечуючи потужні можливості для розуміння та генерації природної мови. Вони значно покращили результати у багатьох NLP-завданнях завдяки своїй здатності навчатися на великих обсягах текстових даних. LLM суттєво підвищили точність та ефективність NER, дозволяючи моделювати складні мовні закономірності та контексти, і, завдяки глибокому навчанню, адаптуватися до різних доменів та мовних особливостей.

РОЗДІЛ 2. РОЗРОБКА МОДЕЛІ ДЛЯ ВИКОНАННЯ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ В УКРАЇНСЬКОМУ ТЕКСТІ НА ОСНОВІ ПРЕТРЕНОВАНОГО ТРАНСФОРМЕРА GPT-3

2.1. Підпункти- етапи розробки

У роботі [1], продовженням якої є цей проєкт, було розроблено тегувальну модель на базі GPT-3.5 за допомогою файн-тюнінгу. Минулорічна модель, як і API ,який був використаний для її файн-тюнінгу стали застарілими, оскільки, GPT стрімко розвивається.Для оптимальної роботи тепер було вирішено працювати з GPT-3.5, яка на даний момент є актуальною. Completions API,згідно з OpenAI [31] , використаний для файн тюнінгу в роботі [1] вже є застарілим, як і ітерація GPT-3 [30], яка раніше використовувалася.Тому потрібно провести міграцію коду для тренування та тренувального датасету до актуального Chat Completions API. Різниця між Completions API та Chat Completions API полягає в їх функціональності та призначеннях .

Completions API зосереджувався на генеруванні завершень тексту, а Chat Completions API спрямований на введення інтерактивних діалогів. Legacy моделі, що використовували Completions API, мали доволі обмежені можливості для взаємодії у діалогах,в той час як сучасніші моделі, які використовують Chat Completions API, мають здатність підтримувати послідовні розмови та розмови з великою кількістю контекстів, забезпечуючи більш ефективну та природну взаємодію з користувачами.

Отже, формат датасету змінився з {"prompt": "<prompt text>", "completion": "<ideal generated text>"} на {"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."}, {"role": "user", "content": "What's the capital of France?"}, {"role": "assistant", "content": "Paris, as if everyone doesn't know that already."}]}, тобто додалося системне повідомлення для моделі.

Як і в минулорічній [1] роботі для файн тюнінгу використано 1/10 частину датасету NER-UK [17, 25] для української мови. Цей датасет є найпоширенішим та найдоступнішим для української мови і був використаний і при тренуванні інших моделей з якими ми проводимо порівняння у цьому проєкті, як от SpaCy NER, Flair NER, XLM-roBERTa NER.

Датасет NER-UK [24] має вигляд директоріїв із файлами .txt та .ann. В перших знаходяться тексти, в других – PIC-анотація до них. NER-UK використовує теги,які наведені на рисунку 2.1:

Тег	Значення
PERS	Персона
LOC	Місце
ORG	Організація
MISC	Різне

Рис. 2.1. Позначки тегів різних іменованих сутностей

Текстові файли мають розмір в середньому 7КБ та приблизно 2000 токенів, якщо застосовувати токенізатор GPT-3.5, що була використана для виконання поставленого завдання, має контекстне вікно в 4096 токенів [34]. У результаті першої спроби було виявлено, що надто велика кількість текстів не поміщаються до контекстного вікна моделі. Отже було вирішено переформатувати датасет до вигляду «<речення> – <теги ІС цього речення>», адже речення з більшою ймовірністю мають меншу за максимальну кількість токенів. Для виконання цього завдання було написано скрипт мовою програмування Python, повна версія коду якого знаходиться знаходиться в гітхаб-репозиторії проєкту [12, 13, 14].

Скрипт працює наступним чином:

1. Виконується імпортування необхідних модулів os та Json.Після цього визначається функція read_annotations, яка приймає аргумент annotation_file і

повертає список анотацій. У цій функції відкривається файл annotation_file для зчитування, і кожен рядок перевіряється на початок з символу «T». Якщо це умова виконується, то рядок розбивається на частини, і отримані значення тегу та сутності додаються до списку анотацій. На кінці функція повертає список анотацій.

2. Функція process_files приймає аргументи text_folder , annotation_folder і output_file. У цій функції відкривається файл output_file у режимі запису, і для кожного файлу у директорії text_folder перевіряється, чи закінчується назва файлу на «.txt». Якщо це умова виконується,то формується шлях до текстового файлу і файлу анотації, використовуючи os.path.join.

3. Відкривається текстовий файл для зчитування, і його вміст розбивається на речення. Для кожного речення формується змінна prompt, яка містить поточне речення і додаткові роздільники.

4. Перевіряється, чи існує файл анотації (annotation_file). Якщо файл існує, то викликається функція read_annotations для отримання списку анотацій з файлу анотації. Потім створюється порожній список sentence-annotations, і для кожної анотації перевіряється, чи міститься сутність в поточному реченні. Якщо умова виконується, то тег та сутність додаються до списку sentence_annotations. Після цього список sentence annotations перетворюється у рядок, або якщо список порожній – у значення «NONE».

5. Формується словник json_line з ключами «prompt» і «completion». Значення ключа «prompt» - це змінна prompt, додана до рядка "\n\n\###\n\n", а значення ключа «completion» - це змінна completion, додана до рядка «END». Потім словник перетворюється у рядок JSON і записується у файл jsonl_file за допомогою json.dumps і jsonl_file.write.

6. За допомогою функції migrate Формат датасету змінюється з {"prompt": "<prompt text>", "completion": "<ideal generated text>"} на {"messages": [{"role": "system", "content": ""}, {"role": "user", "content": ""}, {"role": "assistant", "content": ""}]} для дотримання оновлених вимог до датасету моделі GPT-3.5.

Завдяки чат-орієнтованості моделі GPT-3.5, рядок "\n\n\###\n\n" більше не потрібен, оскільки модель розуміє, що після повідомлення користувача повинна бути відповідь, а в полі system вказано, що від моделі завжди очікується відповідь у вигляді NER тегів і нічого більше.

Текст інструкції: “You are NERtagger, an expert bot designed to perform NER accurately for any language, including but not limited to Ukrainian, English, or a mixture of languages. Always respond only with the tagged entities and nothing else. The tagset you will use includes: LOC, ORG, PERS, MISC and others. The response format is ('entity', 'tag')/newline.”

Було зосереджено увагу на орфографічних помилках та надано інструкцію позначати абсолютно всі сутності, навіть невідомі, з контексту.

Для файн-тюнінгу GPT-3.5 на отриманому датасеті було використано скрипт тренування, повна версія якого знаходиться в гітхаб-репозиторії проєкту [14, 13, 14].

Гіперпараметри, які були використані для файн-тюнінгу моделі, наведені нижче:

1. Розмір Пакета(batch_size):12.Цей параметр вказує на кількість прикладів даних, яка обробляється моделлю за один раз під час навчання.

2. Множник швидкості навчання (learning_rate_multiplier): 16. Цей параметр визначає коефіцієнт, який використовується для множення швидкості навчання базової моделі під час процесу файн-тюнінгу. Він дозволяє масштабувати швидкість навчання на основі конкретних вимог до файн-тюнінгу та характеру набору даних. Для використаного набору даних та поставленого завдання обраний показник показав себе найкраще.

3. Кількість епох (n_epochs): 3. Цей параметр вказує на загальну кількість проходів моделі через весь набір даних під час навчання.

2.2. Метрики тренування

Отже, в результаті було виконано файн-тюнінг моделі GPT-3.5 на отриманому датасеті з PIC-розміткою. Метрики під час тренування були обраховані за допомогою OPENAI API. З отриманих графіків можемо побачити, що на останній тисячі тренувальних кроків втрата моделі [7] складала від 0.00287 до 0.00372.

Рис. 2.2. Графік втрати під час навчання

Висновки до розділу 2

У даному розділі дипломної роботи була описана розробка моделі для автоматичного тегування частин мови в українському тексті на основі попередньо натренованого трансформера GPT-3. Оригінальна робота, якою інспіровано цей проєкт, використовувала тегувальну модель, натреновану на базі GPT-3 з використанням Completions API. Однак у зв'язку з швидким розвитком технологій відкритих штучних нейронних мереж, вирішено оновити модель до GPT-3.5, що є актуальним на сьогоднішній день.

Legacy моделі, які використовували Completions API, мали свої обмеження у взаємодії, зокрема у веденні діалогів, через їхню спрямованість на генерацію завершень тексту. У сучасних моделей, що використовують Chat

Completions API, вдалося підвищити ефективність та природність діалогу завдяки підтримці послідовних розмов та великої кількості контекстів.

Оновлений формат дозволяє моделі краще розуміти контекст та підтримувати натуральну послідовність діалогу, що робить взаємодію з користувачем більш природною та ефективною.

Отже, з урахуванням цих інновацій, оновлення до GPT-3.5 та використання Chat Completions API підвищили якість та функціональність тегувальної моделі, забезпечуючи їй можливість ефективно працювати з українським текстом і відповідати на потреби сучасного інформаційного середовища.

РОЗДІЛ 3. ТЕСТУВАННЯ

Для тестування та порівняння отриманої моделі було вирішено порівнювати її з найбільш поширеними конвенційними засобами автоматичної розмітки частин мови. До списку увійшли: Stanza NER, SpaCy NER,Flair NER та XLM-roBERTa NER.

3.1. Збір даних для тестування

Для наочного тестування якості розпізнавання іменованих сутностей моделями, які входять до порівняння, були обрані тексти, відмінні від «стерильних», з яких складається тренувальний набір даних, але для наочності було взято також і художні тексти українською мовою Для цього було взято речення/тексти , які є у відкритому доступі мережі Інтернет та створені за потрібним запитом штучним інтелектом. До списку входили:

1. Тексти з вкрапленням іншомовних слів (нідерландська, німецька, польська, італійська, іспанська, китайська, іврит, англійська).
2. Тексти з вкрапленням вигаданих слів(слова співзвучні з існуючими в українській мові та повністю вигадані слова)
3. Тексти з одруківками/словами з орфографічними та іншими типами помилок.
4. Художні тексти українською мовою.

Усього було зібрано 203 тексти. Повний файл із текстами знаходиться в Додатку А.

3.2. Проблематика

Проблематика обробки природної мови, зокрема розпізнавання іменованих сутностей (NER), що стосується “нестерильних” текстів є складною та багатогранною. Однією з найважливіших проблем є те,що тексти можуть містити слова з різних мов. Ця багатомовність вимагає від моделей NER

здатності ефективно адаптуватися до різних мовних контекстів та правильно ідентифікувати сутності незалежно від мови.

Крім того, реальні тексти,які зустрічаюся в соцмережах або месенджерах часто містять вигадані слова, які можуть бути використані у специфічних контекстах або як частина сленгу. Це також ускладнює завдання для моделей NER, оскільки такі слова можуть не відповідати жодному з існуючих словників або мовних правил.

Іншою важливою проблемою є одруківки та орфографічні помилки та написання з малої букви ,що часто зустрічається у текстах із соціальних мереж, чатів та електронних листів [28]. Тексти з помилками містять найпоширеніші проблеми, які виникають при письмі(вживання е та и, написання фонетичного варіанту слова або просто замінена 1 буква у слові,що часто трапляється при друківці на екрані телефону).

3.3. Розробка застосунку для порівняльного аналізу

Для зручного порівняння моделей було створено скрипт мовою програмування Python, повна версія якого знаходиться на вебсервісі GitHub [12, 13, 14].

Згаданий скрипт реалізує веб-додаток для розпізнавання іменованих сутностей (NER) в українському тексті за допомогою кількох бібліотек та моделей обробки природної мови. Код використовує Streamlit для створення веб-інтерфейсу, де користувач може вводити текст і отримувати результати аналізу від різних моделей NER.

На початку коду імпортуються необхідні бібліотеки, включаючи Streamlit, Stanza, SpaCy, Flair, transformers та OpenAI. Далі, за допомогою CSS, приховуються елементи інтерфейсу Streamlit, такі як панель інструментів і нижній колонтитул, щоб зробити інтерфейс більш мінімалістичним.

Використовуючи Streamlit secrets, завантажується API-ключ для OpenAI. Щоб зберегти ресурси, моделі завантажуються лише один раз за допомогою

декоратора st.cache_resource. Функція load_models завантажує та ініціалізує моделі для Stanza, SpaCy, Flair і transformers, а також налаштовує pipeline для роботи з моделлю XLM-RoBERTa.

Функції stanza_ner, spacy_ner, flair_ner і gpt_ner виконують аналіз введеного тексту за допомогою відповідних бібліотек і моделей, повертаючи знайдені іменовані сутності та їхні типи. Функція gpt_ner використовує OpenAI API для взаємодії з GPT-3.5, що передбачає асинхронне очікування завершення обробки тексту моделлю.

Інтерфейс Streamlit містить заголовок і текстове поле для введення українського тексту. Кнопка "Analyze" запускає процес аналізу введеного тексту, якщо текстове поле не порожнє. Результати аналізу від різних моделей NER виводяться на екран під відповідними підзаголовками, що дозволяє користувачеві порівняти результати роботи різних моделей.

Таким чином, цей код створює інтерактивний інструмент для розпізнавання іменованих сутностей в українському тексті, використовуючи декілька передових бібліотек та моделей, що забезпечують різноманітні підходи до аналізу тексту.

Кінцевий веб-застосунок [37] зображено на рисунку 3.1.

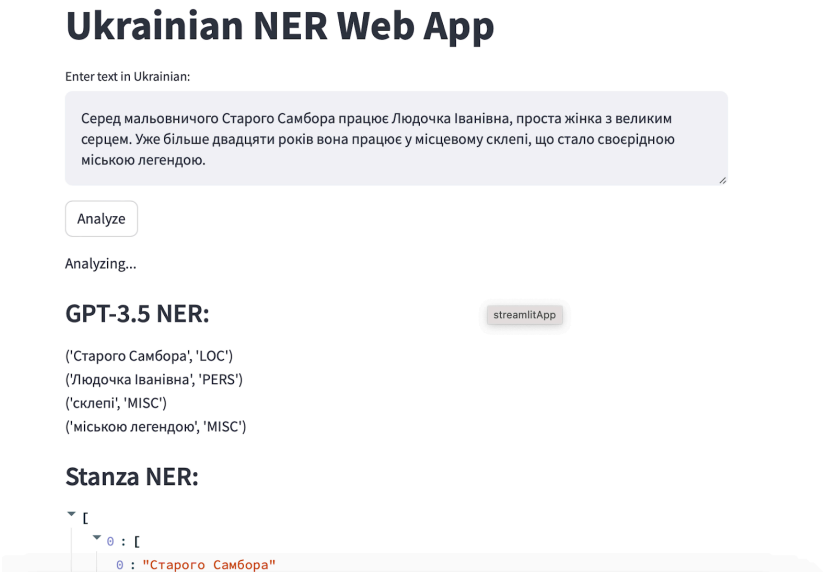

Рис. 3.1 Інтерфейс веб-застосунка

3.4. Опис процедури порівняльного аналізу та оцінки

Усі зібрані тексти було оброблено за допомогою створеної програми-агрегатора. При обробці було використано системне повідомлення відмінне від того, що використовувалося під час файн-тюнінгу. До нього було додано речення “You will provide precise NER tags for all input text, and attempt to tag all entities correctly. Keep an eye out for misspellings and attempt to tag them appropriately based on context. ”, аби спонукати модель до тегування абсолютно всіх сутностей, навіть тих, які у тренувальному датасеті були відсутні при цьому звертаючи увагу на контекст.

Після цього було проведено експертний аналіз отриманих результатів у ручному режимі.

3.5. Тестування та результати експертизи

Було обрано найяскравіші приклади , які показують точність результатів роботи моделей, решта текстів, на яких також проводилось тестування знаходиться у Додатку А.

3.5.1. Вкраплення іншомовних слів в тексті українською мовою.

“Debora вирушила до Markthal, щоб побачити свій geboortehuis.”

Таблиця 3.1

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Debora	PERS			ORG	PER
Markthal	LOC	ORG	ORG	ORG	LOC
geboortehuis	LOC	MISC			

Правильно визначила усі іменовані сутності лише GPT-3.5,XLM-roBERTa NER не змогла розпізнати “geboortehuis”, що повинна мати тег LOC.Інші

моделі показали погані результати, і хоч вони видали результати, але вони не правильні.

“Richard планує поїздку до Bielefeld на наступний місяць.”

Таблиця 3.2

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Richard	PERS				
Bielefeld	LOC	MISC	ORG	ORG	LOC

Найкраще справилась модель GPT-3.5 , а от XLM-roBERTa NER пропустила іменовану сутніть “Richard”(PERS), інші ж моделі хибно визначили “Bielefeld”, оскільки правильним тегом є LOC.

“Jerzy і Kasia зустрілися на PGE Narodowy, щоб обговорити свої плани.”

Таблиця 3.3

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Jerzy	PERS				
Kasia	PERS	ORG	PER	ORG	PER
PGE Narodowy	LOC	ORG			ORG

Лише GPT-3.5 правильно визначив PGE Narodowy як локацію,інші моделі показали неправильні теги,хоча принаймні розпізнали деякі слова польською мовою як іменовані сутності.

“thank god і had наталя семенівна for my current english.”

Таблиця 3.4

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
євробаченні	MISC				
німкеня	PERS				
teardrops	MISC				

Продовження таблиці 3.4

thank god	MISC				
наталя семенівна	PERS	PER			
english	MISC				

GPT-3.5 знову демонструє найкращі результати, правильно визначаючи більшість сутностей. SpaCy змогла ідентифікувати тільки 'наталя семенівна' як персоналію (PER). Інші моделі не дали жодних результатів.

“бен із джетлагу дякую now i'll steal your whole personality”

Таблиця 3.5

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
бен	PERS				

Лише GPT-3.5 змогла розпізнати 'бен' як персоналію (PERS). Інші моделі знову не надали результатів.Тут показано, що іншомовне слово написане з малої ліьери важко розпізнається моделями.

“Минулого тижня я відвідав культурний захід, організований 中国文化中心. Це було неймовірно цікаво! Захід проходив у партнерстві з 北京大学, що додало йому особливої ваги.На початку заходу представники 北京大学 провели лекцію про історію китайської культури. Вони розповіли про давні традиції та звичаї, що існували в різних регіонах Китаю. Після лекції ми мали змогу побачити виступ танцювального колективу 中国舞蹈团, який вразив усіх своєю майстерністю та красою виконання.Однією з найцікавіших частин заходу був майстер-клас з каліграфії, організований 中国书法协会. Під час нього ми навчилися базовим технікам китайської каліграфії та спробували написати кілька ієрогліфів.Завершився захід приємною бесідою за чашкою чаю з представниками 中国文化中心. Ми обговорювали різні аспекти китайської культури та ділилися враженнями від заходу.Цей день залишив у мене незабутні спогади та глибоке враження про багатство і різноманіття китайської культури.

Я дуже вдячний організаторам з 中国文化中心 та 北京大学 за таку чудову можливість ближче познайомитися з культурною спадщиною Китаю.”

Таблиця 3.6

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
中国文化中心	ORG			PERS	ORG
北京大学	ORG	ORG		ORG	ORG
北京大学	ORG	ORG		ORG	ORG
Китаю	LOC	LOC	LOC	LOC	LOC
中国舞蹈团	ORG	ORG		ORG	ORG
中国书法协会	ORG	MISC		ORG	ORG
中国文化中心	ORG			PERS	ORG
Китаю	LOC	LOC	LOC	LOC	LOC
中国文化中心	ORG			PERS	ORG
北京大学	ORG	ORG		ORG	ORG
Китаю	LOC	LOC	LOC	LOC	LOC
Захід			LOC		

GPT-3.5 та XLM-roBERTa показали найкращі результати, правильно ідентифікуючи більшість сутностей. Stanza та SpaCy ідентифікували лише деякі сутності, а Flair зробила помилку, визначивши '中国文化中心' як персоналію (PERS).

“Минулого тижня я мав нагоду відвідати різні міжнародні організації, розташовані в Києві. Спочатку ми зустрілися з представниками United Nations, де обговорили нові ініціативи щодо збереження миру та безпеки в регіоні. Потім ми перейшли до European Union, де обговорювали економічну співпрацю та розвиток інфраструктури.

Після цього ми відвідали представництво World Bank, яке фінансує численні проекти в Україні, зокрема в галузі освіти та охорони здоров'я. Зустріч з представниками International Monetary Fund була присвячена питанням фінансової стабільності та реформування економіки.

Також ми мали цікаву зустріч з представниками UNESCO, де обговорювали важливість збереження культурної спадщини України.

Потім ми відвідали офіс World Health Organization, де обговорювали питання охорони здоров'я та боротьби з епідеміями.На додаток, ми відвідали штаб-квартиру NATO, де обговорювали питання безпеки та співпраці у військовій сфері. Після цього ми мали зустріч з представниками International Red Cross, які розповіли про свої програми гуманітарної допомоги.Завершили ми наш тур в офісі UNICEF, де обговорили усі деталі.”

Таблиця 3.7

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Києві	LOC	LOC	LOC	LOC	LOC
United Nations	ORG	ORG	ORG	ORG	ORG
European Union	ORG	ORG	ORG	ORG	ORG
World Bank	ORG	ORG	ORG	ORG	ORG
Україні	LOC	LOC	LOC	LOC	LOC
International Monetary Fund	ORG	ORG	ORG	ORG	ORG
UNESCO	ORG	ORG	ORG	ORG	ORG
України	LOC	LOC	LOC	LOC	LOC
World Health Organization	ORG	ORG	ORG	ORG	ORG
NATO	ORG	ORG	ORG	ORG	ORG
International Red Cross	ORG	ORG	ORG	ORG	ORG
UNICEF	ORG	ORG	ORG	ORG	ORG

У цьому прикладі всі моделі показали однаково високий рівень точності, успішно розпізнавши всі організації та локації.

“Жив собі в одному лісі маленький зайчик на ім'я Вухань. Він був дуже допитливим і завжди прагнув дізнатися щось нове. Одного дня, прогулюючись лісовою стежкою, він натрапив на дивний камінь із загадковими символами.Зайчик вирішив дізнатися більше про цей камінь і вирушив до мудрого старого сова, який жив у центрі лісу.

Сова, відомий своєю мудрістю, належав до організації "חכמי היער" בשם. Він уважно оглянув камінь і сказав:

- Цей камінь є частиною великої легенди. Він вказує на шлях до прихованого скарбу, який колись заховав у лісі великий маг.

Вухань вирішив піти за підказками і знайти скарб. На своєму шляху він зустрів різних мешканців лісу. Одного разу він натрапив на групу мишей, які належали до клубу "חקרני הטבע".

Миші допомогли зайчику розгадати одну з підказок і разом вони вирушили далі обговорювали питання захисту прав дітей та забезпечення їхнього благополуччя.

Після кількох днів пошуків, Вухань і його нові друзі знайшли старий дуб, під яким був захований скарб. Вони разом викопали його і знайшли не золото чи коштовності, а старовинну книгу знань.

Книга належала до бібліотеки הארגון לשימור מורשת היער і містила багато цікавих фактів про ліс та його мешканців.

Зайчик Вухань був надзвичайно радий своєму відкриттю. Він вирішив поділитися знаннями з усіма мешканцями лісу, організувавши читання в центральній галявині.

Разом з членами "חכמי היער" та "חקרני הטבע" він провів багато цікавих зустрічей, де всі дізнавалися щось нове про свій рідний ліс.

Так маленький зайчик Вухань став відомим дослідником і мудрим радником для всіх мешканців лісу. Він довів, що знання – найцінніший скарб, який можна знайти.”

Таблиця 3.8

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Вухань	PERS	PERS	PER	PERS	PER
совива	PERS	ORG			PER
חכמי היער	ORG	ORG			ORG
Вухань	PERS	PERS	PER	PERS	PER
мишей	PERS			PERS	
חקרני הטבע	ORG	ORG			ORG
Вухань	PERS	PERS	PER	PERS	PER
библиотеки	LOC				
הארגון לשימור מורשת היער	ORG	ORG			ORG
Вухань	PERS	PERS	PER	PERS	PER
חכמי היער	ORG	ORG			ORG
חקרני הטבע	ORG	ORG			ORG
Вухань	PERS	PERS	PERS	PERS	PER

GPT-3.5 та XLM-roBERTa продемонстрували найкращі результати, правильно визначаючи більшість сутностей. Stanza та SpaCy мали проблеми з розпізнаванням деяких сутностей, особливо специфічних для івриту. Flair теж мала обмежені результати, не розпізнавши кілька ключових сутностей.

3.5.2. Результати тестування моделей на текстах з вкрапленнями іншомовних слів

GPT-3.5 та XLM-roBERTa показали найкращі результати в тестах на розпізнавання іменованих сутностей (NER) у текстах, що містять українські, англійські, китайські , івритські та інші іменовані сутності. Ці моделі змогли правильно визначити більшість сутностей у всіх представлених прикладах.

GPT-3.5 показала стабільно високі результати в усіх представлених тестах, розпізнаючи майже всі іменовані сутності різними мовами.Це свідчить про високу якість та надійність моделі.

На основі отриманих даних можна зробити висновок, що розпізнавання іменованих сутностей залишається складним завданням, яке вимагає подальшого вдосконалення моделей та алгоритмів. Використання великих мовних моделей, таких як GPT-3.5 та XLM-roBERTa, демонструє значний потенціал у покращенні точності та якості NER, особливо в багатомовних контекстах.

Слід зауважити, що найкраще себе показали моделі нат розпізнаванні іменованих сутностей англійською мовою. Якщо знову ж таки згадати моделі ,які видавали найкращі результати(а це GPT-3.5 та XLM-roBERTa) можна припустити,що це сталось тому,що Великі мовні моделі можуть мати знання про багато мов завдяки величезним обсягам даних, на яких вони тренувалися. Це дозволяє їм бути більш універсальними у розпізнаванні іменованих сутностей.

3.5.3. Вкраплення вигаданих слів у текстах

“У Греморі живе відомий кулінар Дойніж , який готує дивовижні марциплюшки з фростилічним соусом. Соус в себе різні ягоди, такі як малина, полуниця, чорниця та інші, які змішуються з цукром та іншими інгредієнтами для створення консистенції, що додає текстуру страві.”

Таблиця 3.9

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Греморі	LOC	PERS	LOC	PERS	LOC
Дойніж	PERS	PERS	PER	PERS	PER

Правильні теги видали лише GPT-3.5 та XLM-roBERTa NER, а Stanza та SpaCY помились тегом до “Греморі”, що має мати тег LOC.

“У науковій балонторії Флюмбрика професор Дендорій вивчає властивості нових хімічних сполук, які він називає мікроцитрами.”

Таблиця 3.10

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
науковій балонторії	MISC				
Флюмбрика	ORG	PERS	PER	PERS	ORG
професор Дендорій	PERS	PERS	PER	PERS	PER
хімічних сполук	MISC				
мікроцитрам	MISC				

Найбільше іменованих сутностей розпізнав GPT-3.5, що робить його лідером серед інших моделей, проте “наукова балонторія” це локація, адже це перероблене слово “лабораторія”.

“ У Трибуці Філажанна купила красивий плексофон, щоб грати мелодії для своєї народнильки.”

Таблиця 3.11

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBerTa
Трибуці	LOC	PERS	LOC	PER	LOC

Продовження таблиці 3.11

Філажанна	PERS	PERS	PER	PERS	PER

Доволі добре показали себе GPT-3.5 та XLM-roBERTa NER, але вони не розпізнали іменовану сутність “народнильки” як PERS.Те ,що це персоналія, можна зрозуміти з контексту, проте жодна модель не змогла цього зробити.

“ В рослиняку поблизу Фірелінда місцевий мисливець Вівантер знайшов рідкісного звіра з чудернацьким ім'ям – Гліпсокан.”

Таблиця 3.12

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Фірелінда	LOC	PERS	LOC	PERS	LOC
Вівантер	PERS	PERS	PER	PERS	PER
Гліпсокан	MISC				

Добре справились GPT-3.5 та XLM-roBERTa NER , але Гліпсокан як PERS не розпізнали, також “рослиняку” не було розпізнано як LOC. Це слово було утворено від слова “рослина”.

“Ноксультант в селі дрейковичі працює у гамазині, допомагаючи місцевим жителям з вибором ноксів та консультуючи їх щодо їхніх потреб. він знає кожного клієнта особисто і завжди готовий допомогти з вибором продукції, що найкраще підходить для кожного індивідуального випадку.”

Таблиця 3.13

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
селі дрейковичі	LOC				

Лише GPT-3.5 розпізнала “селі дрейковичі” , але пропустила “ноксультант”(PERS), “у гамазині”(LOC),у яких склади були поміняні місцями. Решта ж моделей взагалі не видали жодних результатів.

“байрактарович вирощує бараболю на своїй плужині. він є досвідченим плужнером і знає всі тонкощі догляду за цією рослиною. його дружина,

вівантівна лагіна, теж активно допомагає йому в господарстві. разом вони вирощують смачні та здорові бараболі, які цінуються серед місцевих мешканців за їхню якість та смакові якості.”

Таблиця 3.14

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
байрактарович	PERS				
вівантівна лагіна	PERS				
плужині		LOC			

Видали результати лише GPT-3.5(2) та Stanza(1) і ці теги є правильними.Інші ж не розпізнали нічого.

“У глибокому лісі Гломірії росла дивовижна рослина на ім'я квітозгуда. Квітозгуда мала високий стеблюр із зеленкуватими блискулями, а її листочки вкривалися мізерянськими краплинками, які світилися у темряві.Легенди Гломірії розповідали, що квітозгуда виростає лише раз на тисячу років під час місячного зіллепіння. Її квітки, звані жупралістами, розкривалися лише вночі й випромінювали дивовижний аромат, що привертав жоруніальних комах.Одного разу, дослідник на ім'я Зумбрагаль вирушив у ліс Гломірії, щоб знайти квітозгуду. Він чув про її магічні властивості і хотів використати її силу для зцілення рідкісної хвороби, яка охопила його село. Зумбрагаль подорожував крізь густі хащі та перепливав клокоплинні річки, поки не дійшов до серця Гломірії. Там, серед величезних дерзолив і древніх кронизмів, він побачив квітозгуду. Її жупралісти світилися, як небесні зорі, а аромат був настільки сильним, що Зумбрагаль відчув, як його охоплює чудесний спокій.Зумбрагаль обережно зібрав кілька жупралістів і повернувся додому. Він створив з них лікарське зілля, яке негайно допомогло хворим одужати. Жителі села були вдячні й прославляли Зумбрагаля як великого героя.Згодом квітозгуда стала

символом надії та зцілення. Люди з різних куточків світу приходили до Гломірії, щоб побачити цю дивовижну рослину. Вони приносили з собою різні дарунки, аби вшанувати її магічну силу.Ліс Гломірії, завдяки квітозгуді, перетворився на місце паломництва. Місцеві жителі, відомі як гломіри, стали хранителями цієї священної рослини. Вони навчали нових поколінь про важливість збереження квітозгуди та її чарівних жупралістів.Квітозгуда продовжувала рости і цвісти, даруючи свої жупралісти світові. Її мізерянські краплинки залишалися загадкою для багатьох дослідників, які намагалися розгадати таємницю її дивовижної краси та магії. Легенди про квітозгуду передавалися з покоління в покоління, і вона залишалася символом надії і зцілення для всіх, хто вірив у її чудесну силу.”

Таблиця 3.15

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Гломірії	LOC	LOC	LOC	LOC	LOC
Зумбрагаль	PERS	PERS	PER	PERS	PER
Гломірії	LOC	LOC	PER	LOC	LOC
Зумбрагаль	PERS	PERS	PER	PERS	PER
Гломірії	LOC	LOC	PER	LOC	LOC
Зумбрагаля	PERS	PERS	PER	PERS	PER
Гломірії	LOC	LOC	LOC	LOC	LOC
гломіри	PERS	LOC	PER		PER

Виходячи з отриманих результатів, Flair NER продемонстрував найкращу продуктивність у ідентифікації об’єктів у тексті, за ним йдуть XLM-roBERTa NER і Stanza NER. GPT-3.5 і SpaCy NER показали меншу точність у цьому конкретному завданні.Тут можливо було розпізнати іменовану сутність лише за контекстом, оскільки слова просто вигадані,а не створені по аналогії.

“Відвідування тренажерної зали завжди викликає в мені відчуття захоплення. Сьогодні я вирішив спробувати нове тренування з використанням

незвичайних тренажерів та вправ, які, як я чув, приносять чудові результати.Першим тренажером, який привернув мою увагу, був гігровелтер. Цей прилад мав великі металеві колеса та гнучкі важелі. Вправа, яку я виконував на ньому, називалася "флеборун". Стоячи на платформі гігровелтера, я тримався за важелі та обережно натискав ногами, змушуючи колеса обертатися. Кожен рух викликав напругу в м'язах ніг і спини, змушуючи їх працювати на повну силу. Я відчував, як кожен м'яз напружується та розслабляється, створюючи гармонійний ритм.Після кількох підходів на гігровелтері, я перейшов до наступного тренажера - міффіглятора. Міффіглятор був схожий на велике коло з ручками по боках. Вправа називалася "крумбітинг". Стоячи всередині кола, я обережно обертав його навколо себе, утримуючи рівновагу. Ця вправа вимагала великої координації та концентрації, оскільки кожен оберт залучав м'язи живота, спини та ніг. Я відчував, як тіло працює разом, створюючи відчуття стабільності та сили.Наступною вправою була "беміфора" на тренажері, відомому як лумінград. Лумінград мав вигляд великого металевого каркаса з мотузками, що звисали з нього. Я встав на платформу та взявся за мотузки, почавши піднімати і опускати своє тіло, ніби виконуючи підтягування, але з додатковою вагою мотузок. Це викликало неймовірне напруження в м'язах рук і грудей, змушуючи їх працювати на межі можливостей.Завершивши з лумінградом, я перейшов до тренажера, який мав назву триломіт. Він складався з великої платформи, яка могла нахилятися в різні боки, та вагових кульок. Вправа називалася "галімбулінг". Я стояв на платформі та обережно нахилявся в різні сторони, утримуючи вагові кульки у рівновазі. Це вимагало значної роботи від м'язів кора та ніг, допомагаючи розвинути баланс та координацію.

На завершення тренування я використав тренажер з назвою фербератор. Він складався з кількох горизонтальних пластин, які рухалися вгору і вниз. Вправа "журфітинг" включала в себе вставання на пластини та виконання присідань. Кожен присід був супроводжений опором пластин, що підсилювало

роботу м'язів ніг та стегон.Завершивши своє незвичайне тренування, я відчував себе виснаженим, але задоволеним. Кожен тренажер і кожна вправа подарували мені нові відчуття та викликали цікавість до подальших занять. Використання гігровелтера, міффіглятора, лумінграда, триломіта та ферберatора дозволило мені відчути, як працюють різні групи м'язів, і подарувало новий погляд на звичні тренування. Я з нетерпінням чекаю на наступний візит до зали, щоб знову спробувати ці незвичайні тренажери та продовжити свій шлях до досягнення фізичної досконалості.”

Таблиця 3.16

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
гігровелтер	MISC				
міффіглятор	MISC			PERS	
крумбітинг	MISC				
лумінград	MISC				
беміфора	MISC				
триломіт	MISC				
галімбулінг	MISC				
фербератор	MISC				
журфітинг	MISC				
Лу					MISC

GPT-3.5 продемонстрував найкращу ефективність у визначенні токенів як "MISC", правильно ідентифікуючи всі з них. Моделі Flair NER та XLM-roBERTa NER показали помилки у класифікації токенів, ідентифікуючи деякі з них як "PERS" або "MISC". Усі слова були вигаданими і не мають жодної схожості з словами,які є в українській мові.

“Минулого зондатя я мав нагоду відвідати різні флімфоричні організації, розташовані в Глімбурзі. Спочатку ми зустрілися з представниками Трілпакс, де обговорили нові фурмінії щодо збереження міролосу та безхорби в регіоні.

Потім ми перейшли до Квантирійської Астролії, де обговорювали екномінну співпрацю та розвиток інфрафтрури.

Після цього ми відвідали представництво Банк Гломікс, яке фінансує численні проекти в Україні, зокрема в галузі освітлони та охорони здорв'я. Зустріч з представниками Фінтракс Монетарного Фунду була присвячена питанням фінальної стабільності та реформування економічних тужин.Також ми мали цікаву зустріч з представниками Культурікс, де обговорювали важливість збереження культурної спадощі України. Потім ми відвідали офіс Організації Здоров'я Марвін, де обговорювали питання охорони здоров'я та боротьби з епідеморіями.На додаток, ми відвідали штаб-квартиру Набітон, де обговорювали питання безпекострі та співпраці у військовій сфері. Після цього ми мали зустріч з представниками Флортих Червоного Хреста, які розповіли про свої програми гумінітарної допомоги.Завершили ми наш тур в офісі Унілор, де обговорювали питання захисту прав дітей та забезпечення їхнього благоплугу.”

Таблиця 3.17

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Глімбурзі	LOC	LOC	LOC	LOC	LOC
Трілпакс	ORG	PERS	ORG	PERS	ORG
Квантирійської Астролії	ORG	LOC	LOC	LOC	ORG
Банк Гломікс	ORG	ORG	ORG	ORG	ORG
Україні	LOC	LOC	LOC	LOC	LOC
Фінтракс Монетарного Фунду	ORG	ORG	ORG	ORG	ORG
Культурікс	ORG	PERS	PER	PERS	ORG

Продовження таблиці 3.17

Організації Здоров'я Марвін	ORG	ORG	ORG	ORG	ORG
Набітон	ORG	LOC	PER	LOC	ORG
Флортих Червоного Хреста	ORG	PERS	PER	ORG	ORG
Унілор	ORG	ORG	LOC	PERS	ORG

Усі моделі видали однакову кількість розпізнаних іменованих сутностей, проте кожна при пустилась помилок як от SpaCy розпізнала “Флортих Червоного Хреста” як особу,хоча насправді це організація.

3.5.4. Результати тестування моделей на текстах з вкрапленнями вигаданих слів

Оцінюючи ефективність розпізнавання іменованих об'єктів (NER) різними моделями (GPT-3.5, Stanza, SpaCy, Flair, XLM-roBERTa) на заданих текстах, можна зробити кілька спостережень:

GPT-3.5 і Stanza продемонстрували порівнянну ефективність, правильно ідентифікувавши значну кількість згаданих сутностей, при цьому обидві моделі пропускали слова .SpaCy також правильно визначила іменовані сутності, проте багато пропустила.

Flair правильно ідентифікував менше всього іменованих сутностей і доволі велику кількість сутностей ця модель розпізнала неправильно.

XLM-roBERTa показала одні з найкращих результатів, правильно ідентифікувавши велику кількість іменованих сутностей у останніх двох заданих текстах.

Загалом, XLM-roBERTa показала найнадійнішу роботу, а GPT-3.5 інколи випереджала інші моделі тим,що видавала результати там, де інші не могли нічого розпізнати , але жодна з моделей не була ідеальною в ідентифікації всіх

названих об'єктів,оскільки іменовані сутності були вигаданими словами,яких не існує в українській мові.

На мою думку, найкраще в кількісному та якісному підході справились GPT-3.5 та XLM-roBERTa NER оскільки моделі можуть використовувати контекстні підказки для розпізнавання іменованих сутностей. Навіть якщо слово є вигаданим, його місце у реченні або відношення до інших слів може вказувати на те, що це іменована сутність. Також GPT-3.5 і RoBERTa були треновані на великих обсягах тексту, що дозволяє їм добре розуміти різноманітні шаблони і контексти. Ці моделі можуть узагальнювати свої знання для нових, вигаданих слів, базуючись на схожих структурах і патернах у даних, на яких вони навчалися. Ще моделі можуть використовувати знання з різних мов, щоб краще розпізнавати нові слова. Вигадані слова можуть бути схожі на реальні слова з інших мов, що допомагає моделям їх правильно класифікувати.Подальше навчання з використанням додаткових маркованих даних і точне налаштування може підвищити їхню точність.

3.5.5. Вкраплення слів з помилками у текстах

“Вируніка прийшла до ківнати своєї систри , щоб взяти у неї книгу.”

Таблиця 3.18

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Вируніка	PERS				PER

Отже, справились з завданням лише GPT-3.5 та XLM-roBERTa NER, проте ніхто не розпізнав “ківната “ як LOC та “систра” як PERS.

“Мекола запросив Вєру на вечерю до ристорану "Золота Фабіта".”

Таблиця 3.19

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Мекола	PERS				PER

Продовження таблиці 3.19

Вєра	PERS	PERS	PER	PERS	PER
Золота Фабіта	ORG	MISC			MISC

Тут добре себе показали GPT-3.5 та XLM-roBERTa NER, але ніхто з них не правильно розпізнав “ристоран “Золота Фабіта” як LOC.

“Сиргій навчається у місцевому університеті, який розташований на вулеці Навчальної.”

Таблиця 3.20

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER
Сиргій	PERS				PER
місцевому університеті	ORG				
вулеці Навчальної	LOC	LOC	LOC	LOC	LOC

Усі іменова сутніості правильно розпізнала модель GPT-3.5, трішки гірше справилась XLM-roBERTa NER , оскільки не розпізнала “місцевому університеті”

Дорогий хавтер,

Не дозволяйте генативу взяти верх над вами у вашій тахі. Краще прагніть до щастя і позитеву, а не до разоб та тикикри інших. Кожна льодина заслуговує на повагу та порозуміння.

З пувагою Мекола”

Таблиця 3.21

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa NER

Продовження таблиці 3.21

Дорогий хавтер	PERS				
генативу	MISC				
тахі	MISC				
Мекола	PERS	PERS	PER	PERS	PER

Усі моделі пропустили слово “тахі”, що означало “хаті” і мало мати тег LOC, але загалом найкраще показала себе модель GPT-3.5, яка розпізнала велику кількість іменованих сутностей у цьому тексті.

3.5.6. Результати тестування текстів з вкрапленнями слів з помилками

GPT-3.5 доволі добре ідентифікує осіб, хоча іноді включає додаткові об'єкти, такі як організації та різні категорії. Він також надає більш детальний контекст, ідентифікуючи фрази на кшталт «Дорогий хавтер».

Stanza NER демонструє хороші результати в ідентифікації осіб, але може пропускати певні об'єкти, які фіксують інші моделі.SpaCy NER має деякі неточності, наприклад, неправильно ідентифікує «Питро» як організацію і не розпізнає «Мурійко».Flair NER загалом добре ідентифікує осіб, але не має знань про додатковий контекст .XLM-roBERTa NER стабільно добре ідентифікує осіб і добре розуміє суб'єктів у контексті.

Загалом, XLM-roBERTa має найкращі результати в цьому порівнянні, за нею йдуть GPT-3.5 і Flair NER. SpaCy NER, хоча загалом непоганий, має більше неточностей порівняно з іншими. Stanza NER є багатообіцяючим, але може потребувати подальшого налаштування для підвищення точності.

Обдумуючи, чому знову ж таки хороше розпізнавання було отримано від XLM-roBERTa , слідом за якою йде GPT-3.5,було зроблено припущення , що так сталось, бо моделі можуть використовувати семантичні подібності між словами для розпізнавання іменованих сутностей. Якщо слово з помилкою схоже на відоме слово, яке часто виступає як іменована сутність, модель може класифікувати його відповідним чином.Також це можливо через поняття

“толерантність до шуму”: GPT-3.5 і RoBERTa тренуються на величезних обсягах даних, що часто включають помилки і неточності. Завдяки цьому вони можуть розпізнавати патерни, навіть якщо дані містять помилки, і правильно класифікувати такі слова.

Підсумовуючи, для загальних завдань NER найкраще використовувати GPT-3.5 або XLM-roBERTa через їхню стабільну ефективність у виявленні об'єктів. Для більш нюансованих завдань, які вимагають детального контексту або специфічної категоризації об'єктів, GPT-3.5 може бути кращим завдяки своїй здатності фіксувати додатковий контекст і різні об'єкти.

3.5.7. Тестування “стерильних” художніх текстів українською мовою

“ Вулиці Ніжина наповнюються сумішшю ароматів: кави з кав'ярні “На розі”, свіжих булочок з пекарні тьоті Каті на розі вулиці, та трохи диму від першої сигарети ранку.”

Таблиця 3.22

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Каті	PERS	PERS	PER	LOC	PER

З однією іменованою сутністю справились GPT-3.5, Stanza NER та Spacy NER, XLM-roBERTa,розпізнавши “Каті” як персону. Проте всі моделі не розпізнали “Вулиці Ніжина”(LOC), “кав’ярні “На розі””(LOC) як сутності взагалі.

“Іванич стояв на порозі свого старого дому в Новому Роздолі, вдивляючись у далечінь, де сонце заходило за обрій. Це містечко завжди було для нього місцем, де він відчував себе вдома. Тут кожен куточок дихав спогадами дитинства, і кожен камінь на дорозі розповідав свою історію. Людосічка, його стара подруга, завжди називала його "Іванич" з особливою ніжністю. Вона зараз жила в Берліні, але Іванич завжди відчував її присутність поруч, немов їхні душі були зв'язані невидимою ниткою.

Валентинчик, молодший брат Людосічки, виріс разом з ними. Він завжди був жартівником, і його жарти були настільки дотепними, що навіть

найпохмуріший день ставав яскравішим.Кузя, старий дворовий пес, зараз лежав на ганку, ліниво розтягнувшись на теплій плитці. Він був вірним другом, який знав всі їхні секрети і завжди був поруч у найважчі моменти.”

Таблиця 3.23

Entity	GPT-3.5	Stanza	SpaCy	Flair	XLM-roBERTa
Іванич	PERS	PERS	LOC	MISC	PER
Новому Роздолі	LOC	LOC	LOC	LOC	LOC
Людосічка	PERS			PERS	PER
Берліні	LOC	LOC	LOC	LOC	LOC
Іванич	PERS	PERS		PERS	PER
Людосічки	PERS	PERS	PER	PERS	PER
Валентинчик	PERS			PERS	PER
Кузя	PERS			PERS	PER

В розпізнаванні іменованих сутностей у цьому тексті найкраще себе проявили GPT-3.5 та XLM-roBERTa, правильно розпізнавши усі наявні іменовані сутності. Flair зробила помилку у “Іванич”(скорочена форма по батькові Іванович)- визначила цю сутність як MISC. У Stanza та Spacy виникли труднощі з розпізнаванням “Людосічка” та “Валентинчика”, які є пестливими формами імені Людмила та Валентин. Також Spacy визначила “Іванич” як локацію, тим самим опинившись на останньому місці в тестуванні на даному тексті.

“У тіні Волновахи бійці АТО, серед яких виділяється Козак, готуються до ночі, налаштовуючи свої серця на ритм війни. Коханий Каті знову на війні. ”

Таблиця 3.24

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Каті	PERS	PERS	PER	LOC	PER
Волновахи	LOC	LOC	LOC	LOC	LOC

Продовження таблиці 3.24

АТО	ORG	LOC	ORG	LOC	LOC
Козак	PERS	PERS	PER	PERS	PER

Найменш точний результат було отримано від Flair NER, оскільки “Каті” було розпізнано як локацію. Іменована сутність “АТО” є спірним питанням, оскільки розшифровується як Антитерористична операція. Робим припущення, що моделі , які розпізнали цю іменовану сутність як локацію, “знають” цю абревіатуру з контекстом “зона АТО”.

“Серед мальовничого Старого Самбора працює Людочка Іванівна, проста жінка з великим серцем. Уже більше двадцяти років вона працює у місцевому склепі, що стало своєрідною міською легендою.”

Таблиця 3.25

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Старого Самбора	LOC	LOC	LOC	LOC	LOC
Людочка Іванівна	PERS	PERS	PER	PERS	PER

Отже, на рівні справились всі моделі, але ніхто з них не розпізнав слово “склеп”, яке означає магазин і є запозиченим з польської мови .Це слово часто вживається жителями західних областей України.

3.5.8. Результати тестування “стерильних” художніх текстів

За результатами порівняння роботи різних моделей NER на українських текстах, можна сказати, що GPT-3.5, XLM-roBERTa, Stanza NER та SpaCy NER демонструю досить непогані результати у розпізнаванні іменованих сутностей. Вони здатен розпізнати особи та деякі локації, але має проблеми з ідентифікацією організацій та інших типів сутностей.Flair NER інколи плутає

локацію та персону, тим самим займаючи останнє місце серед інших моделей в тестуванні на “стерильних текстах”.

3.6. GPT-3.5 для обробки природної мови так чи ні?

3.6.1. Переваги GPT-3.5 для обробки природної мови

GPT-3.5 є потужною моделлю для обробки природної мови (NLP), яка володіє численними перевагами у контексті розпізнавання іменованих сутностей та виконання інших лінгвістичних завдань.

Однією з ключових переваг GPT-3.5 є його тренування на великих корпусах текстів, які охоплюють багато мов та жанрів. Це дозволяє моделі володіти широким культурним контекстом і різноманіттям лінгвістичних варіацій, що є важливим для ефективного розпізнавання іменованих сутностей (NER) та інших завдань, які вимагають розуміння мови в різних контекстах.

Додатково, GPT-3.5 володіє величезним обсягом знань, накопичених під час тренування на масштабних корпусах текстів. Ці знання включають синтаксичні структури, семантичні відношення, а також нюанси використання мови в різних контекстах. Це дозволяє моделі не лише точно визначати іменовані сутності, але й розуміти тонкощі мови, що можуть бути важливими для аналізу текстів у реальному світі, таких як соціальні мережі, чати та електронні листи.

Ще однією перевагою є здатність моделі до контекстуального розуміння. GPT-3.5 може адаптуватись до різних лінгвістичних контекстів, враховуючи попередні частини тексту при генерації відповідей. Це робить модель ефективнішою у порівнянні з традиційними методами, що базуються на правилах або статистиці, особливо в складних завданнях NLP, де важлива точність та здатність до узагальнення.

Таким чином, GPT-3.5 представляє собою потужний інструмент для обробки природної мови, який володіє не лише великою кількістю знань, накопичених в ході тренування на масштабних наборах даних, але й здатністю до контекстуального розуміння і адаптації до різних мовних умов. Це робить

його потенційно важливим інструментом для багатьох лінгвістичних завдань у сучасному NLP.

3.6.2. Недоліки GPT-3.5 для обробки природної мови

GPT-3.5 є потужною моделлю для обробки природної мови (NLP), проте має ряд недоліків, які слід враховувати у наукових дослідженнях та практичних застосуваннях.

Одним з головних недоліків GPT-3.5 є залежність від початкового значення випадкових чисел (seed). Це означає, що різні запускання моделі з різними початковими значеннями можуть давати різні результати, що може ускладнювати відтворюваність результатів. У контексті NLP стабільність і відтворюваність є критично важливими, оскільки вони забезпечують надійність і валідність наукових досліджень. Видача різних результатів при однакових запитах може призвести до невизначеності і ускладнити процес аналізу даних.

Ще одним суттєвим недоліком є обмеженість у збереженні контексту. GPT-3.5 може мати труднощі з обробкою дуже довгих текстів або текстів, які вимагають утримання великої кількості контексту. В таких випадках модель може втрачати важливу інформацію або генерувати відповіді, які не повністю враховують попередні частини тексту. Це особливо проблематично для завдань, що потребують довгострокового збереження контексту, таких як аналіз великих документів або проведення діалогів.

Високі вимоги до обчислювальних потужностей є ще одним значним обмеженням GPT-3.5. Модель вимагає значних ресурсів для тренування і використання, що може бути недоступним для багатьох дослідників та організацій. Використання GPT-3.5 у великомасштабних застосуваннях або в реальному часі потребує значних обчислювальних ресурсів, що створює бар'єри для широкого впровадження цієї технології, особливо у менш розвинених або фінансово обмежених контекстах.

Таким чином, незважаючи на свої значні можливості, GPT-3.5 має ряд обмежень, які варто враховувати при її використанні у задачах NLP. Важливо розуміти ці обмеження та враховувати їх при плануванні та проведенні наукових досліджень. Подальші дослідження повинні бути спрямовані на подолання цих обмежень, зокрема шляхом розробки моделей, які менш залежні від початкового значення випадкових чисел, мають покращену здатність до збереження контексту та вимагають менших обчислювальних ресурсів.

Висновки до розділу 3

На підставі аналізу переваг і недоліків моделі GPT-3.5 для обробки природної мови (NLP) можна зробити висновок щодо доцільності її використання у практичних застосуваннях. Модель GPT-3.5 володіє значним потенціалом у сфері NLP завдяки своїм великим знанням, натренованим на масштабних корпусах текстів, та здатністю до контекстуального розуміння мови. Вона ефективно виконує завдання розпізнавання іменованих сутностей та інших лінгвістичних аналізів, що вимагають розуміння мови в різних контекстах. Однак модель має певні обмеження, зокрема недостатню стабільність у відтворюваності результатів і високі обчислювальні вимоги, що ускладнюють її використання у великомасштабних проектах. Загальною рекомендацією є використання GPT-3.5 у ситуаціях, де важливо здійснювати детальний аналіз текстових даних та враховувати широкий спектр мовних варіацій та контекстів, однак необхідно враховувати її обмеження і здійснювати оцінку відповідно до конкретних потреб проекту та ресурсних можливостей.

ВИСНОВКИ

Дипломна робота присвячена вивченню ефективності великих мовних моделей (LLM), зокрема GPT-3.5, у завданні розпізнавання іменованих сутностей (NER). Проведене дослідження підтвердило значний вплив LLM на галузь обробки природної мови (NLP) і виявило їхні переваги над традиційними моделями NER.

У вступі наголошено на зміні парадигми у NLP з появою LLM, які завдяки навчанню на великих корпусах текстів продемонстрували високу здатність до вирішення різноманітних лінгвістичних завдань. Модель GPT-3.5, завдяки своїй розширеній архітектурі, показала здатність ефективно розпізнавати і класифікувати іменовані сутності в тексті, що значно підвищує автоматизацію обробки текстової інформації.

Аналіз у першому розділі показав, що NER є ключовою технологією для структурування неструктурованої інформації, а великі мовні моделі значно підвищують точність і ефективність цього процесу. Розгляд історії розвитку технології та наявних інструментів підтвердив важливість LLM для сучасних NLP-систем.

Другий розділ був присвячений розробці моделі для автоматичного тегування частин мови в українському тексті на основі GPT-3.5. Виявлено, що використання Chat Completions API дозволило моделі краще розуміти контекст і підтримувати природність діалогу, що суттєво покращило функціональність і якість обробки українського тексту.

У третьому розділі оцінено переваги та недоліки використання GPT-3.5 для NLP-завдань. З'ясовано, що модель володіє значним потенціалом завдяки великим знанням та контекстуальному розумінню мови, проте має певні обмеження, зокрема недостатню стабільність результатів і високі обчислювальні вимоги. Рекомендовано використовувати GPT-3.5 у завданнях, що потребують детального аналізу текстових даних та врахування широкого спектру мовних варіацій і контекстів.

Отже, великі мовні моделі, такі як GPT-3.5, демонструють високий рівень контекстуального розуміння і здатність до узагальнення лінгвістичних знань, що робить їх потужними інструментами для лінгвістичного аналізу і NER-завдань. Дослідження підтвердило їхній трансформаційний потенціал у комп'ютерній лінгвістиці, сприяючи розвитку NLP-технологій і глибшому розумінню мови. Водночас для ефективного використання LLM необхідно враховувати їхні обмеження, зокрема обчислювальні вимоги та стабільність результатів, і застосовувати їх відповідно до конкретних потреб проектів та ресурсних можливостей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. А. Страп, К. Кучинський, «Дотренування GPT-3 для виконання лінгвістичних завдань української мови», Курсова робота, Київський нацональний університет імені Тараса Шевченка, Київ, 2023.
2. Розпізнавання іменованих об'єктів (NER) з генеративним ШІ. Advanced Artificial Intelligence API. URL: https://nlpcloud.com/uk/nlp-named-entity-recognition-ner-api.html
3. Що таке розпізнавання іменованих сутностей (NER): визначення, приклади, типи та застосування. Shaip. URL: https://uk.shaip.com/blog/named-entity-recognition-and-its-types/
4. A comprehensive guide to named entity recognition (NER). AI-Powered Engineering Services, LLM Training, Teams | Turing. URL: https://www.turing.com/kb/a-comprehensive-guide-to-named-entity-recognition
5. A machine learning approach to POS tagging - machine learning. SpringerLink. URL: https://link.springer.com/article/10.1023/A:1007673816718
6. Awan A. A. What is named entity recognition (NER)? Methods, use cases, and challenges. Learn Data Science and AI Online | DataCamp. URL: https://www.datacamp.com/blog/what-is-named-entity-recognition-ner
7. Azure OpenAI Service fine-tuning gpt-3.5-turbo - Azure OpenAI. Microsoft Learn: Build skills that open doors in your career. URL: https://learn.microsoft.com/en-us/azure/ai-services/openai/tutorials/fine-tune?tabs=python-new,command-line
8. Brockopp D. Y. What is NLP?. AJN, American Journal of Nursing. 1983. Т. 83, № 7. С. 1012–1014. URL: https://doi.org/10.1097/00000446-198383070-00012
9. Chan B. XLM-RoBERTa: The multilingual alternative for non-english NLP. Medium. URL: https://medium.com/deepset-ai/xlm-roberta-the-multilingual-alternative-for-non-english-nlp-cf0b889ccbbf

10. Cupani M. Advanced NER With GPT-3 and GPT-J. Medium. URL: https://towardsdatascience.com/advanced-ner-with-gpt-3-and-gpt-j-ce43dc6cdb9c
11. Doc · spaCy API Documentation. Doc. URL: https://spacy.io/api/doc
12. Dyplom/app.py at main · andeashkkk/Dyplom. GitHub. URL: https://github.com/andeashkkk/Dyplom/blob/main/app.py
13. Dyplom/dataset_creation_migration.py at main · andeashkkk/Dyplom. GitHub. URL: https://github.com/andeashkkk/Dyplom/blob/main/dataset_creation_migration.py
14. Dyplom/training.py at main · andeashkkk/Dyplom. GitHub. URL: https://github.com/andeashkkk/Dyplom/blob/main/training.py
15. EntityRecognizer · spaCy API Documentation. EntityRecognizer. URL: https://spacy.io/api/entityrecognizer/
16. GitHub - flairNLP/flair: A very simple framework for state-of-the-art Natural Language Processing (NLP). GitHub. URL: https://github.com/flairNLP/flair.
17. GitHub - lang-uk/ner-uk: Ukranian NER annotation project. GitHub. URL: https://github.com/lang-uk/ner-uk.
18. GitHub - ShuheWang1998/GPT-NER. GitHub. URL: https://github.com/ShuheWang1998/GPT-NER.
19. GitHub - soutsios/pos-tagger-bert: BERT fine-tuning for POS tagging task (Keras). GitHub. URL: https://github.com/soutsios/pos-tagger-bert
20. GPT-NER: named entity recognition via large language models. arXiv.org. URL: https://arxiv.org/abs/2304.10428
21. Hussein D. M. E.-D. M. A survey on sentiment analysis challenges. Journal of King Saud University - Engineering Sciences. 2018. Т. 30, № 4. С. 330–338. URL: https://doi.org/10.1016/j.jksues.2016.04.002
22. Improving large language models for clinical named entity recognition via prompt engineering / H. Yan та ін. OUP Academic. URL:

https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocad259/7590607

23. Improving large language models for clinical named entity recognition via prompt engineering / H. Yan та ін. OUP Academic. URL: https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocad259/7590607#437656653
24. Introducing NER-UK 2.0: a rich corpus of named entities for ukrainian. ACL Anthology. URL: https://aclanthology.org/2024.unlp-1.4/
25. Mudadla S. What is parts of speech (POS) tagging natural language processing?In. Medium. URL: https://medium.com/@sujathamudadla1213/what-is-parts-of-speech-pos-tagging-natural-language-processing-in-2b8f4b07b186.
26. Named entity recognition (NER): what it is & how it is used. AIMultiple: High Tech Use Cases & Tools to Grow Your Business. URL: https://research.aimultiple.com/named-entity-recognition/
27. Named entity recognition. Stanza. URL: https://stanfordnlp.github.io/stanza/ner.html
28. Ner and pos when nothing is capitalized. arXiv.org. URL: https://arxiv.org/abs/1903.11222
29. Named entities / ред.: S. Sekine, E. Ranchhod. Amsterdam: John Benjamins Publishing Company, 2009. URL: https://doi.org/10.1075/bct.19
30. OpenAI GPT-3: everything you need to know [updated]. Springboard Blog. URL: https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/
31. Optimizing large language models for openapi code completion. arXiv.org. URL: https://arxiv.org/abs/2405.15729
32. Overview. Stanza. URL: https://stanfordnlp.github.io/stanza/

33. Pambou J. Fine-tuning LLMs for domain specific NLP tasks: techniques and best practices. Bejamas. URL: https://bejamas.io/hub/guides/fine-tuning-llms-for-domain-specific-nlp-tasks
34. Please tell me the maximum number of tokens for gpt-3.5-turbo-1106. OpenAI Developer Forum. URL: https://community.openai.com/t/please-tell-me-the-maximum-number-of-tokens-for-gpt-3-5-turbo-1106/582766
35. POS(Parts-Of-Speech) Tagging in NLP - GeeksforGeeks. GeeksforGeeks. URL: https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/
36. Python | Named Entity Recognition (NER) using spaCy - GeeksforGeeks. GeeksforGeeks. URL: https://www.geeksforgeeks.org/python-named-entity-recognition-ner-using-spacy/
37. URL: https://ner-highlighter-6usvrahuugxur5rim8it8q.streamlit.app
38. What is NLP (natural language processing)? | IBM. IBM - United States. URL: https://www.ibm.com/topics/natural-language-processing.
39. XLM-RoBERTa – PyText documentation. PyText Documentation – PyText documentation. URL: https://pytext.readthedocs.io/en/master/xlm_r.html.
40. Zero-shot named entity recognition with flair. URL: https://rubrix.readthedocs.io/en/stable/tutorials/07-zeroshot_ner.html

Додаток А

Тексти з вкрапленнями іншомовних слів

“Jerzy і Kasia зустрілися на PGE Narodowy, щоб обговорити свої плани.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Jerzy	PERS	-	-	-	PER
Kasia	PERS	ORG	PER	ORG	PER
PGE Narodowy	LOC	ORG	-	ORG	ORG

“Bella дуже любить відвідувати parchi Todi під час літніх канікул.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Jerzy	PERS	-	-	-	rzy (PER)
Kasia	PERS	ORG	PER	ORG	PER
PGE Narodowy	LOC	ORG	-	ORG	ORG
Bella	PERS	-	-	-	PER
parchi Todi	LOC	MISC	ORG	ORG	LOC
Todi	-	-	-	ORG	LOC

“Alejandro працює в Casa Batlló вже більше десяти років.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Jerzy	PERS	-	-	-	PER
Kasia	PERS	ORG	PER	ORG	PER
PGE Narodowy	LOC	ORG	-	ORG	ORG
Bella	PERS	-	-	-	PER
parchi Todi	LOC	MISC	ORG	ORG	LOC

Todi	-	-	-	ORG	LOC
Alejandro	PERS	-	-	ORG	PER
Casa Batlló	ORG	ORG	ORG	ORG	ORG

Тексти з вкрапленнями вигаданих слів

“На вечірці у Лумбаї всі танцювали під ритми жовтокрапів, а Каваліна милувалася світлом кристаломлітів.”

Entity	**GPT-3.5**	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Лумбаї	LOC	LOC	PER	LOC	LOC
Каваліна	PERS	PERS	PER	PERS	PER
жовтокрапів	MISC	-	-	-	-
кристаломлітів	MISC	-	-	-	-

“У Греморі живе відомий кулінар Дойніж , який готує дивовижні марциплюшки з фростилічним соусом.”

Entity	**GPT-3.5**	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Лумбаї	LOC	LOC	PER	LOC	LOC
Каваліна	PERS	PERS	PER	PERS	PER
жовтокрапів	MISC	-	-	-	-
кристаломлітів	MISC	-	-	-	-
Греморі	LOC	PERS	LOC	PERS	LOC
Дойніж	PERS	PERS	PER	PERS	PER

“Діти в школі Келтрон навчалися мистецтву малювання фантастичних луридів, яких навчателька ліні Заніна називала казковими створіннями.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Лумбаї	LOC	LOC	PER	LOC	LOC
Каваліна	PERS	PERS	PER	PERS	PER
жовтокрапів	MISC	-	-	-	-
кристаломлітів	MISC	-	-	-	-
Греморі	LOC	PERS	LOC	PERS	LOC
Дойніж	PERS	PERS	PER	PERS	PER
школі Келтрон	ORG	PERS	PER	PERS	LOC
луридів	MISC	-	-	-	-
Заніна	PERS	PERS	PER	PERS	PER
казковими	MISC	-	-	-	-

“У науковій балонторії Флюмбрика професор Дендорій вивчає властивості нових хімічних сполук, які він називає мікроцитрами.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Лумбаї	LOC	LOC	PER	LOC	LOC
Каваліна	PERS	PERS	PER	PERS	PER
жовтокрапів	MISC	-	-	-	-
кристаломлітів	MISC	-	-	-	-
Греморі	LOC	PERS	LOC	PERS	LOC
Дойніж	PERS	PERS	PER	PERS	PER
школі	ORG	PERS	PER	PERS	LOC

Келтрон					
луридів	MISC	-	-	-	-
Заніна	PERS	PERS	PER	PERS	PER
казковими	MISC	-	-	-	-
науковій балонторії	MISC	-	-	-	ORG
Флюмбрика	ORG	PERS	PER	PERS	ORG
професор Дендорій	PERS	PERS	PER	PERS	PER
хімічних сполук	MISC	-	-	-	-
мікроцитрами	MISC	-	-	-	-

“В Халмідії відкрили новий розважник атракціонів з велетенськими карусями, де Антель вперше спробував піднятися на торопивіт.”

“У Трибуці Філажанна купила красивий плексофон, щоб грати мелодії для своєї народнильки.”

Entity	**GPT-3.5**	**Stanza NER**	**SpaCy NER**	**Flair NER**	**M-roBERTa N**

“На фестивалі у Грантурі всі захоплювалися танцями чаріпобців, які виконували акробатичні трюки під музику еллібідів.”

Entity	**GPT-3.5**	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Грантурі	LOC	LOC	LOC	LOC	LOC

“В навчальнику Міграміда студенти вивчають давню мову флеридіанців, яка має складну систему гліптокодів.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Міграміда	ORG	PERS	PER	PERS	LOC
флеридіанців	MISC	-	-	-	-

“В рослиняку поблизу Фірелінда місцевий мисливець Вівантер знайшов рідкісного звіра з чудернацьким ім'ям – гліпсокан.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Фірелінда	LOC	PERS	LOC	PERS	LOC
Вівантер	PERS	PERS	PER	PERS	PER
гліпсокан	MISC	-	-	-	-

“там складний лор виявляється він не ухилянт, він не підлягає по хорошому призову”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
лор	PERS	-	-	-	-

“тепер бюрократінша необмежена ґуґл світом”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
ґуґл	ORG	-	-	-	-

“У лісі Журмурії, де зростають густі шемельдрові дерева і протікає річка Гіроглатка, жив незвичайний заєць на ім'я Тумбелрік. Він був відомий своєю здатністю знаходити найрідкісніші жмурквітки, які росли тільки в тіньових куточках Гіроглатки.Одного разу, прогулюючись шемельдровими хащами, Тумбелрік натрапив на дивний камінь, вкритий химерними знаками. Він був

дуже зацікавлений і вирішив звернутися до старого мудреця Лургіфорта, який жив на узгір'ї Залімбуру.

- Лургіфорте, що означають ці знаки? - запитав Тумбелрік, показуючи камінь.

Лургіфорт уважно оглянув камінь і відповів:

- Ці знаки є частиною древнього закляття Зумбрія. Легенда говорить, що воно може привести до скарбів Аркамбріола, які зберігаються в глибині Скурмальду.

Тумбелрік вирішив відправитися на пошуки скарбів. На своєму шляху він зустрів багато чудернацьких істот, таких як шипрогласи та муркливці. Вони допомагали йому розгадувати загадки і долати перешкоди.Після багатьох пригод, Тумбелрік дістався до глибин Скурмальду, де знайшов старовинний скринь. Відкривши його, він побачив не золоті монети чи дорогоцінні камені, а дивовижні кристали Хурміфлонду, які мали чарівні властивості. Тумбелрік повернувся до Журмурії, де поділився своїми знахідками з іншими мешканцями лісу. Всі були вражені його мужністю та наполегливістю. Кристали Хурміфлонду допомогли зробити життя в лісі ще кращим, і Тумбелрік став справжнім героєм Журмурії.”

Entity	**GPT-3.5**	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Журмурії	LOC	LOC	LOC	LOC	LOC
Гіроглатка	LOC	LOC	LOC	LOC	LOC
Тумбелрік	PERS	LOC	PER	PERS	PER
Гіроглатки	LOC	PER	PER	LOC	LOC
Лургіфорта	PERS	PERS	PER	PERS	PER
Залімбуру	LOC	LOC	LOC	LOC	LOC
Лургіфорте	PERS	PERS	PER	PERS	PER
Зумбрія	MISC	PERS	PER	PERS	LOC
Аркамбріола	MISC	LOC	PER	PERS	PER
Скурмальду	LOC	LOC	PER	LOC	LOC
Хурміфлонду	MISC	PER	PERS	PERS	LOC

“ноксультант в селі дрейковичі працює у гамазині, допомагаючи місцевим жителям з вибором ноксів та консультуючи їх щодо їхніх потреб. він знає кожного клієнта особисто і завжди готовий допомогти з вибором продукції, що найкраще підходить для кожного індивідуального випадку.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
селі дрейковичі	LOC	-	-	-	-

“байрактарович вирощує бараболю на своїй плужині. він є досвідченим плужнером і знає всі тонкощі догляду за цією рослиною. його дружина, вівантівна лагіна, теж активно допомагає йому в господарстві. разом вони вирощують смачні та здорові бараболі, які цінуються серед місцевих мешканців за їхню якість та смакові якості.”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
байрактарович	PERS	-	-	PER	-
вівантівна лагіна	PERS	-	PER	-	-
плужині	-	LOC	-	-	LOC

“Минулого зондатя я мав нагоду відвідати різні флімфоричні організації, розташовані в Глімбурзі. Спочатку ми зустрілися з представниками Трілпакс, де обговорили нові фурмінії щодо збереження міролосу та безхорби в регіоні. Потім ми перейшли до Квантирійської Астролії, де обговорювали екномінну співпрацю та розвиток інфрафтрури.Після цього ми відвідали представництво Банк Гломікс, яке фінансує численні проекти в Україні, зокрема в галузі освітлони та охорони здорв'я. Зустріч з представниками Фінтракс Монетарного

Фунду була присвячена питанням фінальної стабільності та реформування економічних тужин.Також ми мали цікаву зустріч з представниками Культурікс, де обговорювали важливість збереження культурної спадощі України. Потім ми відвідали офіс Організації Здоров'я Марвін, де обговорювали питання охорони здоров'я та боротьби з епідеморіями.На додаток, ми відвідали штаб-квартиру Набітон, де обговорювали питання безпекострі та співпраці у військовій сфері. Після цього ми мали зустріч з представниками Флортих Червоного Хреста, які розповіли про свої програми гумінітарної допомоги.Завершили ми наш тур в офісі Унілор, де обговорювали питання захисту прав дітей та забезпечення їхнього благоплугу.”

Entity	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Глімбурзі	LOC	LOC	LOC	LOC
Трілпакс	PERS	ORG	PERS	ORG
Квантирійської Астролії	LOC	LOC	LOC	ORG
Банк Гломікс	ORG	ORG	ORG	ORG
Україні	LOC	LOC	LOC	LOC
Фінтракс Монетарного Фунду	ORG	ORG	ORG	ORG
Культурікс	PERS	PER	PERS	ORG
України	LOC	LOC	LOC	LOC
Організації Здоров'я Марвін	ORG	ORG	ORG	ORG
Набітон	ORG	LOC	LOC	ORG
Флортих Червоного	PERS	PER	ORG	ORG

Хреста				
Унілор	ORG	LOC	PERS	ORG

Тексти зі словами,які містять помилки

“Вируніка прийшла до ківнати своєї систри , щоб взяти у неї книгу.”

Entity	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Вируніка	PER	PER	PER	PER

“Торас зустрівся з дрогом у кав'ярні на вулеці Вешняковій.”

Entity	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Торас	PERS	PER	-	PER
Вешнякові й	PERS	PER	LOC	LOC

“Вуни пішли на екскурсію до полацу в Горобському районі.”

Entity	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Горобському районі	LOC	-	-	LOC

“Улена працює в офесі на пруспекті Сітіліній, де щодня бачить одну й ту ж льодину.”

Entity	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Улена	PERS	-	-	PER
офесі	LOC	-	-	LOC
пруспекті Сітіліній	LOC	LOC	LOC	LOC

“Мекола запросив Вєру на вечерю до ристорану "Золота Фабіта".”

Entity	Stanza NER	SpaCy NER	Flair NER	M-roBERTa N
Мекола	PERS	-	-	PER
Вєра	PERS	PER	PERS	PER
Золота Фабіта	ORG	MISC	MISC	-

“Питрові вирушили на відпочинок до кородту Береговилля.”

Entity	Stanza NER	SpaCy NER	Flair NER	LM-roBERTaE
Питрові	PERS	-	PERS	-
Береговилля	LOC	PER	LOC	LOC

“Сиргій навчається у місцевому університеті, який розташований на вулеці Навчальної.”

Entity	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Сиргій	PERS	-	PERS	PER
місцевому університеті	ORG	-	-	-
вулеці Навчальної	LOC	LOC	LOC	LOC

“Катирина любить гуляти біля озира Вербань, де багато спорцменів.”

Entity	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Катирина	PERS	-	PERS	PERS

Вербань	LOC	LOC	LOC	LOC

“Мутвуій купив квіти на ринку в Львіві, щоб подарувати їх матирі.”

Entity	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa F**
Львіві	LOC	LOC	LOC	LOC

“Брій працює лікарем у клініці на пруспекті Відроджиння, де лікує багато панцієнтів.”

Entity	**GPT-3.5**	**Stanza NER**	**SpaCy NER**	**Flair NER**	**XLM-roBERTa**
Брій	PERS	-	-	PERS	PER
клініці	ORG	-	-	-	-
проспекті Відроджен ня	LOC	ORG	-	-	LOC
пацієнтів	MISC	-	-	-	-

“Дорога Мурійко,

Я давно хотів тобі написати цього лєста, але завжди не вистачало смілості. Тепер, коли нарешті зібрався з думками, хочу висловити свої почуття до тебе.Вже багато часу я відчуваю до тебе щось особливе, що важко описати словами. Твої очі, твої усмішка, і твій голос – все це зачарувало мене з першого ж моменту, як ми познайомілись. Ти єдина, хто може розвеселити мене навіть у найгірші дні.Коли ми проводимо час разом, я відчуваю себе найщасливішою людиною на землі. Ти робеш мій світ яскравішим і надихаєш на краще. Я ніколи не зустрічав такої доброї і щируї льодини, як ти.

З кожним днем мої почуття до тебе стають все сильнішими. Я не можу уявити свого життя без тебе. Ти – моє сонце, яке освітлює моє життє, і я хочу провести з тобою кожну мить.Я знаю, що це може бути несподіванкою для тебе, але я не

міг більше тримати це в собі. Я кохаю тебе, Мурійко, всім своїм серцем. Спудіваюся, що ти відчуваєш те ж саме, і ми зможемо разом пободувати наше щасливе майбутнє.

З любов'ю,Питро”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	XLM-roBERTa
Питро	PERS	PERS	ORG	PERS	PER
Мурійко	PERS	PERS	PER	PERS	PER

“Дорогий хавтер,

Побачив ваш лист і вирішив відповісти. Дякую за ваші васло але я не згоден з вашими думками. Людина може писати про речі, яких вона не знає, але це не означає, що вона має рацію. Краще спрямовуйте свою енергію на щось позитивне і корисне. Життя надто коротке, щоб витрачати його на негатив та критику. Можливо, ви зможете змінити свої погляди, якщо спробуїте підійти до справ з іншого боку.Не дозволяйте генативу взяти верх над вами у вашій тахі.

З пувагою Мекола”

Entity	GPT-3.5	Stanza NER	SpaCy NER	Flair NER	M-roBERTa N
Дорогий хавтер	PERS	-	-	PERS	-
генативу	MISC	-	-	-	-
тахі	MISC	-	-	-	-
Мекола	PERS	PERS	PER	PERS	PER