Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему *Метод ідентифікації подій в україномовних текстах засобами обробки природної мови*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент 2 курсу, група КНм-22-1* _____ *Н.С. Домбровський*
Курс, група виконавця — Підпис — Ініціали, прізвище

Керівник: *старший викладач кафедри КН* _____ *Т.К. Скрипник*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* _____ *Р.О. Багрій*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор _____ *О.В. Бармак*
Підпис — Ініціали, прізвище

18 грудня 2023 р.

Хмельницький 2023

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *магістр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

(підпис)
д.т.н., професор О.В. Бармак
« 01 » вересня 2023 року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: *«Метод ідентифікації подій в україномовних текстах засобами обробки природної мови»*

2. Завдання видано студенту *Домбровському Назарію Сергійовичу*
(прізвище, ім'я, по батькові)

3. Керівник роботи *ст. викладач кафедри КН Скрипник Тетяна Казимирівна*
(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від « *15* » *серпня* *2023* р. № *30*

5. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета кваліфікаційної роботи магістра – вирішення задачі автоматизованої ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє за вхідними даними у вигляді україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Для досягнення мети необхідно також виконати дослідження предметної області ідентифікації подій засобами обробки природної мови, спроєктувати інформаційну систему що використовує розроблений метод та створити відповідну програмну реалізацію для дослідження ефективності розробленого методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Реферат

Кваліфікаційна робота магістра розв’язує задачу ідентифікації подій в україномовних текстах засобами обробки природної мови, що дає можливість за вхідними даними у вигляді україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Для досягнення мети використовується дисперсійне оцінювання важливості слів і нейромережева модель Stanza для виокремлення іменованих сутностей та формування словника термінів подій і формування речень з опису події. Також було створено відповідну програмну реалізацію для апробації методу.

Актуальність теми. У сучасному цифровому світі обсяги текстової інформації, включаючи новини, соціальні медіа, блоги, наукові публікації, постійно зростають. Це створює потребу в ефективних інструментах для їх аналізу, особливо для мов, які мають менше ресурсів, таких як українська. У контексті новин та соціальних медіа швидке та точне розпізнавання подій є ключовим для інформування громадськості та прийняття важливих рішень.

Зрозуміле та структуроване виявлення подій дозволяє компаніям та організаціям ефективніше аналізувати тренди, настрої та впливи, що сприяє кращому плануванню та стратегічному прийняттю рішень. Особливо в умовах збільшення кіберзагроз та інформаційних війн, здатність швидко ідентифікувати та аналізувати події в україномовних текстах має важливе значення для національної безпеки та суспільного благополуччя.

Розвиток та застосування методів ідентифікації подій в україномовних текстах є важливим кроком у напрямку розширення можливостей обробки природної мови, забезпечуючи більшу доступність та розуміння цифрового контенту для україномовних користувачів.

Мета і задачі роботи. *Мета кваліфікаційної роботи магістра* – вирішення задачі ідентифікації подій в україномовних текстах засобами обробки природної мови, що дає можливість за вхідними даними у вигляді

україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Також необхідно створити відповідну програмну реалізацію для апробації методу. Для досягнення мети необхідно виконати такі завдання:

– Дослідити предметну область ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Дослідити існуючі методи та засоби ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Створити метод ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Спроєктувати інформаційну систему на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Обрати засоби розробки для спроєктованої архітектури інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Розробити відповідну програмну реалізацію методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Дослідити практичну ефективність застосування методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Об’єкт дослідження – процес визначення обговорюваних подій в україномовних текстах засобами обробки природної мови.

Предмет дослідження – моделі, методи, алгоритми та засоби для визначення подій в україномовних текстах засобами обробки природної мови.

Методи дослідження, що застосовані для вирішення поставлених завдань: використовуються основні положення методів аналізу даних й теорії множин, для реалізації інформаційної системи визначення подій в україномовних текстах засобами обробки природної мови – методології проектування інформаційних систем, а також було використано об’єктно-орієнтований підхід.

Наукова новизна одержаних результатів. Результати виконання кваліфікаційної роботи магістра містять інновації й наукову новизну, зокрема було удосконалено метод ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для поданого україномовного текстового контенту визначити ключові події, що згадуються в тексті на базі перетворення вхідних даних у вигляді тестового текстового допису у вихідні дані у вигляді тексту, що стисло відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі для виокремлення іменованих сутностей та сформованого словника термінів подій для формування речень з опису події.

Створений метод, будучи застосований для ідентифікації подій в україномовних текстах, має інноваційну властивість враховувати не лише ключових дійових осіб, а й обставини події, подаючи в стислому форматі висновок щодо наявних у тексті подій. Результати дослідження можуть опосередковано сприяти підвищенню ефективності процесу швидкої агрегації новин для ЗМІ та інформаційних порталів, що робить процес збору та обробки інформації більш ефективним. Також опосередковано може мати практичне застосування для систем моніторингу та безпеки, де можна виявляти події, які можуть вказувати на можливі загрози або проблеми.

Практичне значення одержаних результатів. Було розроблено інформаційну систему ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового допису перетворити у вихідні дані у вигляді тексту, що найточніше відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій.

Проведені дослідження ефективності розробленого методу ідентифікації подій в україномовних текстах засобами обробки природної мови з використанням розробленої відповідної інформаційної системи. Розроблена інформаційна система автоматизованого визначення подій в україномовних

текстах засобами обробки природної мови є ефективною та допомагає коректно визначати в текстових дописах події та виводити стислий текст користувачеві. Система спроможна коректно ідентифікувати ключові події, виявляючи іменовані сутності, тематичні ключові слова та контекстуальні зв’язки в текстах.

Результати дослідження свідчать, що розроблений метод спроможний працювати із україномовним контентом, ідентифікувати події, про які написано в тестовому текстовому дописі. Окрім того, реалізований застосунок на базі методу ідентифікації подій в україномовних текстах повертає множину ідентифікованих в тексті іменованих сутностей.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові й практичні результати кваліфікаційної роботи опубліковані:

Домбровський Н.С., Скрипник Т.К., Вознюк Л.О. Метод ідентифікації подій в україномовних текстах засобами обробки природної мови. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». Хмельницький, 2023. с. 80-83.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 42 найменувань та 7 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 109 сторінки, з них 82 сторінки основного тексту та 26 сторінок додатків. У роботі наведено 33 рисунків та 10 таблиць.

Ключові слова: визначення подій, визначення ключових термінів, визначення іменованих сутностей, інформаційна система.

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
ІС	Інформаційна система
КН	Комп’ютерні науки
ШІ	Штучний інтелект
МН	Машинне навчання
ІТ	Інформаційні технології
ПП	Програмний продукт
NLP	Natural Language Processing
NER	Named Entity Recognition
ML	Machine Learning

Вступ

Кваліфікаційна робота магістра розв’язує задачу ідентифікації подій в україномовних текстах засобами обробки природної мови, що дає можливість за вхідними даними у вигляді україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Для досягнення мети використовується дисперсійне оцінювання важливості слів і нейромережева модель Stanza для виокремлення іменованих сутностей та формування словника термінів подій і формування речень з опису події. Також було створено відповідну програмну реалізацію для апробації методу.

Актуальність теми. У сучасному цифровому світі обсяги текстової інформації, включаючи новини, соціальні медіа, блоги, наукові публікації, постійно зростають. Це створює потребу в ефективних інструментах для їх аналізу, особливо для мов, які мають менше ресурсів, таких як українська. У контексті новин та соціальних медіа швидке та точне розпізнавання подій є ключовим для інформування громадськості та прийняття важливих рішень.

Зрозуміле та структуроване виявлення подій дозволяє компаніям та організаціям ефективніше аналізувати тренди, настрої та впливи, що сприяє кращому плануванню та стратегічному прийняттю рішень. Особливо в умовах збільшення кіберзагроз та інформаційних війн, здатність швидко ідентифікувати та аналізувати події в україномовних текстах має важливе значення для національної безпеки та суспільного благополуччя.

Розвиток та застосування методів ідентифікації подій в україномовних текстах є важливим кроком у напрямку розширення можливостей обробки природної мови, забезпечуючи більшу доступність та розуміння цифрового контенту для україномовних користувачів.

Мета і задачі роботи. *Мета кваліфікаційної роботи магістра* – вирішення задачі ідентифікації подій в україномовних текстах засобами обробки природної мови, що дає можливість за вхідними даними у вигляді

україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Також необхідно створити відповідну програмну реалізацію для апробації методу. Для досягнення мети необхідно виконати такі завдання:

– Дослідити предметну область ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Дослідити існуючі методи та засоби ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Створити метод ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Спроєктувати інформаційну систему на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Обрати засоби розробки для спроєктованої архітектури інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Розробити відповідну програмну реалізацію методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

– Дослідити практичну ефективність застосування методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Об’єкт дослідження – процес визначення обговорюваних подій в україномовних текстах засобами обробки природної мови.

Предмет дослідження – моделі, методи, алгоритми та засоби для визначення подій в україномовних текстах засобами обробки природної мови.

Методи дослідження, що застосовані для вирішення поставлених завдань: використовуються основні положення методів аналізу даних й теорії множин, для реалізації інформаційної системи визначення подій в україномовних текстах засобами обробки природної мови – методології проектування інформаційних систем, а також було використано об’єктно-орієнтований підхід.

Наукова новизна одержаних результатів. Результати виконання кваліфікаційної роботи магістра містять інновації й наукову новизну, зокрема було удосконалено метод ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для поданого україномовного текстового контенту визначити ключові події, що згадуються в тексті на базі перетворення вхідних даних у вигляді тестового текстового допису у вихідні дані у вигляді тексту, що стисло відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі для виокремлення іменованих сутностей та сформованого словника термінів подій для формування речень з опису події.

Створений метод, будучи застосований для ідентифікації подій в україномовних текстах, має інноваційну властивість враховувати не лише ключових дійових осіб, а й обставини події, подаючи в стислому форматі висновок щодо наявних у тексті подій. Результати дослідження можуть опосередковано сприяти підвищенню ефективності процесу швидкої агрегації новин для ЗМІ та інформаційних порталів, що робить процес збору та обробки інформації більш ефективним. Також опосередковано може мати практичне застосування для систем моніторингу та безпеки, де можна виявляти події, які можуть вказувати на можливі загрози або проблеми.

Практичне значення одержаних результатів. Було розроблено інформаційну систему ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового допису перетворити у вихідні дані у вигляді тексту, що найточніше відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій.

Проведені дослідження ефективності розробленого методу ідентифікації подій в україномовних текстах засобами обробки природної мови з використанням розробленої відповідної інформаційної системи. Розроблена

інформаційна система автоматизованого визначення подій в україномовних текстах засобами обробки природної мови є ефективною та допомагає коректно визначати в текстових дописах події та виводити стислий текст користувачеві. Система спроможна коректно ідентифікувати ключові події, виявляючи іменовані сутності, тематичні ключові слова та контекстуальні зв’язки в текстах.

Результати дослідження свідчать, що розроблений метод спроможний працювати із україномовним контентом, ідентифікувати події, про які написано в тестовому текстовому дописі. Окрім того, реалізований застосунок на базі методу ідентифікації подій в україномовних текстах повертає множину ідентифікованих в тексті іменованих сутностей.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові й практичні результати кваліфікаційної роботи опубліковані:

Домбровський Н.С., Скрипник Т.К., Вознюк Л.О. Метод ідентифікації подій в україномовних текстах засобами обробки природної мови. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». Хмельницький, 2023. с. 80-83.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 42 найменувань та 7 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 109 сторінки, з них 82 сторінки основного тексту та 26 сторінок додатків. У роботі наведено 33 рисунків та 10 таблиць.

Розділ 1 Дослідження предметної області ідентифікації подій засобами обробки природної мови

1.1 Сучасний стан предметної області обробки природної мови

Аналіз природної мови (Natural Language Processing, або NLP) – це галузь комп’ютерних наук та штучного інтелекту, яка займається розумінням та обробкою людської мови комп'ютерами. Основною метою аналізу природної мови є створення систем, які можуть взаємодіяти з користувачем через природну мову, а також розуміти, інтерпретувати та використовувати інформацію, представлену у текстовій формі [1].

Обробка природної мови (NLP) є однією з найбільш динамічних і значущих галузей в області штучного інтелекту та комп'ютерних наук. Її актуальність полягає у здатності машин ефективно розуміти, інтерпретувати та відтворювати людську мову, що відкриває безліч можливостей для автоматизації, поліпшення комунікацій та аналізу великих обсягів текстових даних [2].

NLP постійно розвивається, вбираючи в себе новітні досягнення в галузі машинного навчання, глибинного навчання та семантичного аналізу. Наукові дослідження та технологічні інновації в цій сфері привели до створення алгоритмів, здатних аналізувати, генерувати та відповідати на людську мову з небувалою точністю [3].

Обробка природної мови (NLP) відіграє ключову роль у сучасному світі, де величезні обсяги інформації потребують розуміння та обробки на швидкості, недоступній для людських можливостей. Ця галузь стала необхідною для багатьох аспектів повсякденного життя, від особистих комунікацій до великих бізнес-операцій.

У сфері обслуговування клієнтів, NLP втілюється в чат-ботах і віртуальних асистентах, які можуть вести природньо звучні розмови з людьми, відповідаючи на запитання та надаючи інформацію. Це не тільки підвищує

ефективність обслуговування, але й забезпечує доступність цих послуг 24/7, розширюючи можливості для підприємств та зручності для споживачів [4].

В освітньому секторі, NLP вносить революційні зміни, дозволяючи створювати індивідуалізовані навчальні програми та автоматизувати оцінювання. Це може бути особливо корисним у мовному навчанні, де NLP-інструменти забезпечують зворотній зв'язок щодо вимови, граматики та семантики, допомагаючи студентам удосконалювати свої навички [5].

У медицині, NLP використовується для аналізу великих обсягів клінічних даних, таких як медичні записи пацієнтів, дослідницькі звіти та наукові статті. Це дозволяє лікарям та дослідникам швидше виявляти тренди, робити точніші діагнози та розробляти ефективніші лікувальні стратегії [6].

У соціальних медіа та цифровому маркетингу, NLP використовується для аналізу настроїв та виявлення трендів серед великих груп користувачів, що допомагає компаніям зрозуміти потреби та вподобання своїх клієнтів, а також вдосконалити свої маркетингові стратегії [7].

У контексті ідентифікації подій в україномовних текстах, аналіз природної мови використовується для розпізнавання та класифікації подій, які відбуваються в текстах. Це включає в себе розуміння контексту, виявлення ключових слів та фраз, визначення відношень між елементами тексту та інші аспекти, що допомагають автоматизувати процес ідентифікації подій у великих обсягах текстової інформації.

Сучасні методи аналізу природної мови включають в себе використання машинного навчання, глибокого навчання та інших технік обробки даних для покращення точності та ефективності ідентифікації подій в текстах українською мовою.

У сучасній галузі ідентифікації подій в україномовних текстах існує кілька ключових викликів та тенденцій, які визначають напрямки подальших досліджень та розвитку цього сегменту.

Мовна різноманітність української мови, включаючи синоніми, діалекти та арго, ускладнює завдання ідентифікації подій [8]. Розробка моделей, які ефективно враховують цю різноманітність, стає важливим завданням.

Також важливий виклик пов'язаний із недостатньо розміченою інформацією для тренування моделей машинного навчання. Україномовні дані часто обмежені та менш розмічені порівняно з іншими мовами, що ускладнює розробку ефективних моделей.

Специфічні характеристики подій, зокрема розрізнення різних типів подій та їхніх контекстів, представляють собою ще один важливий аспект. Однакові слова можуть мати різний смисл у різних контекстах, і точне розрізнення вимагає високої точності моделей.

Не останнім викликом є забезпечення етичної обробки даних враховуючи зростання усвідомленості щодо етичних питань використання даних. Розглядання та вирішення питань конфіденційності та безпеки при обробці текстової інформації стає важливим аспектом роботи в цьому напрямку.

Тенденції включають розвиток глибокого навчання, збільшення уваги до розпізнавання контексту, використання многомовних моделей та збільшення уваги до розробки систем для конкретних завдань, таких як визначення сутностей та класифікація текстів [9].

В результаті виконання аналізу сучасного стану області обробки природної мови було визначено, що NLP є не просто технологією, а однією із найбільш динамічних і значущих галузей в області штучного інтелекту та комп’ютерних наук. Актуальність розвитку NLP полягає у здатності машин ефективно розуміти, інтерпретувати та відтворювати людську мову, що відкриває безліч можливостей для автоматизації, поліпшення комунікацій та аналізу великих обсягів текстових даних

1.2 Аналіз сучасних методів та засобів ідентифікації подій в україномовних текстах засобами обробки природної мови

Методи та засоби інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови є важливою частиною сучасних досліджень у галузі комп’ютерних наук та штучного інтелекту. Основна мета цих методів полягає в автоматичному виявленні та класифікації подій, описаних у текстах, які можуть включати новини, статті, блоги чи навіть соціальні медіа.

У контексті обробки природної мови, подія зазвичай визначається як взаємопов'язана група концептів, яка описує певну ситуацію або подію, що має місце у часі і просторі. Подія складається з ключових компонентів, таких як учасник, дія, час та місце [10].

Ідентифікація подій – це процес виявлення, класифікації та розуміння подій, які відбуваються в текстах або інших джерелах інформації. Ця галузь важлива в області обробки природної мови та штучного інтелекту, оскільки дозволяє системам автоматично аналізувати та визначати події з текстових даних.

Ідентифікація подій може бути класифікована за кількома аспектами [11].

1. *Класифікація за типом подій:*

– Специфічні події: Визначення конкретних подій, таких як зустрічі, конференції, спортивні події тощо.

– Загальні події: Виявлення загальних подій, таких як новини, зміни в економіці, соціокультурні явища.

2. Класифікація за рівнем деталізації:

– Грубий рівень: Визначення обширних категорій подій без докладного уточнення.

– Детальний рівень: Аналіз подій з високою деталізацією та врахуванням контексту.

3. Класифікація за методами аналізу:

– Статистичні методи: Використання статистичних моделей та правил для ідентифікації ключових слів та фраз, що вказують на події.

– Машинне навчання: Застосування алгоритмів машинного навчання для автоматичного виявлення та класифікації подій на основі тренувальних даних.

Ідентифікація подій в текстах може вказувати на актуальність інформації, допомагати в ефективному відслідковуванні подій у реальному часі, а також використовується для створення систем аналізу новин, моніторингу соціальних мереж та інших застосувань [12]. Застосування ідентифікації подій може бути широким, включаючи сфери від журналістики та фінансів до наукових досліджень та військового аналізу.

Різноманітні методи визначення подій в текстах використовують різні стратегії, особливо з використанням обробки природної мови та штучного інтелекту. Методи, засновані на правилах, ґрунтуються на визначених правилах та шаблонах, які враховують ключові слова, фрази та структуру тексту.

Машинне навчання включає класифікацію, де алгоритми, такі як наївний Баєсів класифікатор, метод опорних векторів (SVM) або глибоке навчання, автоматично класифікують тексти на події та не події. Використання нейронних мереж дозволяє виявляти складні взаємозв'язки та контексти для точного визначення подій.

Статистичний аналіз базується на аналізі частоти вживання слів чи фраз у тексті, або використовує метод TF-IDF для визначення важливості слів у тексті, сприяючи виявленню ключових елементів, пов'язаних із подіями [13].

Використання дисперсійної оцінки може бути корисним при ідентифікації подій у текстах. Цей метод можна використовувати для визначення важливості слів у текстових документах, що можуть вказувати на ключові терміни або інформацію, пов'язану з певними подіями чи темами. Розрахунок дисперсійної оцінки може допомогти виділити найбільш значущі слова для подальшого аналізу та ідентифікації подій.

При розробці методу ідентифікації подій в україномовних текстах за допомогою обробки природної мови, важливо вибрати підхід для оцінки

важливості слів та токенів у текстах. У даному випадку, використання дисперсійної оцінки виявляється більш обґрунтованим. Цей метод краще враховує різноманітність слів та токенів у тексті, що особливо важливо для ідентифікації подій.

Оцінка на основі дисперсії, така як BM-25, сприяє уникненню перенавчання моделі. Це означає, що модель не буде занадто вагатися на специфічних словах чи фразах, які можуть бути унікальними для певних текстів, і зосередиться на більш загальних та інформативних частинах тексту, що стосуються подій [14].

Глибоке навчання включає використання попередньо тренованих моделей, таких як BERT або GPT, які навчені на великих обсягах даних та можуть враховувати широкий контекст тексту.

Комбінаційні стратегії, такі як ансамблі, об'єднують кілька моделей або методів для досягнення більшої точності та стійкості результатів. Ці стратегії можуть використовуватися як окремо, так і в комбінації, залежно від конкретного завдання та характеристик текстових даних, що обробляються.

На рисунку 1.1 наведено основні кроки та методи для пошуку подій в текстах.

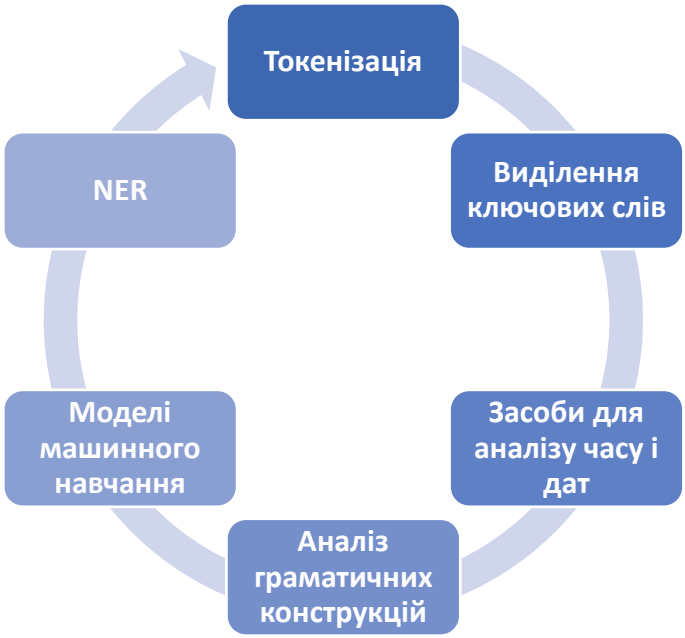


Рисунок 1.1 – Основні кроки та методи для пошуку подій в текстах

Обробка природної мови (NLP) може визначати події в тексті шляхом аналізу лінгвістичних ознак і контексту, що вказують на те, що дія або подія сталася. Перш ніж розглядати текст як ціле, він має бути поділений на окремі токени (слова, фрази, речення). Токенізація – це перший і важливий крок у роботі з текстом. Вона включає розділення тексту на окремі токени, які можуть бути словами, фразами або реченнями. Це спрощує подальший аналіз тексту, оскільки текст розбивається на легко оброблювані одиниці. Токенізація допомагає розділити текст на окремі одиниці, які можуть бути в подальшому оброблені. Перед аналізом тексту важливо провести препроцесинг, такий як перетворення тексту у нижній регістр, лематизація (зведення слів до базової форми), видалення стоп-слів (зайвих слів, таких як "і", "в", "на", які не несуть значущої інформації), і видалення пунктуації.

Аналіз ключових слів допомагає визначити, які слова або фрази в тексті є ключовими для розуміння події. Наприклад, слова "вибух", "пожежа", "аварія" можуть вказувати на подію. Алгоритми виділення ключових слів визначають основні терміни або фрази, які мають велике значення для змісту тексту. Вони можуть вказувати на потенційні події у тексті, так як ключові слова часто вказують на важливі концепти.

Аналіз граматичних структур допомагає визначити, які слова в тексті є суб'єктами, дієсловами, об'єктами тощо. Це допомагає встановити синтаксичні зв'язки між словами і зрозуміти структуру події.

Моделі машинного навчання, такі як багатокласові класифікатори або рекурентні нейронні мережі (RNN), можуть використовуватися для класифікації тексту за наявністю певних подій [15]. Наприклад, модель може бути навчена визначати текст, який описує пожежу або демонстрацію.

Визначення іменованих сутностей, таких як імена осіб, місця або організації, допомагає встановити, які сутності є пов'язаними з подією. Наприклад, ім'я особи може вказувати на участь цієї особи в певній події. Також, важливо визначити, коли відбулася подія або якісь дії. Для цього

використовуються методи виявлення часових виразів, такі як "вчора", "сьогодні", або визначення дат та часів у тексті.

Розуміння контексту тексту є важливим аспектом визначення подій, оскільки слова можуть мати різне значення в різних контекстах. Для обробки більшого обсягу текстових даних можуть використовуватися інструменти для аналізу текстових документів, такі як Elasticsearch або Apache Solr.

Також для ідентифікації подій в текстах може використовуватись дисперсійна оцінка. Цей параметр може бути корисним в якості підтримуючого інструменту для виявлення потенційних подій або тем, які стають предметом зростаючого обговорення в певний період часу. Наприклад, раптове збільшення частотності певних слів чи фраз може вказувати на важливі події, що сталися або зміни у суспільних настроях.

Для створення ефективного інструменту визначення подій у тексті, дисперсійну оцінку можна поєднати з рядом інших методів та технік обробки природної мови (NLP) та аналізу даних. На рисунку 1.2 наведено одні з можливих існуючих інструментів, з якими можна поєднати дисперсійну оцінку для визначення подій в україномовних текстах.

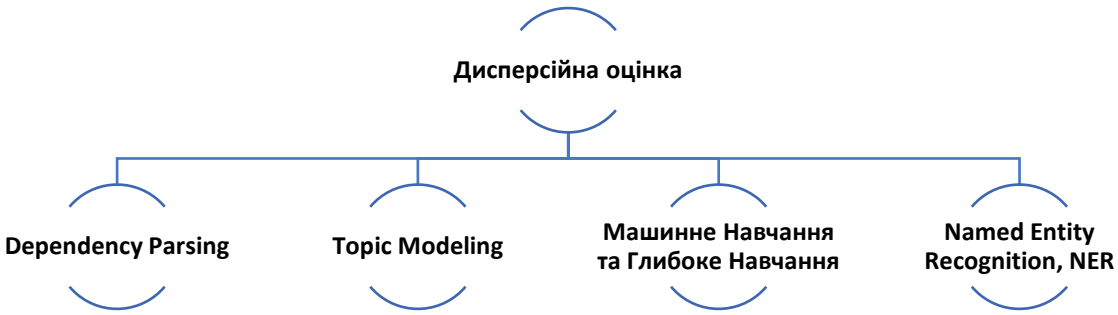


Рисунок 1.2 – Одні з можливих існуючих інструментів, з якими можна поєднати дисперсійну оцінку

Визначення іменованих сутностей може дозволити ідентифікувати та класифікувати важливі елементи у тексті, такі як імена осіб, географічні назви,

організації, дати та інші специфічні сутності, що є критичними для розуміння контексту подій.

Аналіз залежностей допомагає розуміти граматичні відносини між словами у реченні, що важливо для визначення структури та смислу висловлювань. Використання алгоритмів тематичного моделювання, таких як LDA, дозволить виявити основні теми чи сюжети, які домінують у тексті, і може вказувати на основні події або питання.

Використання алгоритмів машинного навчання та глибокого навчання для класифікації текстів, прогнозування та виявлення патернів. Моделі на зразок BERT або GPT можуть бути особливо корисними для розуміння контексту та нюансів мови [16].

Інтеграція таких методів з дисперсійною оцінкою дозволяє створити більш комплексний та багатовимірний підхід до аналізу тексту, здатний ефективно виявляти та аналізувати події. Однак, для досягнення високої точності та ефективності, важливо правильно налаштувати та оптимізувати використання цих інструментів відповідно до специфіки даних та цілей аналізу.

Таким чином, було проведено аналіз методів та засобів інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови. Використання різноманітних методів, включаючи статистичний аналіз, машинне навчання, глибоке навчання та аналіз залежностей, дозволяє розпізнавати, класифікувати та аналізувати події. Важливою складовою є використання дисперсійної оцінки та іменованих сутностей для виділення ключових подій, а інтеграція різних підходів забезпечує комплексний аналіз текстів для ефективного їх виявлення та розуміння.

З проведеного аналізу для ідентифікації подій в україномовних текстах були обрані підходи визначення дисперсійної оцінки для визначення важливості слів у текстових документах та іменованих сутностей в тексті, що дозволяє ідентифікувати та класифікувати важливі елементи у тексті, такі як імена осіб, географічні назви, організації, дати та інші специфічні сутності, що є критичними для розуміння контексту подій.

1.3 Аналіз наукових публікацій напряму ідентифікації подій засобами обробки природної мови

Для розуміння подальшої побудови методу інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови необхідно провести детальний аналіз існуючих наукових публікацій напряму інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови.

Було проведено пошук наукових досліджень та статей на базі Scopus [17] та Google Scholar [18]. В дослідженні [19] було використано набір даних TimeBank для вирішення проблем ідентифікації подій та визначення їхніх семантичних класів у текстах. Метод базується на класифікації слів, визначаючи, чи вони входять у подію, та вказуючи їхній семантичний клас. За цільову задачу прийнято два завдання: визначення слів, що відображають події, та визначення їхніх семантичних класів.

Автори детально розглядають методику розпізнавання подій як завдання класифікації, використовуючи слова та їх контекст. Для цього кожному слову присвоюється мітка (B-I-O), що вказує на його роль у події. Запропоновано ряд важливих властивостей, які допомагають у визначенні, чи слово є частиною події, та який саме тип події воно представляє.

Автори використовують різні групи ознак, такі як текстові, афіксальні, морфологічні, класи слів, управлінські, часові, визначені за допомогою зовнішніх ресурсів, таких як WordNet. Важливою є інтеграція ознак, які кодують контекст та семантику слів, для досягнення високої точності в класифікації.

Результати експериментів свідчать про успішність моделей, зокрема вони перевершують базові системи. Отримана точність в ідентифікації подій та їхніх семантичних класів свідчить про ефективність підходу, але вказує і на потребу у подальших дослідженнях для розширення обсягу тренувальних даних та удосконалення ознак для подальшого поліпшення моделей.

Автори [20] та [21] реалізували технологію пошуку тенденцій твітів на основі кластеризації, яка формує потік даних у вигляді коротких представлень кластерів та їх популярності для подальшого дослідження громадської думки1. Для досягнення точності результату впливає природна мовна особливість потоку інформації твітів. Описано ефективний підхід до збору, фільтрації, очищення та попередньої обробки твітів на основі порівняльного аналізу алгоритмів Bag of Words, TF-IDF та BERT1.

Також авторами визначено вплив стемінгу та лематизації на якість отриманих кластерів. Стемінг та лематизація дозволяють значно зменшити вхідний словник українських слів на 40,21% та 32,52% відповідно. Знайдено оптимальні комбінації методів кластеризації (K-Means, Agglomerative Hierarchical Clustering та HDBSCAN) та векторизації твітів на основі аналізу 27 кластерів одного зразка даних.

Вибрано метод представлення кластерів твітів у короткому форматі. Алгоритми, які використовують відстань Левенштейна, тобто розмитість сортування, розмитість набору та Левенштейн, показали найкращі результати.

Ці алгоритми швидко виконують перевірки, мають більшу різницю в подібності, тому можна більш точно визначити межу подібності.

Згідно з результатами кластеризації, оптимальними рішеннями є використання алгоритму кластеризації HDBSCAN та алгоритму векторизації BERT для досягнення найточніших результатів, а також використання K-Means разом з TF-IDF для досягнення найкращої швидкості з оптимальним результатом. Автори [20], [21] також зазначають, що вибір статистичних підходів та NER можуть бути ефективними для вирішення задачі визначення подій в тексті.

В результаті виконання аналізу існуючих наукових статей, можна зробити висновок, що проведені дослідження підтверджують, що статистичні методи машинного навчання можуть успішно застосовуватися для задачі ідентифікації подій у текстах. Дана методика може бути застосована для подальших розвитків

у сфері обробки природної мови та відзначається потенціалом для подальших досліджень та вдосконалення алгоритмів розпізнавання подій у текстах.

1.4 Аналіз існуючого програмного забезпечення напряму ідентифікації подій засобами обробки природної мови

Для реалізації методу інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови необхідно дослідити існуюче програмне забезпечення цього напряму для виявлення в них сильних та слабких сторін.

Одним із популярних засобів для аналізу є «Hume AI» [22]. Hume AI – це інструмент, який допомагає розуміти емоційний стан людей, а також виявляти тон та настрій тексту. Hume AI пропонує набір моделей, які дозволяють розпізнавати емоційні вирази в мовленні, відтворювати відтінки голосу, аналізувати міміку обличчя та багато іншого (рисунок 1.3). Цей інструмент може бути корисним для підприємств, які хочуть зрозуміти, як їхні клієнти реагують на їхні продукти або послуги.

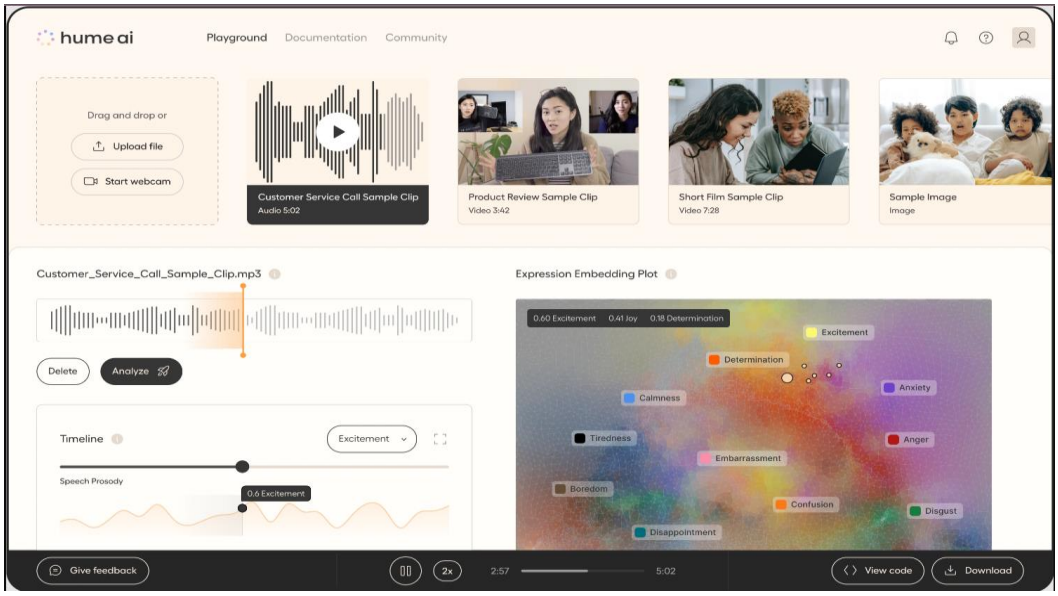


Рисунок 1.3 – Інтерфейс застосунку «Hume AI» [22]

Застосунок має багато функцій (рисунок 1.4), що можуть бути корисні не лише в аналізі тексту, а й мовлення, розпізнавання обличь та емоцій.

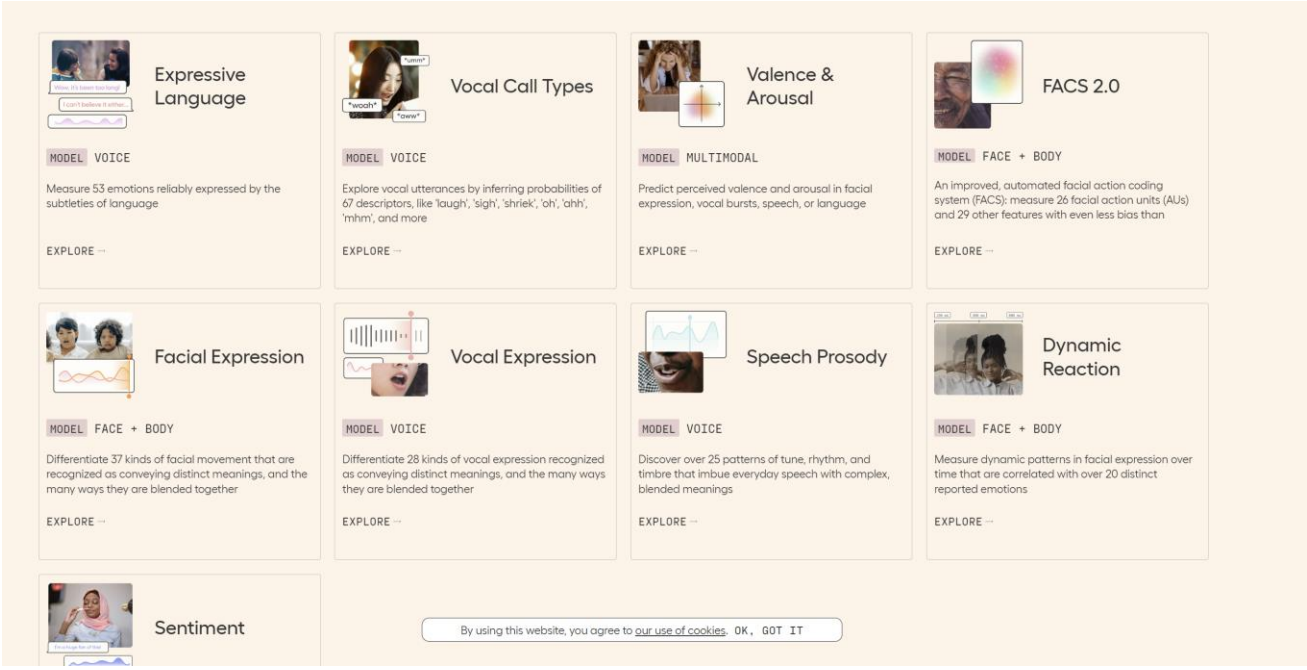


Рисунок 1.4 – Група функцій застосунку «Hume AI» [22]

Платформа надає API для наукових досліджень і розробки додатків, які реагують на експресивну поведінку людини, дотримуючись етичних норм і найкращих наукових практик. Серед можливостей моделей Hume AI – визначення виразу обличчя, просодії мови (тон, ритм і тембр мовлення), голосових сплесків (таких як сміх і зітхання) та емоційного тону транскрибованого тексту.

Однак, виходячи з наявної інформації, Hume AI більше фокусується на емоційних та експресивних аспектах мови, а не на розпізнаванні подій у традиційному розумінні NLP. Платформа вимірює емоційну мову за 53 параметрами, але окремо не згадує про здатність ідентифікувати окремі події в тексті.

Однак, Hume AI має і свої недоліки:

– сервіс не підтримує українську мову;

– не має функцій для визначення подій в тексті;

– не є безкоштовним і може бути досить дорогим для деяких груп користувачів.

Hume AI дозволяє розпізнавати емоційні вирази в мовленні, відтворювати відтінки голосу, аналізувати міміку обличчя та багато іншого. Це може бути корисним для підприємств, які хочуть зрозуміти, як їхні клієнти реагують на їхні продукти або послуги. Однак, Hume AI може бути досить дорогим для деяких користувачів.

Також одним з потужних інструментів є Komprehend [23]. Це інструмент, який допомагає розуміти текстову інформацію, виявляти тон та настрій тексту, а також класифікувати документи. Komprehend пропонує набір моделей, які дозволяють розпізнавати емоційний стан тексту, виявляти ключові слова та фрази, аналізувати сентимент та багато іншого. Цей інструмент може бути корисним для підприємств, які хочуть зрозуміти, як їхні клієнти реагують на їхні продукти або послуги (рисунок 1.5). Крім того, Komprehend є безкоштовним для використання на початковому етапі, але може бути дорогим для користувачів, які потребують більш високої точності.

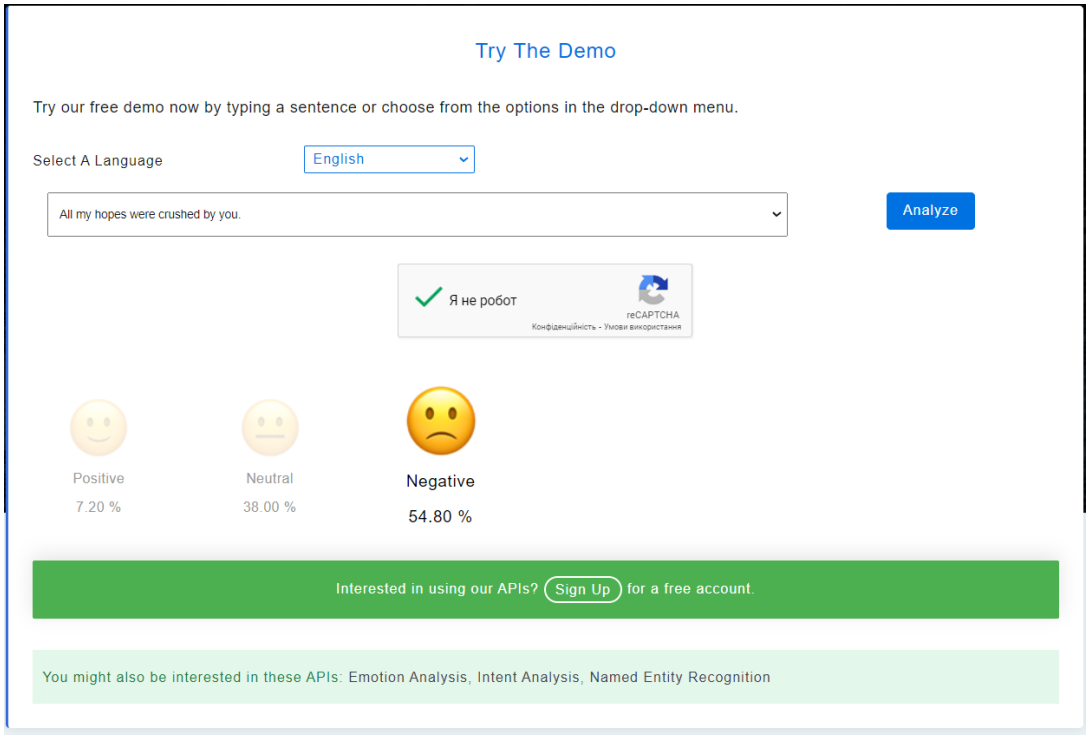


Рисунок 1.5 – Інтерфейс застосунку «Komprehend» [23]

Komprehend дозволяє розуміти текстову інформацію, виявляти тон та настрій тексту, а також класифікувати документи. Це може бути корисним для підприємств, які хочуть зрозуміти, як їхні клієнти реагують на їхні продукти або послуги. Крім того, Komprehend є безкоштовним для використання на початковому етапі, але може бути дорогим для користувачів, які потребують більш.

API платформи можуть аналізувати різні типи даних, що може бути особливо корисним у різних галузях, таких як фінанси та охорона здоров'я. Komprehend забезпечує конфіденційність і відповідність директивам GDPR, пропонуючи гнучкі варіанти розгортання, включаючи розгортання в приватній хмарі через контейнери Docker або локальне розгортання, гарантуючи безпеку і конфіденційність даних.

Для розробників Komprehend пропонує низку викликів API на день, починаючи з безкоштовного тарифного плану, який дозволяє до 1 000 викликів API, та різні рівні доступу в залежності від обраного тарифного плану. Їхні API підтримують кілька мов для аналізу настроїв, аналізу емоцій та генерації ключових слів, тоді як інші API НЛП наразі обмежені англійською мовою.

Компанія Amazon, яка розробляє Komprehend, заявляє, що ця платформа має високу точність у виявленні настрою та емоцій у тексті. Однак, деякі користувачі відзначають, що точність може бути нижчою, коли використовується для аналізу текстів, що містять багато сленгу, або текстів, що містять багато граматичних помилок. Крім того, користувачі повинні платити значні суми коштів за використання платформи, якщо вони потребують більш складних функцій.

Таким чином, провівши дослідження існуючих програмних продуктів, було встановлено, що жоден з них не надає можливості визначення подій в україномовних текстах, хоч застосунки сильні в області визначенні емоційної тональності, не пропонують необхідних засобів для визначення подій в тексті.

1.5 Постановка задачі

Метою кваліфікаційної роботи магістра є вирішення задачі ідентифікації подій в україномовних текстах засобами обробки природної мови, що дає можливість за вхідними даними у вигляді україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Також необхідно створити відповідну програмну реалізацію для апробації методу.

Для досягнення мети слід вирішити наступні завдання:

1. Дослідити сучасний стан підходів щодо інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови.

2. Розробити метод інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити події за вхідними даними у вигляді текстового документу та відповідної математичної моделі перетворити у вихідні дані у вигляді формування висновку щодо подій, визначених в тексті.

3. Створити тестову програмну реалізацію розробленого методу.

4. Дослідити практичну ефективність застосування методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Висновки до розділу 1

В результаті виконання першого розділу кваліфікаційної роботи магістра, було проведено дослідження предметної області ідентифікації подій в україномовних текстах засобами обробки природної мови та сучасного стану цією області досліджень. В рамках виконання розділу було оглянуто актуальні методи та інструменти, які використовуються в цій галузі, а також проведено аналіз їхнього ефективності та можливостей застосування.

Зокрема, було виявлено, що ідентифікація подій в україномовних текстах вимагає високого рівня обробки мовленнєвих особливостей та семантики мови.

Однією з ключових складових успішного методу ідентифікації подій є розпізнавання іменованих сутностей, таких як імена людей, назви організацій, місця тощо, які є важливими у контексті подій.

Було здійснено аналіз сучасних методів та засобів ідентифікації подій в україномовних текстах засобами обробки природної мови та встановлено, що застосування методів статистичної оцінки в поєднанні із сучасними засобами глибокого навчання може дозволити досягнути значного покращення точності ідентифікації подій в україномовних текстах засобами обробки природної мови.

Основною метою цього дослідження є розробка та реалізація методу ідентифікації подій в україномовних текстах. Для досягнення цієї мети потрібно розробити програмний продукт, який здатний аналізувати тексти, виділяти та визначати події, що відіграють ключову роль в тексті.

Важливим аспектом цього дослідження є також можливість адаптації методу до різних контекстів та мовних особливостей української мови. Це дозволить отримувати точні та об’єктивні результати ідентифікації подій у різних сферах, включаючи новинні статті, соціальні медіа та інші джерела інформації.

Також визначено мету кваліфікаційної роботи магістра, якою є вирішення задачі інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити ключові події за вхідними даними у вигляді текстового документу та відповідного методу перетворити у вихідні дані у вигляді формування висновку щодо подій, визначених в тексті. Також необхідно створити відповідну програмну реалізацію для апробації методу.

Розділ 2 Метод ідентифікації подій в україномовних текстах засобами обробки природної мови

2.1 Схема та кроки методу ідентифікації подій в текстах

Метод ідентифікації подій в україномовних текстах засобами обробки природної мови є актуальною задачею в галузі обчислювальної лінгвістики та штучного інтелекту. Цей метод спрямований на автоматичне визначення подій, які відбуваються в тексті, зокрема, діяльності, подій, процесів та їхніх атрибутів, вхідними даними методу слугуватимуть текстові дописи із корпусу україномовних текстів або ж текст, введений чи обраний з дискового простору користувачем. Ідентифікація подій є важливою для багатьох застосувань, включаючи аналіз новин, моніторинг соціальних мереж, розробку систем автоматичного розуміння тексту, аналізу семантики та багато інших областей.

Необхідно чітко окреслити послідовність кроків виконання методу ідентифікації подій в україномовних текстах. Схема роботи методу наведена на рисунку 2.1. Вхідними даними роботи методу є текстовий допис.

На першому кроці роботи методу, для подальшої роботи із текстом відбувається завантаження сформованого словника термінів подій, що вказуватимуть на події в текстовому дописі.

На другому кроці відбувається попередня обробка тексту та приведення його до нормалізованої форми, тобто відбувається видалення стоп-слів, видалення зайвих символів та приведення тексту до нижнього регістру.

Третій крок передбачає використання нейромережевої моделі Stanza для визначення іменованих сутностей в тексті. За допомогою Stanza можна ідентифікувати та класифікувати ключові елементи в тексті, такі як імена людей, організацій, географічних локацій тощо. Крок повертає перелік іменованих сутностей та відповідних міток, ідентифікованих в тексті.

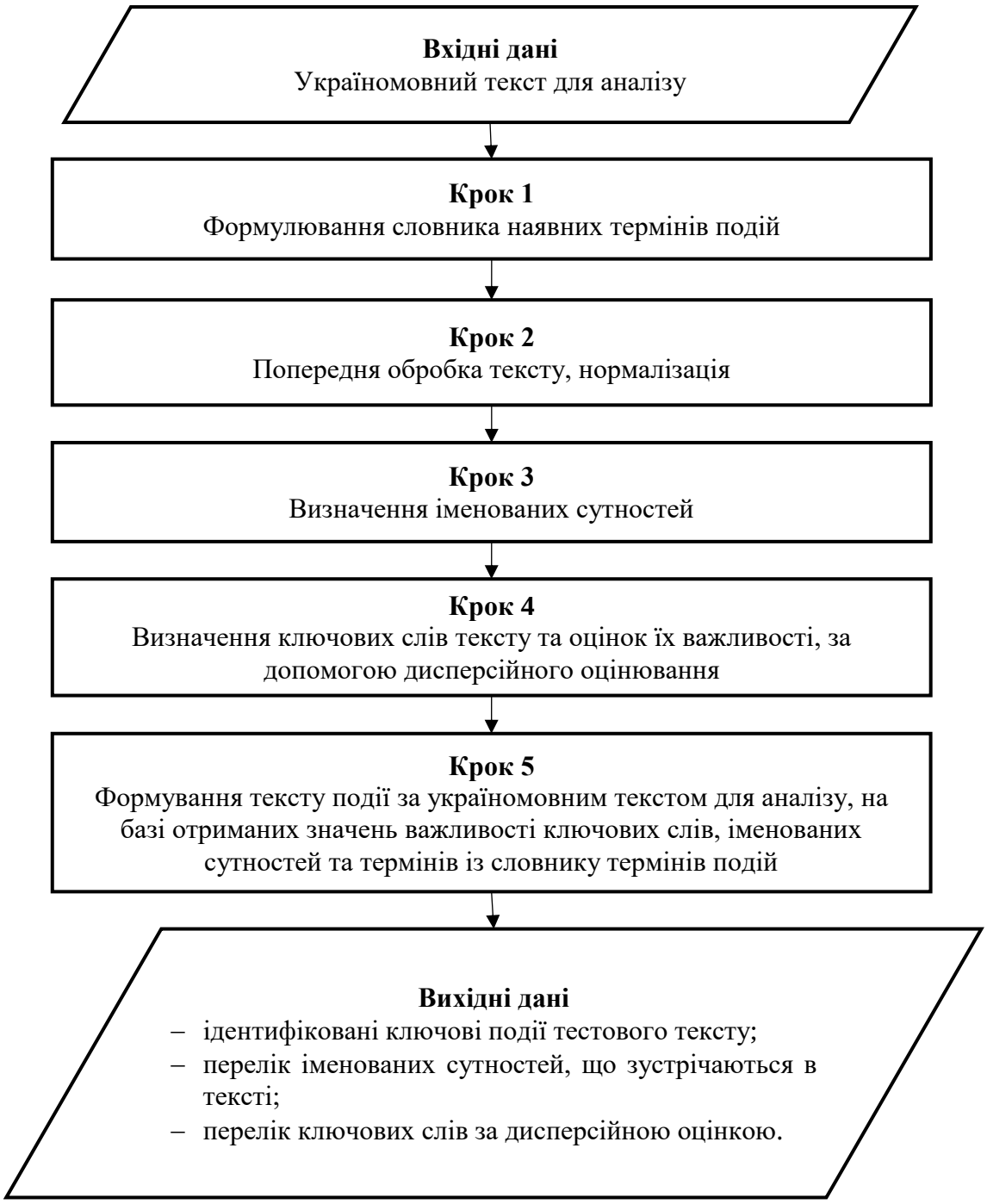


Рисунок 2.1 – Схема методу ідентифікації подій в україномовних текстах засобами обробки природної мови

Крок четвертий передбачає визначення ключових термінів в тексті за допомогою визначення дисперсійної оцінки. Крок повертатиме множину слів та відповідні дисперсійні оцінки.

Крок 5 на базі отриманих результатів з кроку 3 та 4 формує множину термінів, доповнюючи їх словником термінів подій, аналізує текстовий допис на

ключові події. В результаті виконання кроку повертається значення текстового абзацу чи речення, де згадуються іменовані сутності, визначені в тексті та ключові слова із підкріпленням змісту словника термінів подій.

Вихідними даними роботи методу є множина іменованих сутностей, визначених в тексті, множина ключових термінів та сформований висновок щодо ключових подій в тестовому тексті.

Таким чином, було наведено схему та кроки послідовності виконання методу ідентифікації подій в україномовних текстах засобами обробки природної мови із використанням дисперсійної оцінки для визначення ключових термінів тексту, нейронної мережі Stanza для визначення іменованих сутностей та словника термінів подій. В результаті роботи методу буде повертатись множина іменованих сутностей, визначених в тексті, множина ключових термінів та сформований висновок щодо ключових подій в тестовому тексті.

2.2 Формування словника термінів подій для їх ідентифікації в україномовних текстах засобами обробки природної мови

Для формування словника термінів подій для їх ідентифікації в україномовних текстах засобами обробки природної мови необхідно визначити, якого роду записи можуть міститись у таких словниках. Це можуть бути:

– специфічні ключові слова, терміни, що безпосередньо вказують на події (наприклад, «вибори», «конференція»);

– асоційовані слова, які часто вживаються разом із ключовими словами подій (наприклад, «голосування» для «виборів»).

– синоніми і варіації, різні форми ключових слів, що допомагають ураховувати різноманітність мови.

– контекстні правила, що допомагають розрізняти різні значення слів залежно від контексту.

Для реалізації україномовного словника було обрано дослідження та датасет [24]. Створений Айзенбергом та іншими науковцями у 2020 році, корпус

даних Personal Events in Dialogue Corpus містить анотовані транскрипти діалогів з чотирнадцяти епізодів подкасту «Це американське життя». Він містить 1 038 висловлювань, що складаються з 16 962 лексем, з яких 3 664 представляють події, англійською мовою. Містить 1,038 у форматі текстового файлу. Анотації були розмічені в текстових файлах. Кожен текстовий файл містить транскрипт епізоду, сформований так, щоб кожне висловлювання було в окремому рядку. Проміжки тексту, які скрипт вважав подіями, були оточені дужками. Зазвичай події складалися окремими словами, але іноді події були багатослівними виразами. Зразок файлу в наборі даних із виокремленими подіями наведено нижче:

Alan: Due to safety {concerns}, safety {purposes}.... But I mean, I can {type out} a little bit of, like, whatever you {want} to {tell} them, {tell} the shelter, and I can {make sure} they {get} the {message} if that'll {work} for you.

Рисунок 2.2 – Результати формування словника.

У датасеті наведено приклади текстів із визначеними подіями у фігурних дужках. Засобами Python було виокремлено ті слова та за допомогою лематизації машинного перекладу сформовано словник термінів подій для української мови.

Таким чином, нижче наведено приклад того ж фрагменту тексту для україномовного контенту.

Алан: З застережень безпеки {застерігати}, з міркувань {міркувати} безпеки... Але я можу надрукувати {надрукувати} дещо з того, що ви хочете {хотіти} їм сказати {сказати}, передати притулку {передати}, і я можу переконатися {переконатися}, що вони отримають {отримати} це повідомлення, якщо це вам допоможе {допомагати}.

Результат формування словника наведено на рисунку 2.2.

Таким чином, було сформовано словник термінів подій за допомогою перетворення датасету «Personal Events in Dialogue Corpus Dataset», що містить 1038 висловлювань, що складаються з 16962 лексем, з яких 3664 представляють події, українською мовою.

2.3 Підхід до визначення ключових термінів із використанням дисперсійної оцінки

Одним із кроків роботи методу ідентифікації подій в україномовних текстах засобами обробки природної мови є визначення ключових термінів, що зустрічаються в тексті.

Для вирішення поставленої задачі було обрано метод дисперсійної оцінки, що є статистичним методом для визначення різноманітності та варіабельності даних у вибірці. Дисперсія відіграє важливу роль у численних аналітичних завданнях, включаючи оцінку значущості слів у текстових документах для подальшого використання у методах обробки природної мови, таких як ідентифікація подій.

Дисперсійна оцінка ґрунтується на обчисленні середнього квадратичного відхилення (стандартного відхилення) даних від їхнього середнього значення. Цей метод вимірює, наскільки дані розподілені відносно середнього значення. Вище значення дисперсії вказує на більший розкид даних, тоді як нижча дисперсія вказує на менший розкид [25].

Формула для обчислення дисперсії наведена нижче

$$\sigma^2 = \frac{\sum(xi - \mu)^2}{N} \quad , \tag{2.1}$$

де σ^2 – дисперсія, xi – кожне окреме значення вибірки, μ – середнє значення (середнє арифметичне) вибірки, N – кількість значень вибірки.

Основна ідея полягає у вимірюванні дисперсії, яка вказує на те, наскільки значення вибірки відхиляються від середнього значення. Велика дисперсія свідчить про великий розкид даних, тоді як мала дисперсія вказує на подібність значень вибірки.

Дисперсійна оцінка зберігає смисловий зміст тексту, оскільки вона виділяє ключові слова та фрази, які мають семантичний зв’язок із змістом подій. Це важливо для правильної ідентифікації та розуміння подій в текстах. Також дисперсійна оцінка може бути легко та ефективно реалізована та розрахована, що робить її застосування доцільним в методах обробки природної мови.

Отже, для реалізації методу ідентифікації подій в україномовних текстах засобами обробки природної мови, головним підходом для визначення ключових термінів слугуватиме дисперсійна оцінка.

2.4 Нейромережева архітектура Stanza для обробки природної мови

Одним із кроків роботи методу ідентифікації подій в україномовних текстах засобами обробки природної мови є визначення іменованих сутностей, що зустрічаються в тексті.

Для вирішення поставленої задачі було обрано нейромережеву модель Stanza [26].

Stanza – це бібліотека обробки природної мови, розроблена Стенфордським університетом, яка забезпечує інструменти для виконання широкого спектру завдань NLP. Вона включає модулі для токенізації, морфологічного аналізу, визначення іменованих сутностей, аналізу залежностей та інших функцій. Stanza підтримує багато мов і використовує передові техніки

глибинного навчання для обробки тексту, роблячи її корисною для дослідників та розробників у галузі NLP.

Структуру архітектури нейромережевої моделі Stanza наведено на рисунку 2.3.

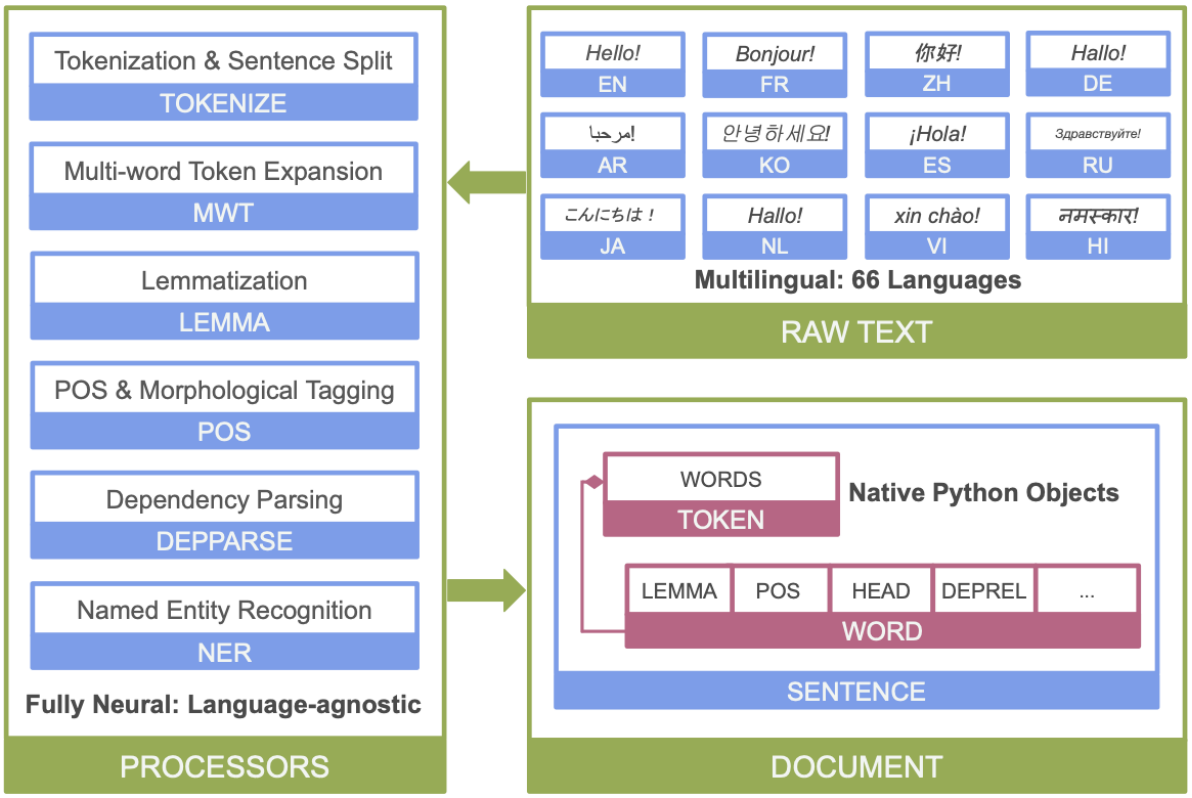


Рисунок 2.3 – Схема архітектури нейромережевої моделі Stanza [26]

Головні шари нейромережі, що виконують фундаментальні операції при обробці даних наведено на рисунку 2.4.

Вхідний шар (Input Layer) приймає вхідні дані, які можуть бути у формі векторів, зображень або текстових даних. Цей шар перетворює дані у формат, зручний для подальшої обробки нейромережею.

Приховані шари (Hidden Layers) виконують складні обчислення за допомогою ваг та активаційних функцій. Конволюційні шари (Convolutional Layers) використовують фільтри для виявлення характеристик у вхідних даних, часто застосовуються в обробці зображень. Рекурентні шари (Recurrent Layers), такі як LSTM, ефективні для обробки послідовностей, зберігаючи інформацію з попередніх кроків.

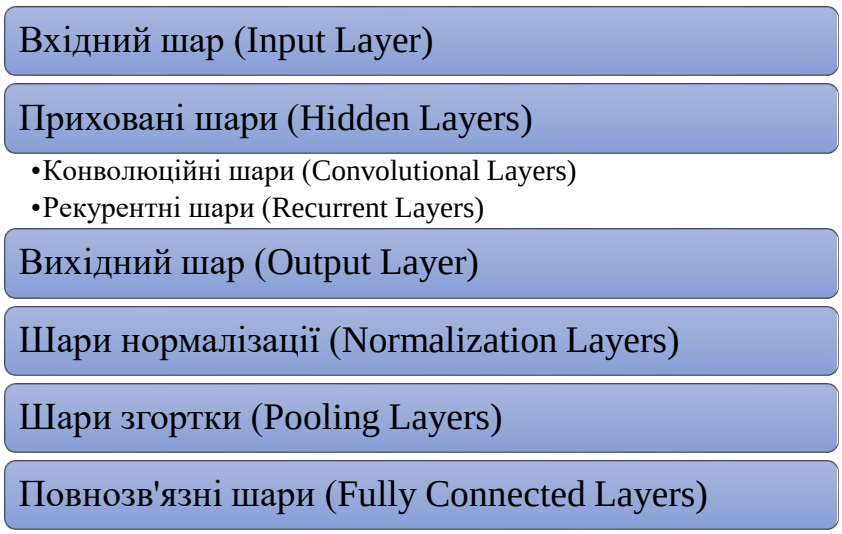


Рисунок 2.4 – Головні шари нейромережі Stanza, що виконують фундаментальні операції

Вихідний шар (Output Layer) виводить кінцевий результат, часто у формі класифікаційних міток або інших видів виходів.

Шари нормалізації (Normalization Layers) стандартизують вхідні дані, покращуючи стабільність та швидкість навчання. Шари згортки (Pooling Layers) зменшують розмір даних, виділяючи важливі ознаки.

Повнозв’язні шари (Fully Connected Layers) з’єднують кожен нейрон з усіма нейронами в попередньому шарі, забезпечуючи комплексний аналіз даних.

Отже, використання дисперсійної оцінки в методі ідентифікації подій в україномовних текстах обрано через її здатність враховувати різноманітність слів, уникнення перенавчання, збереження смислового змісту та легкість в реалізації, а нейромережева модель Stanza необхідна для визначення іменованих сутностей, що зустрічаються в текстовому дописі. Таким чином, дане поєднання засобів для визначення подій в україномовних текстах підхід сприяє покращенню ефективності та точності ідентифікації подій у текстах.

2.5 Формування тексту ключових подій методу ідентифікації в україномовних текстах засобами обробки природної мови

Для створення потужного інструменту визначення подій у тексті, дисперсійну оцінку можна поєднати з іншими технологіями та методами обробки природної мови (NLP).

На 2.5 наведено приклад інтеграції NER та дисперсійної оцінки.

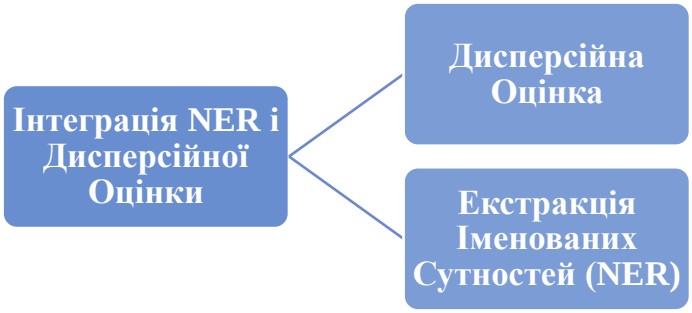


Рисунок 2.5 – Схема прикладу застосування NER та дисперсійної оцінки

Використання NER для ідентифікації ключових іменованих сутностей у тексті, таких як імена осіб, організацій, місцеположень, дат, часових міток тощо. Ці сутності часто є важливими компонентами подій тексту, в якому з'являються ці сутності, допомагає зрозуміти їхню роль у подіях, описаних у тексті.

Використання дисперсійної оцінки для аналізу того, як часто та рівномірно ключові слова або фрази (включаючи іменовані сутності) з'являються у тексті та виявлення слів, які мають низьку дисперсію (тобто часто зустрічаються в певних частинах тексту), що може вказувати на їхню важливість для певної події – основний математичний підхід для побудови методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Після отримання списку іменованих сутностей та аналізу дисперсії, можна об'єднати ці дані для отримання повного розуміння подій. Наприклад, якщо певна іменована сутність (наприклад, назва місця) зустрічається переважно в контексті певних ключових слів або фраз з низькою дисперсією, це може вказувати на специфічну подію. Використовуючи інформацію про іменовані

сутності та їх розподіл у тексті, можна робити висновки про характер, місцеположення, час та учасників подій. Для реалізації NER та дисперсійного аналізу можна скористатися існуючими бібліотеками NLP, такими як spaCy, NLTK або Stanford NLP. Методи можуть бути адаптовані під конкретні завдання, наприклад, для аналізу новинних текстів, наукових публікацій, блогів тощо.

Процес формування тексту ключових подій методу ідентифікації в україномовних текстах наведено на рисунку 2.6.

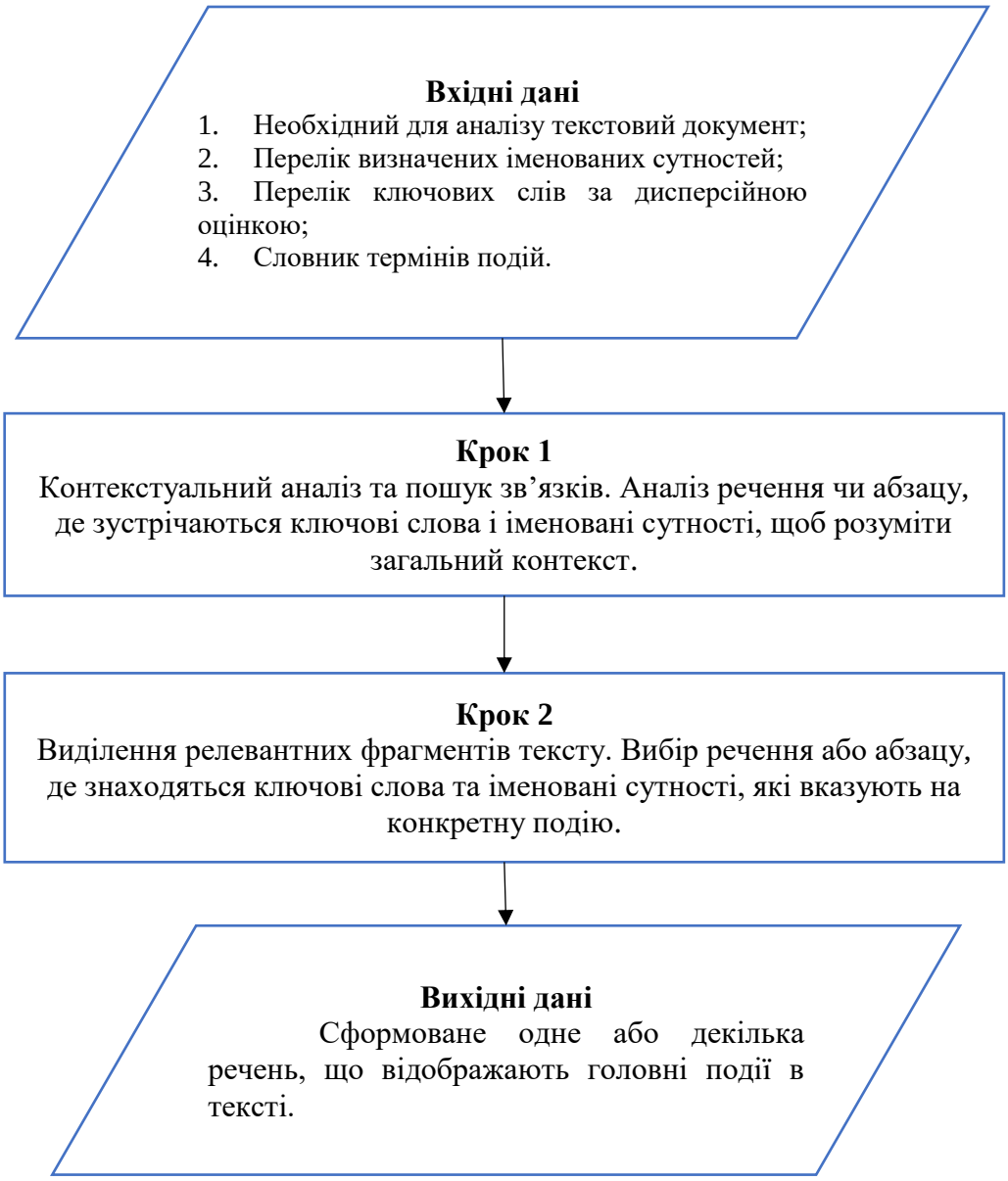


Рисунок 2.6 – Схема процесу формування тексту ключових подій

Вхідним даними є необхідний для аналізу текстовий документ, перелік визначених іменованих сутностей, перелік ключових слів за дисперсійною оцінкою та словник термінів подій.

На першому кроці відбувається контекстуальний аналіз та пошук зв’язків, аналіз речення та абзаци, де зустрічаються ключові слова і іменовані сутності, щоб розуміти загальний контекст.

Програма аналізує, де і як використовуються ключові слова і іменовані сутності. Наприклад, визначає, чи слово «конференція» є частиною оголошення про майбутню подію.

Далі встановлюються взаємозв’язки між словами, тобто з’ясовується, чи пов’язані між собою слово «конференція» і конкретна дата, що допомагає визначити час події.

Крок 2 – виділення релевантних фрагментів тексту. На цьому етапі відбирається речення або абзаци, які містять ключові слова і іменовані сутності, вказуючи на конкретну подію, таким чином виділяючи найбільш важливі частини тексту для подальшого аналізу.

Таким чином, було описано структуру формування тексту ключових подій методу ідентифікації в україномовних текстах засобами обробки природної мови. Результат роботи цього методу повертатиме висновок у вигляді одного чи декількох речень, де зустрічається найбільше термінів, що вказують на ключові події, про які йдеться в тексті.

2.6 Формування та підготовка корпусу текстів для методу ідентифікації подій в україномовних текстах

Для реалізації методу ідентифікації подій в україномовних текстах засобами обробки природної мови на базі поєднання роботи методів NER та дисперсійної оцінки, необхідно сформувати корпус текстів.

Для української мови існує збалансований корпус-мільйонник сучасної мови «БрУК». Репозиторій [27] містить Браунський корпус української мови

(БрУК), який є відкритим, проанотованим корпусом сучасної української мови. Корпус побудований на засадах, що були покладені в основу відомого корпусу англійської мови Brown. Репозиторій містить допоміжні файли, фрагменти текстів, зібрані для корпусу, перевірені фрагменти, написані літературною українською мовою, перевірені фрагменти, що містять помилки, перевірені фрагменти, що зовсім не відповідають вимогам (наприклад, усне мовлення), фрагменти, що чекають на перевірку, документацію, скрипти та список доданих творів.

Цей корпус є корисним для дослідження сучасної української мови, а також для розробки програмних засобів, які працюють з текстами українською мовою. «БрУК» містить близько 1 мільйона слововживань, структура набору даних наведена на рисунку 2.7.

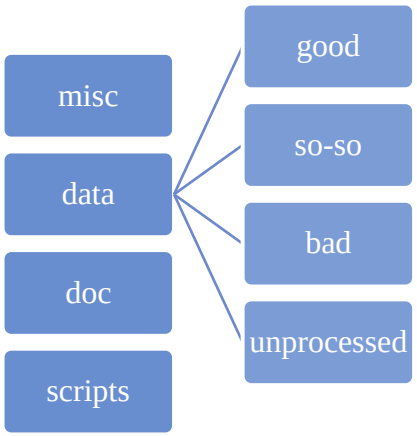


Рисунок 2.7 – Структура корпусу «БрУК»

Зміст датасету розділено на декілька частин. Основні папки містять наступне:

Категорія «misc» (допоміжні файли). В цій категорії зберігаються файли, які не є прямо пов’язані з основними текстовими даними корпусу, але є важливими для проекту. Це можуть бути файлові налаштування, шаблони, допоміжні таблиці або інструкції для користувачів.

Категорія «data» (фрагменти текстів). Тут збережено основний обсяг текстів для корпусу, зокрема:

– Категорія «good» (перевірені фрагменти) – якісні фрагменти, що відповідають встановленим критеріям. Їх можна додатково категоризувати за стилями, авторами чи часовими періодами.

– Категорія «so-so» (фрагменти з помилками) – фрагменти, які містять помилки, але все ще можуть бути корисні для аналізу мовних особливостей або помилок.

– Категорія «bad» (фрагменти, що не відповідають вимогам) – тексти, які не відповідають встановленим критеріям, наприклад, усне мовлення або неструктурований текст, і можуть бути використані для специфічних досліджень.

Категорія «unprocessed» містить фрагменти, що очікують на перевірку. Ці тексти чекають на класифікацію та аналіз, і для них слід встановити чіткі процедури обробки.

Категорія «doc» містить важливі інструкції, стандарти та методологію роботи з корпусом, включаючи анотацію, оцінювання якості та формати файлів.

Категорія «scripts» містить скрипти для автоматизації процесів аналізу, обробки даних, перетворення форматів та вирішення типових мовних завдань.

Така структура забезпечує ефективну організацію та управління даними в рамках роботи, сприяє зручності доступу до інформації та оптимізує процеси аналізу та дослідження.

Таким чином, було сформовано та підготовано корпус текстів для подальшої роботи методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Висновки до розділу 2

В результаті виконання розділу було реалізовано метод ідентифікації подій в україномовних текстах засобами обробки природної мови, у якому вхідними даними роботи методу є текстовий допис із корпусу україномовних дописів або введений власноруч користувачем, вихідними, в свою чергу –

множина іменованих сутностей, визначених в тексті, множина ключових термінів та сформований висновок щодо ключових подій в тестовому тексті.

Також для вирішення поставленої задачі було обрано метод дисперсійної оцінки, що є статистичним методом для визначення різноманітності та варіабельності даних у вибірці. Використання дисперсійної оцінки в методі ідентифікації подій в україномовних текстах обрано через її здатність враховувати різноманітність слів, уникнення перенавчання, збереження смислового змісту та легкість в реалізації, а нейромережева модель Stanza необхідна для визначення іменованих сутностей, що зустрічаються в текстовому дописі. Таким чином, дане поєднання засобів для визначення подій в україномовних текстах підхід сприяє покращенню ефективності та точності ідентифікації подій у текстах.

В результаті виконання розділу також було отримано словник термінів подій за допомогою перетворення датасету «Personal Events in Dialogue Corpus Dataset», що містить 1038 висловлювань, що складаються з 16962 лексем, з яких 3664 представляють події, українською мовою.

Окрім того, було описано структуру формування тексту ключових подій методу ідентифікації в україномовних текстах засобами обробки природної мови. Результат роботи цього методу повертатиме результат у вигляді одного чи декількох речень, де зустрічається найбільше термінів, що вказують на ключові події, про які йдеться в тексті. Також було обрано датасет для подальшої реалізації методу подій в україномовних текстах засобами обробки природної мови, корпус-мільйонник сучасної мови «БрУК».

Розділ 3 Проєктування інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови

3.1 Схема інформаційної системи

Інформаційна система автоматизованої ідентифікації подій в україномовних текстах засобами обробки природної мови виконана на базі методу автоматизованого визначення ідентифікації подій в україномовних текстах засобами обробки природної мови, вхідними даними для роботи якого є текстовий допис, а в якості вихідних даних є висновок щодо відповідності введеного токену до подій та ключових слів, що містяться в тексті. Інформаційна система ідентифікації подій в україномовних текстах засобами обробки природної мови складається із 3 підсистем: «Підсистеми для роботи із корпусами текстів», «Підсистеми для визначення частовживаних слів та подій», «Підсистеми для визначення відповідності токенів» та відповідної бази даних.

Схему інформаційна система ідентифікації подій в україномовних текстах засобами обробки природної мови наведено на рисунку 3.1.

Підсистема для роботи з текстом призначена для роботи із змістом текстових документів, що містяться в БД. В підсистемі реалізовано наступні функції:

– Вибір текстового корпусу. Користувач має змогу вибрати певний набір текстів для аналізу, який може бути заздалегідь структурованим і категоризованим за різними параметрами (наприклад, тематика, жанр, автор).

– Перегляд корпусу. Ця функція дає змогу переглядати вміст обраного корпусу текстів, що є необхідним для ручного аналізу та оцінки якості текстів перед подальшою автоматизованою обробкою.

– Редагування корпусу. Можливість редагування дозволяє коригувати текст у корпусі, видаляти чи додавати фрагменти, а також вносити зміни для поліпшення якості аналізу.

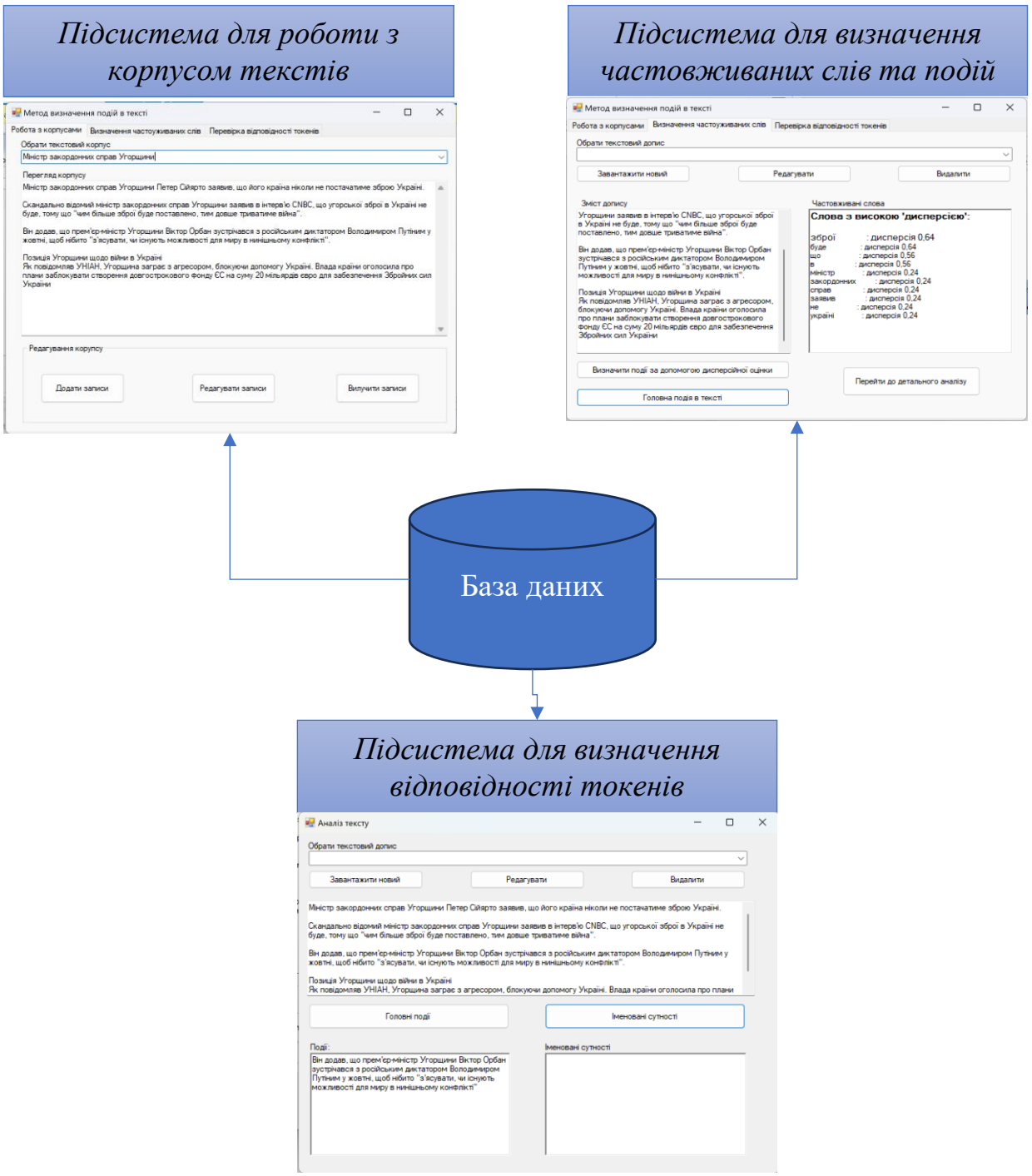


Рисунок 3.1 – Схема інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови

Підсистема для визначення частовживаних слів та подій реалізує можливість визначати частовживані слова й терміни текстах, переглядати та визначати події, що в них описані. Нижче наведено перелік функцій, що надає користувачеві підсистема для визначення частовживаних слів та подій.

– Робота із вибором та редагуванням текстового корпусу при потребі.

– Відображення текстового корпусу для попереднього перегляду в спеціально відведеній області.

– Обчислення та відображення результатів аналізу частоти вживання слів у вибраному тексті або корпусі.

– Визначення події за допомогою дисперсійної оцінки: Ця функція аналізує текст на наявність подій, використовуючи дисперсійну оцінку для визначення, наскільки розподілені ключові слова чи фрази, що можуть вказувати на події.

– Аналіз та перегляд подій, що зустрічаються в тексті.

Реалізація переходу до наступної підсистеми «Підсистема для визначення відповідності токенів» для більш глибокого аналізу вибраного тексту або корпусу.

Підсистема для визначення відповідності токенів має найширший функціонал та є основною в програмному застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови. Реалізовано наступні функції.

– Робота із вибором та редагуванням текстового корпусу при потребі.

Відображення текстового корпусу для попереднього перегляду в спеціально відведеній області.

– Текстове поле «Головні події призначене відображення ключових подій, виявлених у тексті за допомогою методу дисперсійної оцінки.

– Текстове поле «Іменовані сутності» відображатиме іменовані сутності, які були ідентифіковані в тексті за допомогою NER-інструментів.

Ця підсистема надає інструменти для комплексного семантичного та змістового аналізу текстів, що включає виявлення подій, ідентифікацію ключових осіб, організацій або географічних локацій. Поєднання підсистем та результат роботи програмного застосунку забезпечує значні можливості для дослідників, лінгвістів, маркетологів, аналітиків даних та інших фахівців, які працюють з великими обсягами текстової інформації.

Підсумовуючи, було реалізовано схему інформаційної системи для ідентифікації подій в україномовних текстах засобами обробки природної мови, реалізовано необхідні підсистеми, базу даних та налагоджено зв’язок між ними.

3.2 Проєктування бази даних для інформаційної системи ідентифікації подій в україномовних текстах

Для програмної реалізації методу ідентифікації подій в україномовних текстах засобами обробки природної мови необхідно створити базу даних, зв’язки між таблицями та наповнити початковими даними для побудови програмного застосунку та подальшого його тестування.

Та рисунку 3.2 наведено діаграму бази даних для методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

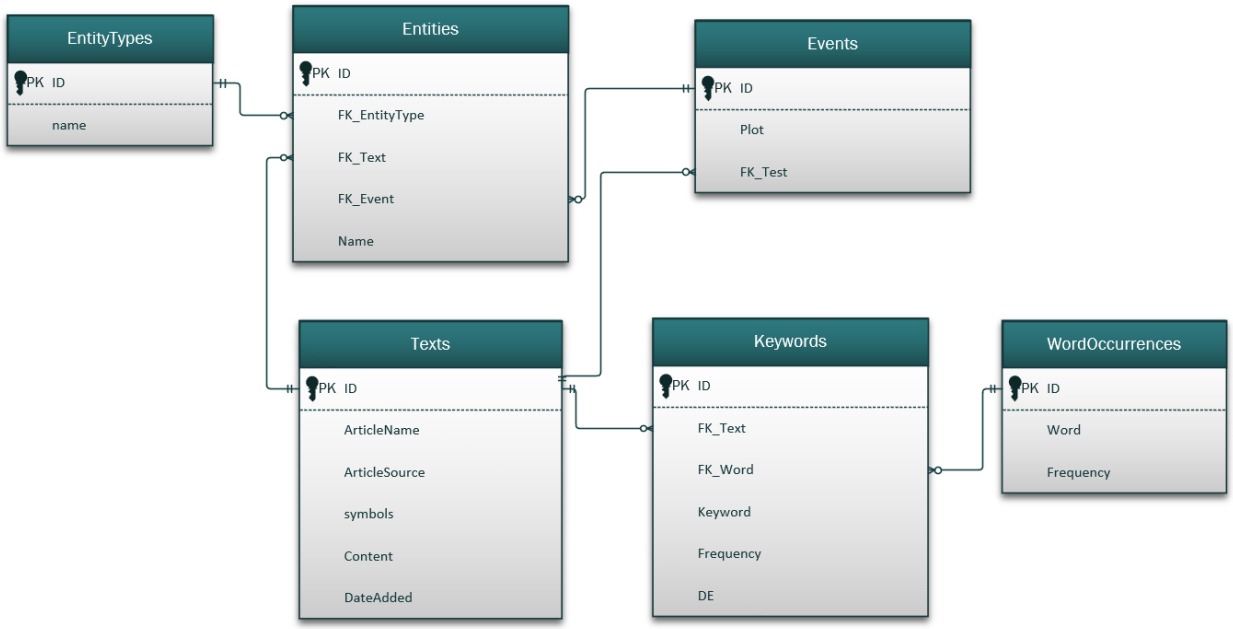


Рисунок 3.2 – Даталогічна модель даних інформаційної системи ідентифікації подій в україномовних текстах

Таблиця «Texts» (таблиця 3.1) призначена для збереження текстових корпусів та основної інформації про них.

Таблиця 3.1 – Атрибути таблиці «Texts»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	ArticleName	text	Назва текстового допису (заголовок статті)
3.	ArticleSource	text	Джерело текстового допису
4.	symbols	int	Кількість символів в текстовому дописі
5.	Content	text	Зміст текстового допису
6.	DateAdded	datetime	Дата й час додання текстового допису

Таблиця «Entities» (таблиця 3.2) призначена для збереження даних щодо іменованих сутностей, що зустрічаються у тексті. Таблиця містить поля для збереження вторинного ключа-посилання на запис таблиці «Texts» для співставлення із відповідним текстом, вторинного ключа-посилання на запис таблиці «Events» для співставлення із відповідною подією, та вторинного ключа-посилання на запис таблиці «EntityTypes» для співставлення із відповідним типом іменованої сутності.

Таблиця 3.2 – Атрибути таблиці «Entities»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	FK_Text	int	Вторинний ключ, посилання на відповідний запис таблиці «Texts» для співставлення із відповідним текстовим дописом
3.	FK_Event	int	Вторинний ключ, посилання на відповідний запис таблиці «Events» для співставлення із відповідною подією
4.	Name	text	Назва іменованої сутності

Таблиця «Events» БД інформаційної системи ідентифікації подій в україномовних текстах (таблиця 3.3) призначена для збереження даних щодо подій, що зустрічаються у тексті. Таблиця містить поля для збереження вторинного ключа-посилання на запис таблиці «Texts» та змісту події.

Таблиця 3.3 – Атрибути таблиці «Events»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	Plot	text	Зміст події
3.	FK_Text	int	Вторинний ключ, посилання на відповідний запис таблиці «Texts» для співставлення із відповідним текстовим дописом

Таблиця «EntityTypes» (таблиця 3.4) призначена для збереження типів подій. Таблиця містить зміст назви типу події.

Таблиця 3.4 – Атрибути таблиці «EntityTypes»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	name	text	Назва типу сутності

Таблиця «KeyWords» (таблиця 3.5) БД інформаційної системи ідентифікації подій в україномовних текстах призначена для збереження ключових слів, що містяться в тексті.

Таблиця 3.4 – Атрибути таблиці «KeyWords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	FK_Text	int	Вторинний ключ, посилання на відповідний запис таблиці «Texts» для співставлення із відповідним текстовим дописом
3.	FK_Word	int	Вторинний ключ, посилання на відповідний запис таблиці «Keywords» для співставлення із відповідним словом
4.	Frequency	double	Значення частоти вживання слова
5.	DE	double	Значення дисперсійної оцінки вживання слова

Таблиця «WordOccurrences» (таблиця 3.6) призначена для збереження ключових слів, що містяться в тексті.

Таблиця 3.6 – Атрибути таблиці «WordOccurrences»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	Word	text	Слово, що є ключовим
3.	Frequency	double	Частота вживання слова

Таким чином, було реалізовано необхідну базу даних для інформаційної системи ідентифікації подій в україномовних текстах. На основі даталогічної моделі БД створено необхідні для роботи таблиці та заповнено їх початковими вхідними даними.

3.3 Вибір спеціалізованих програмних розширень для розробки інформаційної системи

Для коректної роботи програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови необхідно проаналізувати додаткові засоби та модулі, що можуть використовуватись при розробці програмного коду.

Для токенізації та препроцесингу тексту необхідно використати NLTK для .NET. Це версія популярного Natural Language Toolkit (NLTK), адаптована для використання в середовищі .NET, включаючи C#. Бібліотека надає функціональність, подібну до оригінального NLTK на основі Python, який широко використовується для обробки та аналізу тексту в обробці природної мови (NLP).

Для токенізації NLTK для .NET пропонує інструменти, які можуть сегментувати текст на слова, речення або інші одиниці. Це корисно для попередньої обробки в таких завданнях NLP, як синтаксичний аналіз, класифікація тексту та аналіз настроїв. Інструментарій включає різні токенізатори з різними стратегіями для роботи з різними текстовими структурами та мовами.

ML.NET – це фреймворк машинного навчання від Microsoft, який дозволяє розробникам впроваджувати функції машинного навчання в .NET додатки. В контексті ідентифікації подій в текстах, ML.NET можна використовувати для тренування моделей, які аналізують текст і виявляють шаблони та ключові слова, пов'язані з певними подіями.

Такі моделі можуть бути натреновані на анотованих датасетах, де події позначені в тексті, щоб визначити, які слова або фрази є індикаторами подій. ML.NET включає в себе різні типи алгоритмів машинного навчання, які можуть бути використані для класифікації, регресії, пошуку аномалій в текстах та інші завдання NLP.

Для виокремлення іменованих сутностей було обрано нейромережеву модель Stanza. Модель працює із платформою .NET, завдяки додатковим розширенням від Stanford NLP for .NET.

Оскільки Stanza найкраще працює із платформою Python, було вирішено застосувати IronPython для поєднання програмного коду із платформою .NET. Інтеграція Python та C# є важливою темою у розробці сучасного програмного забезпечення. Python, з його потужними бібліотеками та простотою використання, часто вибирається для складних завдань обробки даних та машинного навчання. З іншого боку, C# є вибором для створення стабільних, безпечних і високопродуктивних додатків. Об’єднання цих двох мов програмування в одному проекті може забезпечити значні переваги, зокрема в областях, де необхідно об’єднати швидку обробку даних та надійність.

IronPython дозволяє виконувати Python код безпосередньо всередині C# застосунків. Це розширює можливості C# програм, дозволяючи їм використовувати різноманітні бібліотеки та фреймворки Python, що особливо корисно в галузі обробки природної мови та машинного навчання.

Для передачі даних між Python та C#, використовується об’єкт Scope, який створюється движком Python. Змінні та об’єкти можуть бути встановлені або отримані через Scope, що дозволяє обмін даними між C# та Python кодами.

Для реалізації програмного коду для використання Stanza потрібно визначити додаткові бібліотеки та модулі, що необхідні для визначення іменованих сутностей в тексті. В документації Stanza [28] зазначено, що для розгортання та роботи нейромережевої моделі необхідно додатково встановити PyTorch, NumPy та SciPy.

Оскільки Stanza розроблена для Python, необхідною є встановлена версія Python [29]. Часто використовується Python версії 3.5 або новішої через її сумісність із сучасними бібліотеками та інструментами. Для Stanza необхідно встановити Python версії 3.8 або вище.

Stanza використовує PyTorch, потужний фреймворк для глибокого навчання, який забезпечує необхідне обчислювальне середовище для тренування

та використання нейронних мереж [30]. PyTorch відомий своєю гнучкістю, швидкістю та зручністю у розробці, що робить його ідеальним вибором для NLP завдань.

NumPy вводить поняття багатовимірних масивів, які дозволяють ефективно зберігати та обробляти великі набори даних. В області NLP це може включати вектори ознак, частотні матриці тощо [31]. Використання оптимізованих бібліотек робить NumPy значно швидшим за стандартні Python списки, що критично важливо при обробці великих наборів даних.

NumPy надає розширені можливості для лінійної алгебри, включаючи підтримку векторних та матричних операцій, що є основою багатьох алгоритмів машинного навчання.

Таким чином, для налагодження роботи інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови було обрано спеціалізовані програмні розширення та модулі для розробки, а саме ML.NET, IronPython та для роботи нейромережевої моделі Stanza для ідентифікації іменованих сутностей було визначено NumPy та PyTorch.

3.2 Вибір засобів для реалізації інформаційної системи ідентифікації подій в україномовних текстах

Для реалізації програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови необхідно визначати набір засобів, що забезпечать надійну роботу додатку та передачу даних.

Одним із найпотужніших засобів для реалізації додатків, що використовують машинне навчання, статистиці методи тощо є платформа .NET, розроблена компанією Microsoft. Вона є однією з найбільш універсальних та широко використовуваних платформ для розробки програмного забезпечення, що включає широкий спектр застосувань – від веб-додатків до систем машинного навчання. З огляду на її здатність до інтеграції з багатьма іншими

технологіями, .NET стала популярним вибором для розробників у різних областях, включаючи роботу з текстом та обробку природної мови (NLP).

NET Framework та .NET Core включають великі бібліотеки для роботи з рядками та текстом, надаючи розробникам потужні інструменти для маніпуляції текстом, регулярних виразів, і кодування. В галузі ШІ та МН платформа .NET може легко інтегруватися з бібліотеками та сервісами для обробки природної мови, такими як spaCy, NLTK (через IronPython), або комерційними API, такими як Microsoft Cognitive Services.

.NET є потужною платформою, що підтримує широкий спектр додатків, від веб-розробки до розгортання складних систем машинного навчання. Її здатність легко інтегруватися з іншими технологіями, велика спільнота, постійні оновлення та покращення роблять її ідеальною для розробки сучасних, масштабованих і високопродуктивних додатків (рисунок 3.3).

Рисунок 3.3 – Перелік функцій .NET [32]

Засобом для написання програмного коду було обрано MS Visual Studio 2019. Visual Studio 2022 є останньою ітерацією одного з найпопулярніших інтегрованих середовищ розробки (IDE), створеного Microsoft для розробників програмного забезпечення. Цей масштабний інструментарій розробки забезпечує об'єднаний інтерфейс для створення застосунків для різних платформ, включаючи Windows, macOS, Linux, мобільні та хмарні сервіси.

З огляду на свою багатофункціональність, Visual Studio 2022 розширює свої можливості, включаючи підтримку сучасних мов програмування та фреймворків. Розробники можуть використовувати Visual Studio для створення складних додатків, використовуючи C#, C++, JavaScript, Python та багато інших мов, а також фреймворків, таких як .NET, Angular та React. На рисунку 3.4 наведено інтерфейс середовища розробки програмного коду.

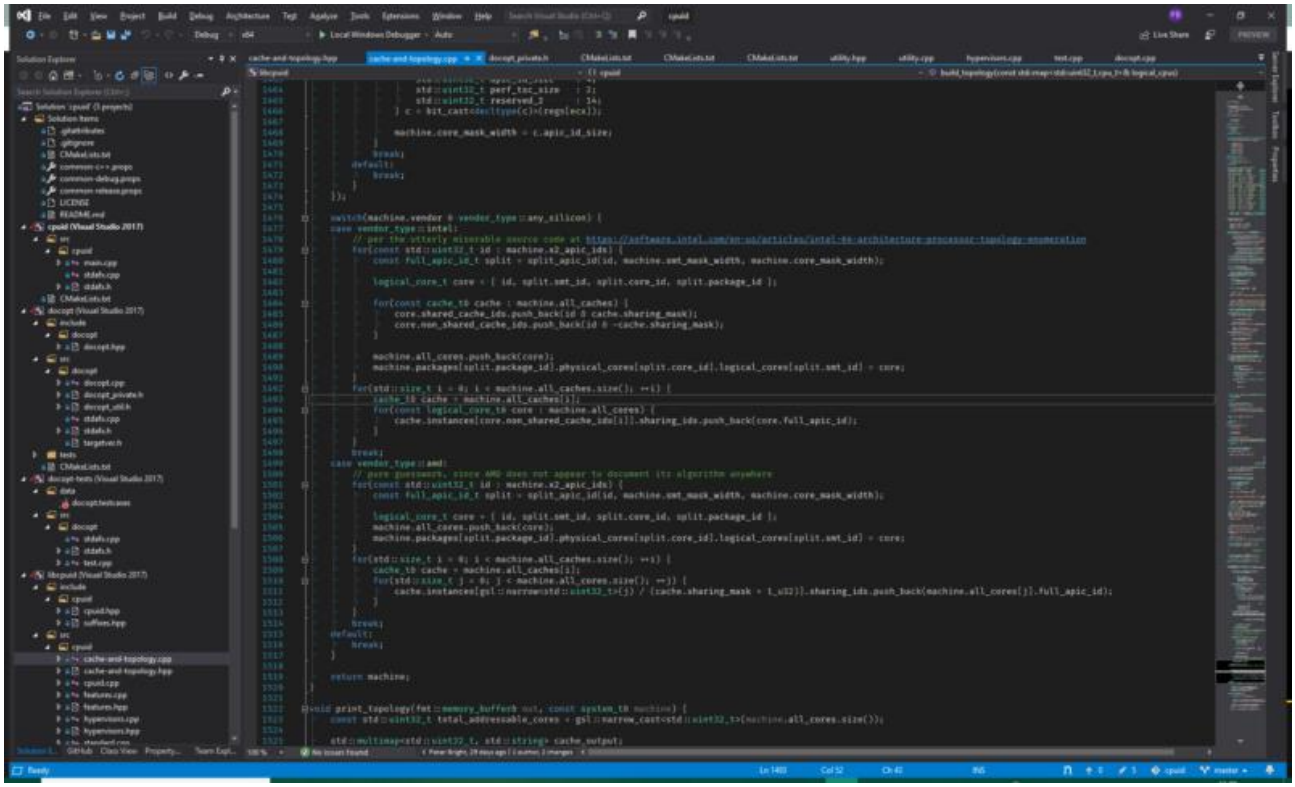


Рисунок 3.4 – Інтерфейс середовища розробки програмного коду MS Visual Studio 2019[33]

Особливість Visual Studio 2022 полягає в її удосконаленій продуктивності та ефективності, дозволяючи розробникам швидше виконувати завдання завдяки покращеному інтелектуальному кодуванню, рефакторингу та пошуку помилок. Редактор коду та інтерфейс були оптимізовані для підвищення чіткості та читабельності, а також для забезпечення більш зручного навігаційного досвіду.

Крім технічних вдосконалень, Visual Studio 2022 включає підтримку для сучасних практик розробки, таких як DevOps і Agile, інтегруючи засоби для співпраці, такі як Git, і функціонал для неперервної інтеграції/неперервного

розгортання (CI/CD). Це IDE є більш доступним, ніж коли-небудь, з вдосконаленнями для людей з особливими потребами та підтримкою різних мовних пакетів.

Загалом, Visual Studio 2022 пропонує розробникам гнучке та потужне середовище, яке не тільки полегшує процес написання коду, але й підтримує всі етапи життєвого циклу розробки програмного забезпечення — від концепції до випуску.

Отже, в якості платформи для реалізації програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови було обрано .NET, а середовище розробки програмного коду – Microsoft Visual Studio 2019.

Висновки до розділу 3

У цьому розділі спроєктовано та реалізовано інформаційну систему для ідентифікації подій в україномовних текстах засобами обробки природної мови, призначену для визначення подій в тексті та оцінки відповідності токена, введеного користувачем до тексту й подій, що містяться в ньому.

Метод використовує статистичні методи обробки природної мови та машинного навчання для визначення подій та іменованих сутностей в тексті. Вхідними даними є текстові документи, вихідними – множина слів та їх оцінок дисперсії, множина іменованих сутностей, визначених в тексті, висновок щодо подій, що є ключовими в тексті.

Було розроблено структуру інформаційної системи ідентифікації подій в україномовних текстах відповідно до послідовності кроків методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Інформаційна система на базі методу автоматизованого визначення ідентифікації подій в україномовних текстах засобами обробки природної мови складається із 3 підсистем: «Підсистеми для роботи із корпусами текстів»,

«Підсистеми для визначення частовживаних слів та подій», «Підсистеми для визначення відповідності токенів» та відповідної бази даних.

У розділі також було проведено аналіз існуючих засобів розробки програмного забезпечення та платформ. В якості платформи для реалізації програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови було обрано .NET, а середовище розробки програмного коду – Microsoft Visual Studio 2019.

Для налагодження роботи інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови було обрано спеціалізовані програмні розширення та модулі для розробки, а саме ML.NET, IronPython та для роботи нейромережевої моделі Stanza для ідентифікації іменованих сутностей було визначено NumPy та PyTorch.

Розділ 4 Дослідження ефективності методу ідентифікації подій в україномовних текстах засобами обробки природної мови

4.1 Програмна архітектура інформаційної системи

Для побудови програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови необхідно створити діаграму класів застосунку, на базі якої буде реалізовано програмний код (рисунок 4.1). Реалізовано три класи для проведення обчислень, кожен з яких має власний функціонал та цільове призначення та дві форми із відповідними вкладками.

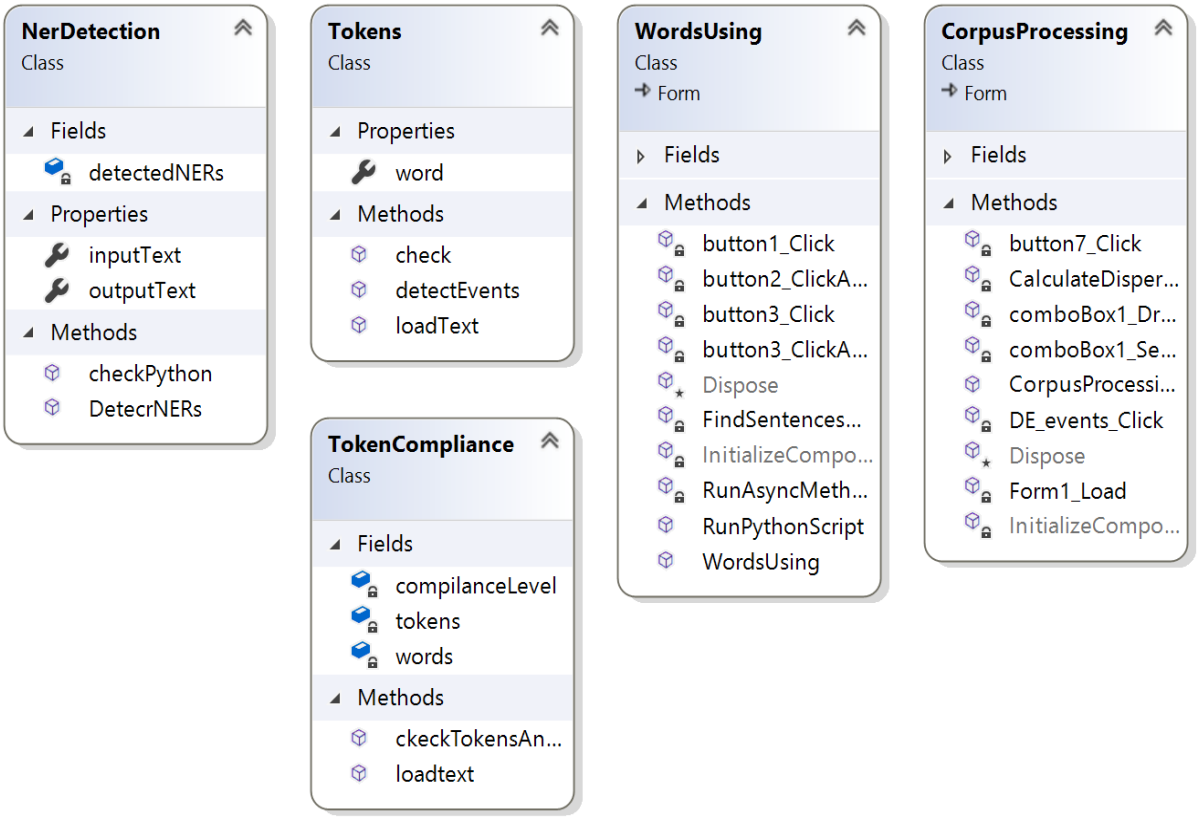


Рисунок 4.1 – Діаграма класів інформаційної системи ідентифікації подій в україномовних текстах

Клас «NerDetection» призначений для роботи із визначенням іменованих сутностей в тексті. Для виконання цієї задачі було використано засіб взаємозв’язку .NET та Python.

Спершу було реалізовано пайтон-скрипт для завантаження Stanford NLP та методу Stanza, що визначатиме іменовані сутності в тексті, переданого з форми C#. Для цього було створено метод checkPython, що перевіряє з’єднання із Python та скриптом методу визначення іменованих сутностей.

Клас «Tokens» призначений для роботи з токенами, словами, що вводить користувач. Клас включає наступні методи:

check – є методом для перевірки валідності токенів;

detectEvents – служить для виявлення подій в тексті, які виражені через токени;

loadText – метод для завантаження тексту, який буде токенізовано.

Клас «WordsUsing» містить інтерфейс користувача (форму) і забезпечує взаємодію з текстом. Реалізовано наступні методи:

– dispose – очистка ресурсів або закриття форми;

– findsentences – метод для виведення змісту тексту на екран;

– runpythonscript – запуск python скриптів, для обробки тексту;

– wordsusing – метод, що демонструє використання слів в тексті.

Клас «WordUsing» призначений для роботи із визначенням частоти вживання слова та основними операціями для подальшого визначення відповідності токенів. В класі реалізовано наступні методи:

– runpythonscript – запуск python скриптів, для обробки тексту;

– wordsusing – метод, що виводить основну інформацію по використанню слів в тексті.

Клас «TokenCompiliance» реалізований для визначення та виведення головних показників: іменованих сутностей, оцінок за дисперсією та оцінки відповідності токена до тексту.

Реалізовано методи:

– checkTokens – метод для перевірки відповідності токена до тексту;

– loadText – метод для завантаження тексту, який буде проаналізовано програмним застосунком.

Таким чином, було створено програмну архітектуру інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє в подальшому використати створену архітектуру реалізації відповідного програмного застосунку.

4.2 Особливості розробки прикладних компонентів інформаційної системи ідентифікації подій в україномовних текстах

Ключовим моментом у роботі інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови є використання нейромережевої моделі Stanza для визначення іменованих сутностей для подальшого пошуку подій в тексті.

На рисунку 4.2 наведено послідовність кроків визначення іменованих сутностей за допомогою Stanza.

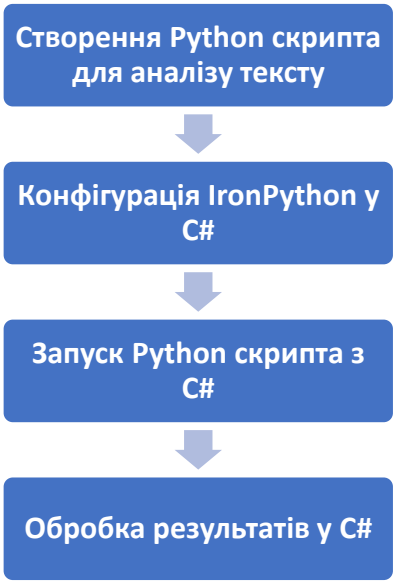


Рисунок 4.2 – Послідовність кроків визначення іменованих сутностей за допомогою Stanza

Інтеграція мов програмування C# та Python за допомогою IronPython для визначення іменованих сутностей в текстах за допомогою бібліотеки Stanza включає складний, але гнучкий підхід. Цей процес починається з встановлення та налаштування IronPython [34, 35], спеціальної реалізації Python для .NET Framework. Після встановлення IronPython, важливо інтегрувати необхідні бібліотеки Python, такі як Stanza, для обробки природної мови. Встановлення залежностей IronPython [36] у C# проекті дозволяє створити міст між двома мовами програмування.

Наступним кроком є розробка Python скрипта, який використовує нейромережеву модель Stanza для виявлення іменованих сутностей у тексті. Цей скрипт приймає вхідні дані, аналізує їх, та повертає результати аналізу. Зберігання скрипта у місці, доступному для C# проекту, забезпечує його легке використання в рамках більшої системи.

Структура алгоритму програмного коду, що використовує Stanza для визначення іменованих сутностей в тексті наведено на рисунку 4.3.

Для взаємодії між C# та Python використовується IronPython.hosting, який створює рушій Python в середовищі C#. Цей двигун дозволяє імпортувати бібліотеки Python, необхідні для виконання скрипта. Завдяки створенню Scope, яке функціонує як контейнер для даних, можлива передача та зберігання інформації між двома мовами.

Запуск Python скрипта з C# здійснюється через рушій IronPython [37, 38]. Текст для аналізу передається в Python скрипт через Scope або як аргумент командного рядка. Після виконання скрипта, результати аналізу можна витягнути та обробити в середовищі C#.

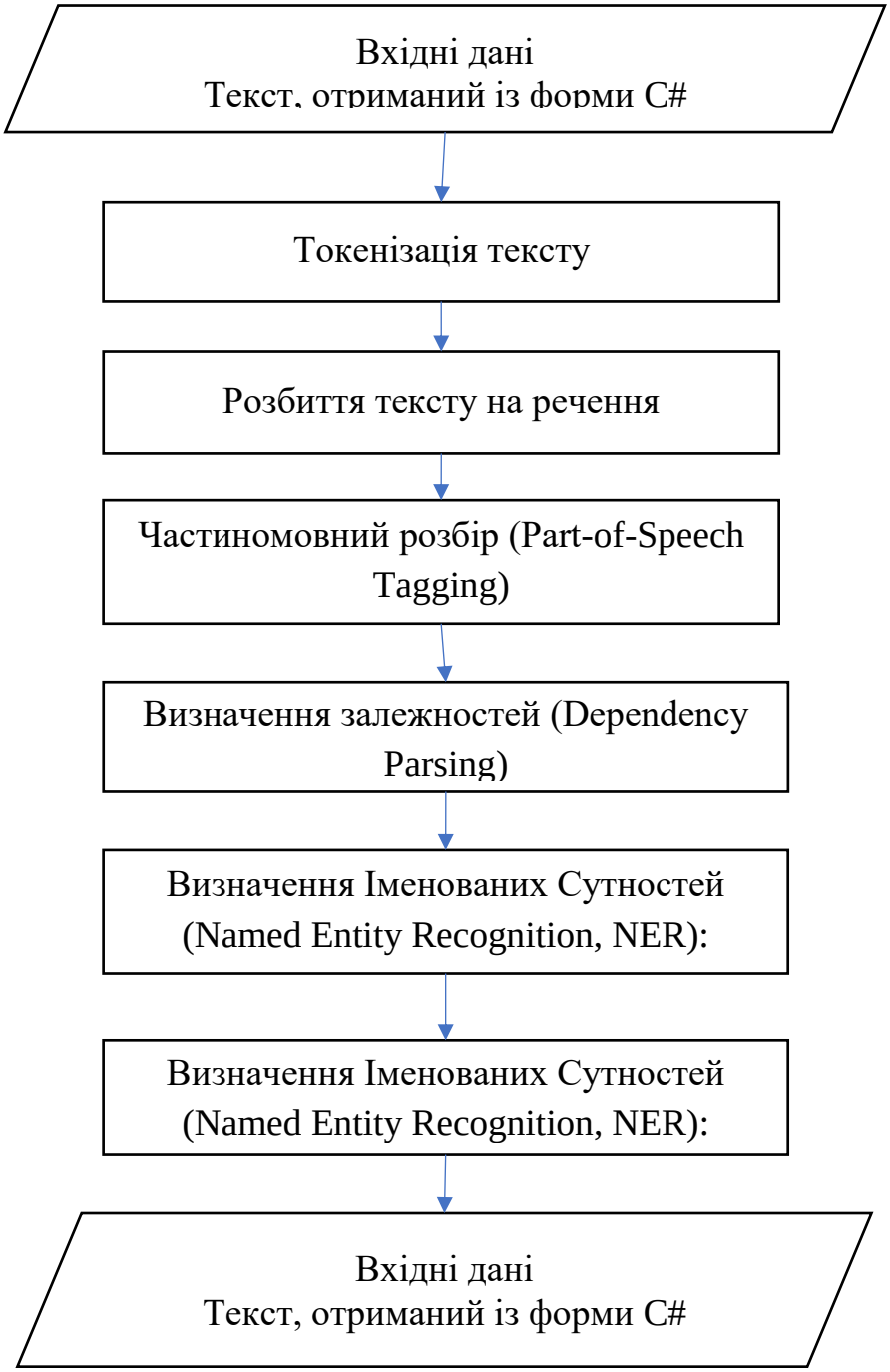


Рисунок 4.3 – Структура алгоритму програмного коду, що використовує Stanza для визначення іменованих сутностей в тексті

Такий підхід до передачі даних з C# використовує сильні сторони обох мов: гнучкість та широкий спектр бібліотек Python у сфері NLP та стабільність та розповсюдженість C# у розробці програмного забезпечення.

Отже, було описано ключовий момент у роботі інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови є

використання нейромережевої моделі Stanza для визначення іменованих сутностей для подальшого пошуку подій в тексті та реалізовано структуру алгоритму програмного коду, що використовує Stanza для визначення іменованих сутностей в тексті

4.3 Прикладне тестування інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови

Для здійснення прикладного тестування інформаційної системи визначення тональності текстової інформації по відношенню до іменованих сутностей було реалізовано ряд тест-кейсів.

Спершу необхідно перевірити, чи установлене з’єднання мж застосунком та корпусом текстів, для цього було реалізовано тест-кейс ТС-0001.

Таблиця 4.1 – Тест-кейс ТС-0001

Тест-кейс ID:ТС0001	**Пріоритет: 1**	**Створено:29.11.23,** Домбровський Н.С.
Назва: Перевірка наявності з’єднання між програмний застосунком та корпусом текстів.		
Кроки	**Очікуваний результат**	
1. Запуск програмного застосунку; 2. Перехід на вкладку «Робота з корпусами текстів»; 3. Натиснення на поле для відображення переліку текстів; 4. Вибір будь-якого файлу; 5. Перегляд вибраного текстового матеріалу із корпусу текстових дописів	Запуск головного екрану застосунку Відображення вмісту текстового корпусу у відповідному полі Текст відповідає очікуваному, тест-кейс пройдено успішно	
Результат виконання тест-кейсу: пройдено успішно		

Результат виконання тест-кейсу наведено на рисунку 4.4.

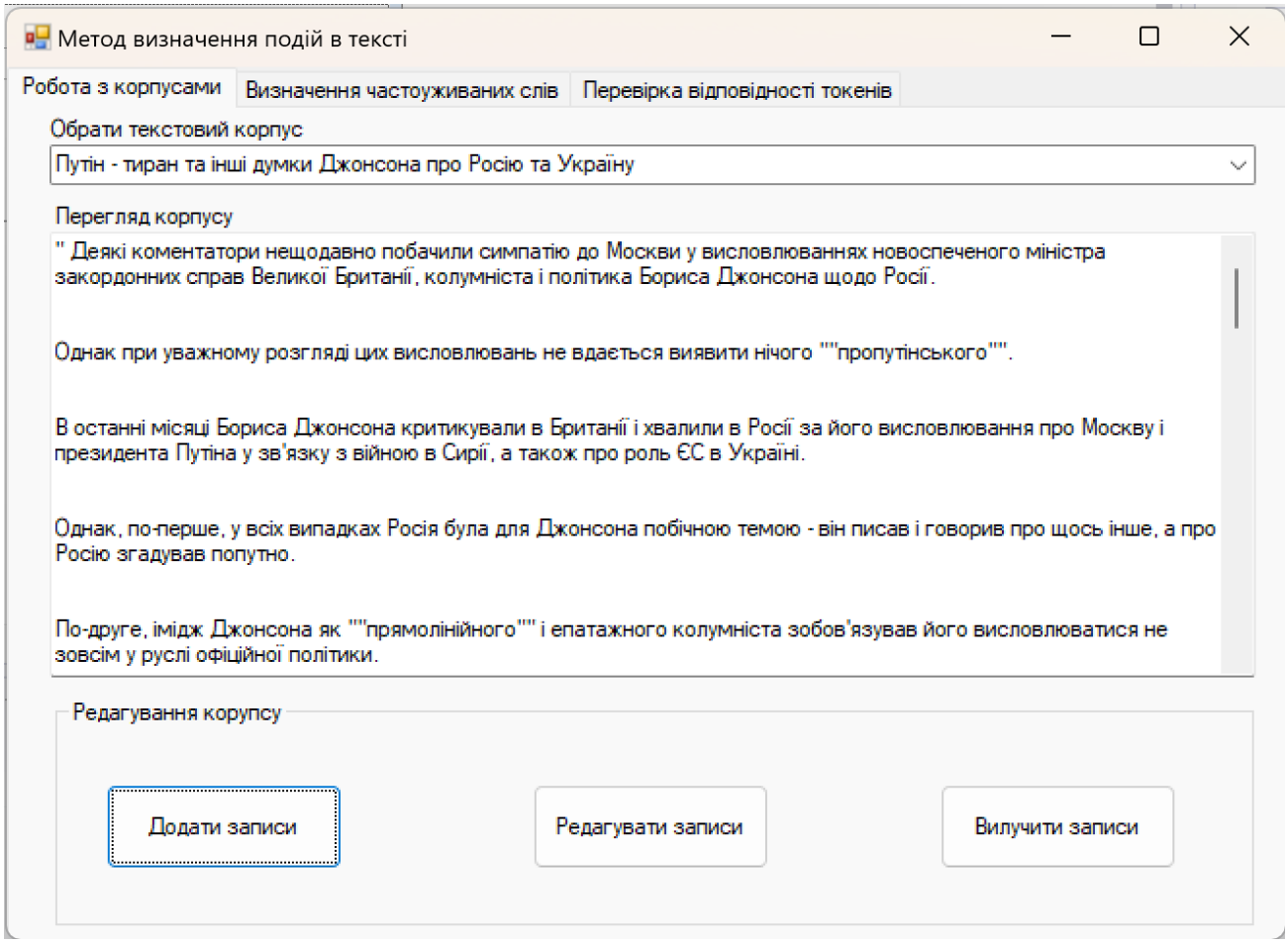


Рисунок 4.4 – Результат проходження тест-кейсу

Також необхідно перевірити ряд функцій, таких як:

– визначення іменованих сутностей за допомогою з’єднання із python скриптом;

– визначення дисперсійної оцінки в тексті;

– визначення подій в тексті;

– визначення відповідності введеного користувачем токена до текстового документу.

Для перевірки функції визначення іменованих сутностей за допомогою з’єднання із python скриптом було реалізовано тест-кейс ТС-0002 (таблиця 4.2). Цей тест-кейс ж дуже важливим, адже за допомогою нього перевірятиметься робота модуля IronPython.

Таблиця 4.2 – Тест-кейс ТС-0002

<table>
<tr><td>Тест-кейс ID:ТС-0001</td><td>Пріоритет: 1</td><td>Створено:29.11.23,
Домбровський Н.С.</td></tr>
<tr><td colspan="3">Назва: Перевірка функції визначення іменованих сутностей за допомогою з’єднання із python скриптом</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Перехід на вкладку "Визначення частовживаних слів" у інтерфейсі програми.
2. Натискання кнопки "Головна подія в тексті" для початку аналізу.
3. Вибір тексту для аналізу з випадаючого списку, якщо це не було зроблено раніше на вкладці "Робота з корпусами".
4. Повторне натискання кнопки "Головна подія в тексті" на формі "Визначення частовживаних слів" для ініціації аналізу обраного тексту.
5. У вікні, що з'явиться після аналізу, вибір опції "Іменовані сутності" для детального аналізу тексту.
6. Порівняння результатів, отриманих від системи, з очікуваними для оцінки точності аналізу.</td><td>Відображення іменованих сутностей в тексті</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

Результат виконання тест-кейсу наведено на рисунку 4.5

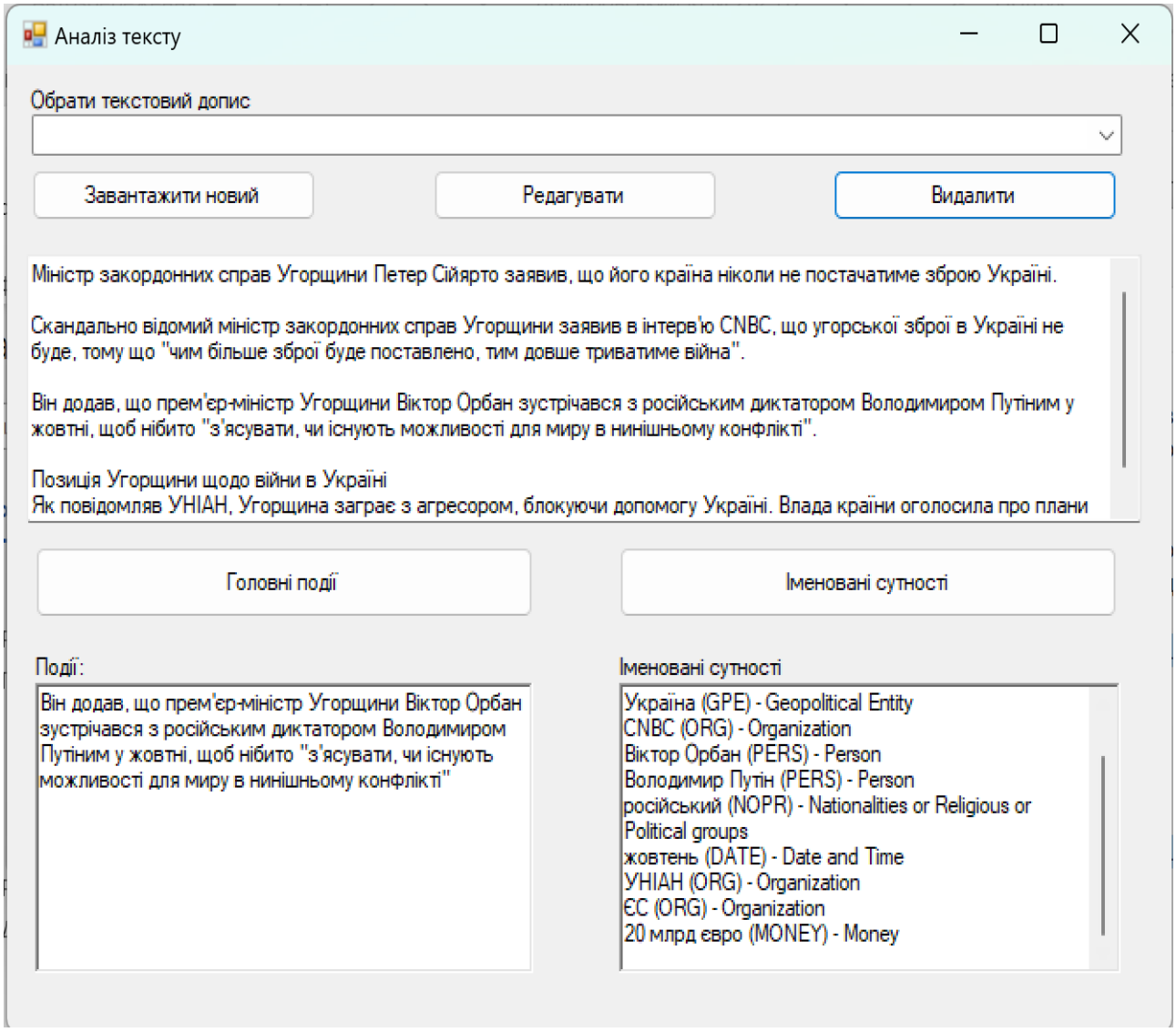


Рисунок 4.5 – Результат виконання тест-кейсу

Необхідно перевірити правильність функції визначення подій в тексті, для цього було реалізовано тест-кейс ТС-0003 (таблиця 4.3).

Цей тест-кейс описує процедуру використання спеціалізованої програми або інформаційної системи для аналізу тексту з метою визначення ключових подій. Користувачеві необхідно обрати конкретний текст з випадаючого списку для того, щоб система могла визначити, який саме текст слід аналізувати. Цей крок актуальний, якщо текст не був попередньо обраний на вкладці «Робота з корпусами».

Після аналізу тексту, користувачу інформаційної системи ідентифікації подій в україномовних текстах відкривається нове вікно, де він може виконати

подальші дії, зокрема натиснути кнопку «Визначити події», що сприяє більш детальному аналізу тексту на предмет виявлення специфічних подій.

Таблиця 4.3 – Тест-кейс ТС-0002

<table>
<tr><td>Тест-кейс ID:ТС-0003</td><td>Пріоритет: 1</td><td>Створено:29.11.23,
Домбровський Н.С.</td></tr>
<tr><td colspan="3">Назва: Перевірка функції визначення подій в тексті</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Перехід на вкладку «Визначення частовживаних слів»;
2. Натиснення на кнопку «Головна подія в тексті»;
3. Обрати в області вибору тексту у випадаючому списку необхідний текст для аналізу (якщо на вкладці «Робота з корпусами» ще не було обрано текст;
4. На формі «Визначення частовживаних слів» натиснути кнопку «Головна подія в тексті»;
5. На новому вікні, що відкрилось натиснути кнопку «Визначити події»;
6. Порівняти отриманий результат з очікуваним.</td><td>Відображення іменованих сутностей в тексті</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

Цей тест-кейс дозволяє перевірити, наскільки програма визначає ключові події в текстах. На рисунку 4.6 наведено результат виконання тест-кейсу.

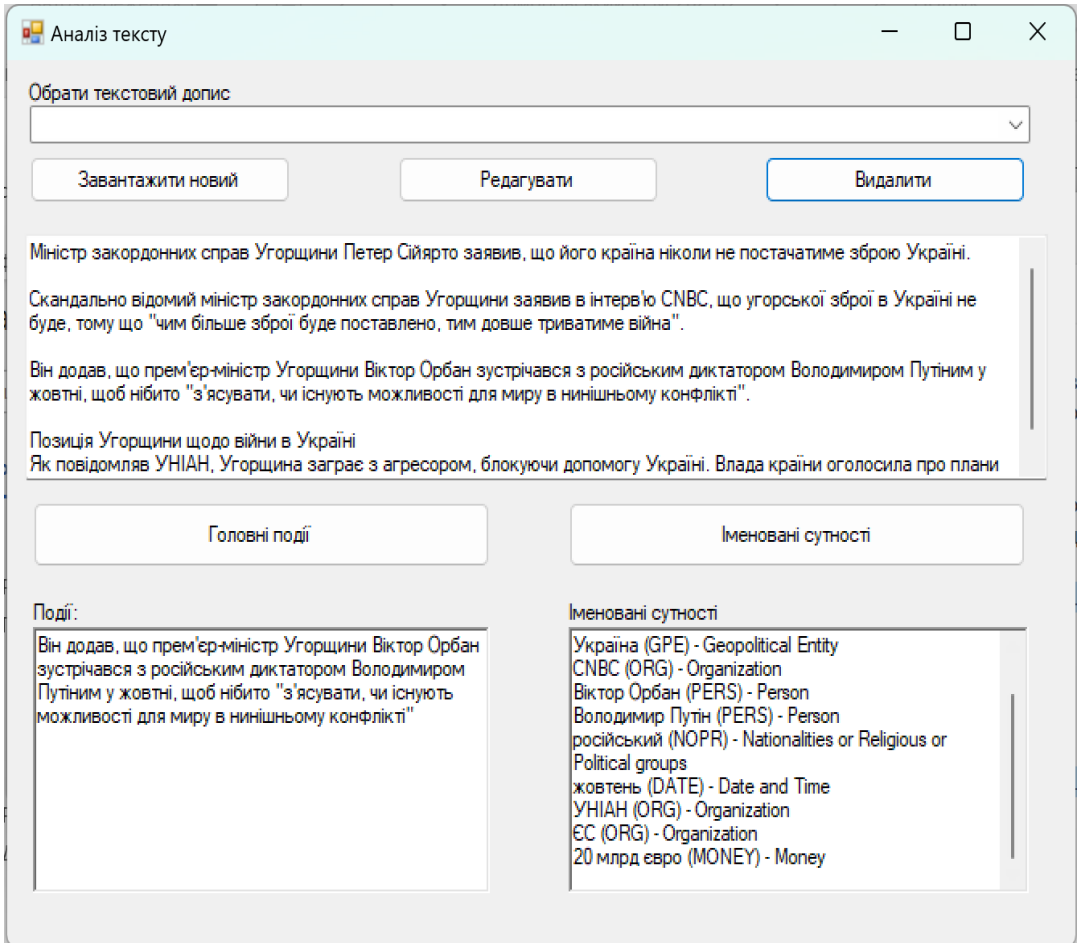


Рисунок 4.6 – Результат виконання тест-кейсу

Таким чином, було проведено прикладне тестування інформаційної системи визначення тональності текстової інформації по відношенню до іменованих сутностей було реалізовано ряд тест-кейсів, що підтверджують можливість програмного застосунку визначати ключові терміни за показником дисперсії, іменовані сутності, що містяться в тексті та знаходити ключові події в тексті, виводячи їх на екран.

4.4 Особливості використання інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови

Було створено програмне забезпечення для автоматизованої ідентифікації подій в україномовних текстах засобами обробки природної мови. Робота

користувача починається із головного вікна застосунку, де користувач може обрати текст, що необхідно дослідити (рисунок 4.7).

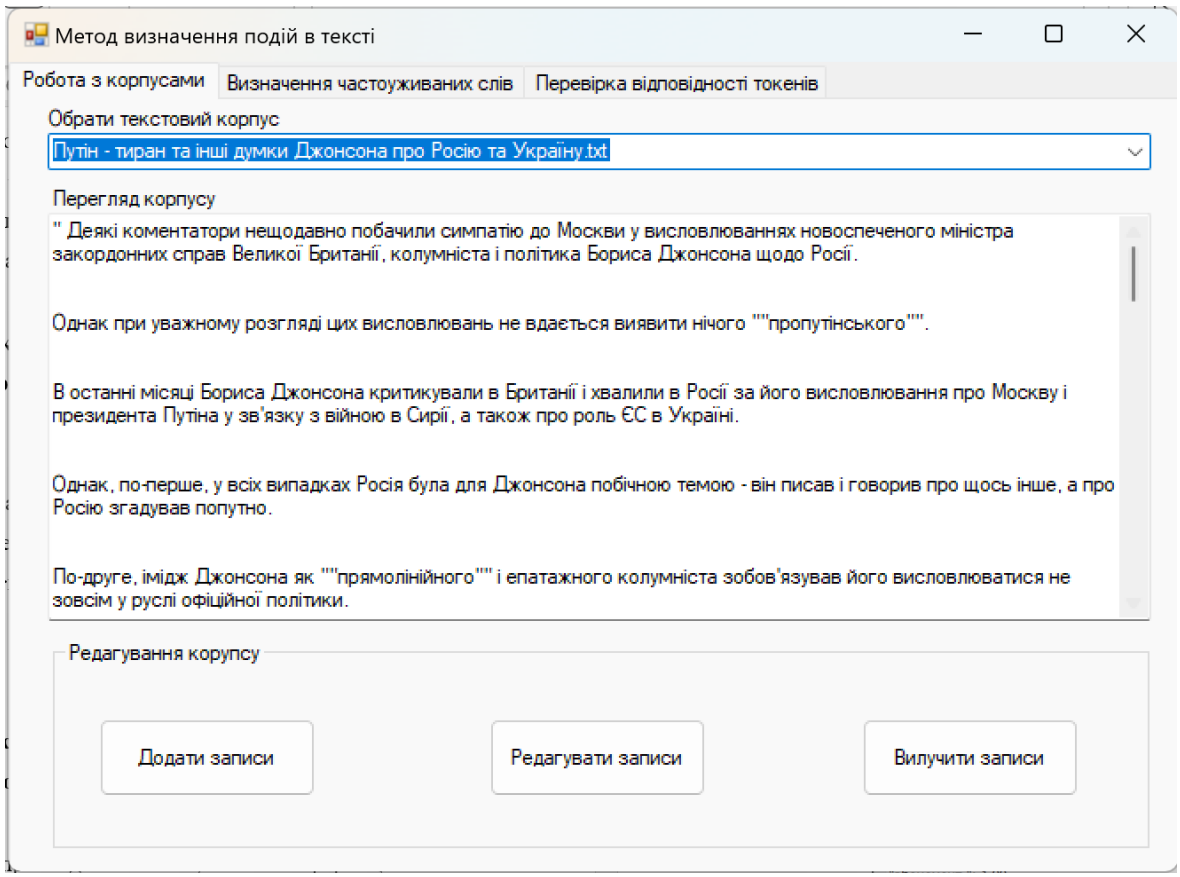


Рисунок 4.7 – Головне вікно застосунку

На цій формі реалізовано функціонал інформаційної системи ідентифікації подій в україномовних текстах роботи з текстами, їх редагування, видалення та збереження.

Програмний застосунок містить три вкладки на головному вікні:

– робота з корпусами;

– визначення частовживаних слів;

– перевірка відповідності токенів.

Далі користувач може перейти до вкладки «Визначення частовживаних слів». Натиснувши на кнопку «Визначити дисперсійну оцінку для слів», користувачеві пропонується переглянути частовживані слова (рисунок 4.8).

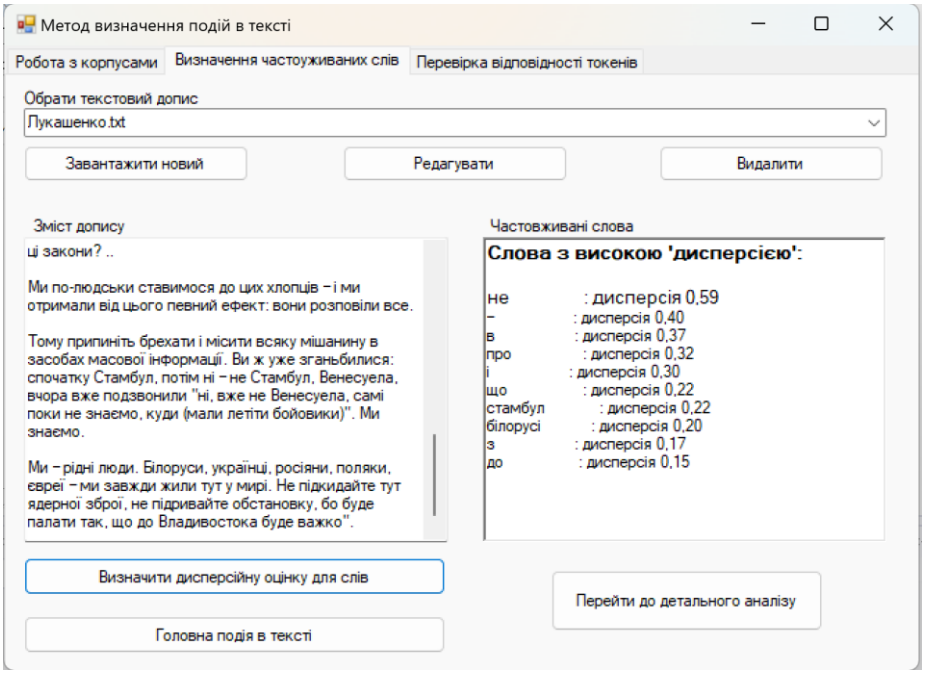


Рисунок 4.8 – Робота із формою «Використання частовживаних слів»

На цій формі користувач інформаційної системи ідентифікації подій в україномовних текстах може натиснути кнопку «Головна подія в тексті», таким чином відкривши форму «Аналіз тексту», де реалізована можливість визначення головної події, що прослідковується в тексті (рисунок 4.9).

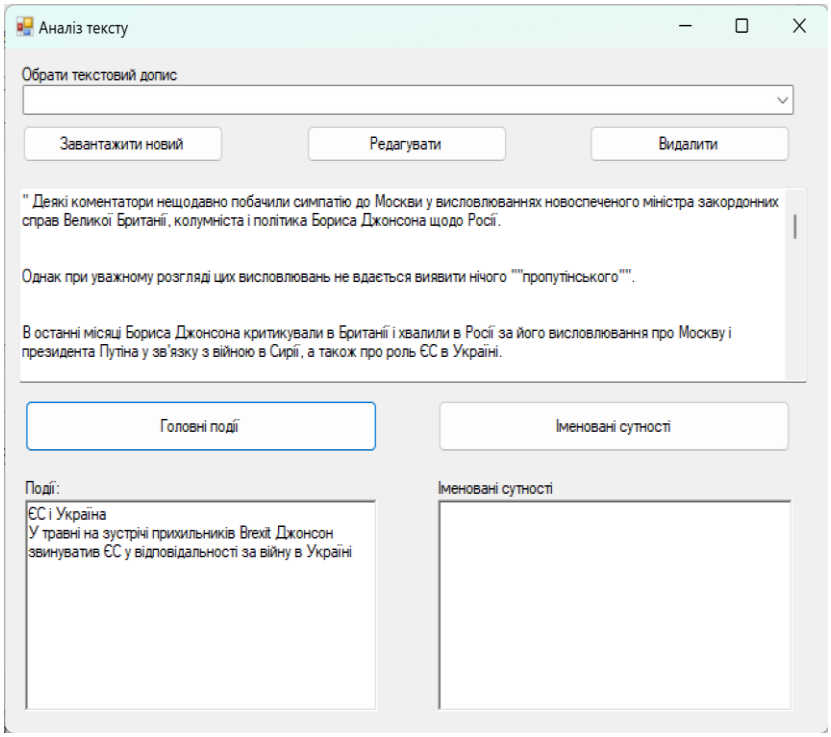


Рисунок 4.9 – Робота із вікном застосунку «Аналіз тексту»

Таким чином, обравши текст, програмний застосунок може повернути головну подію в тексті та вивести її на екран за допомогою методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Також реалізована форма «Перевірка відповідності токенів», де користувач має змогу ввести слово чи декілька слів та перевірити, чи мають вони відношення до тексту (рисунок 4.10).

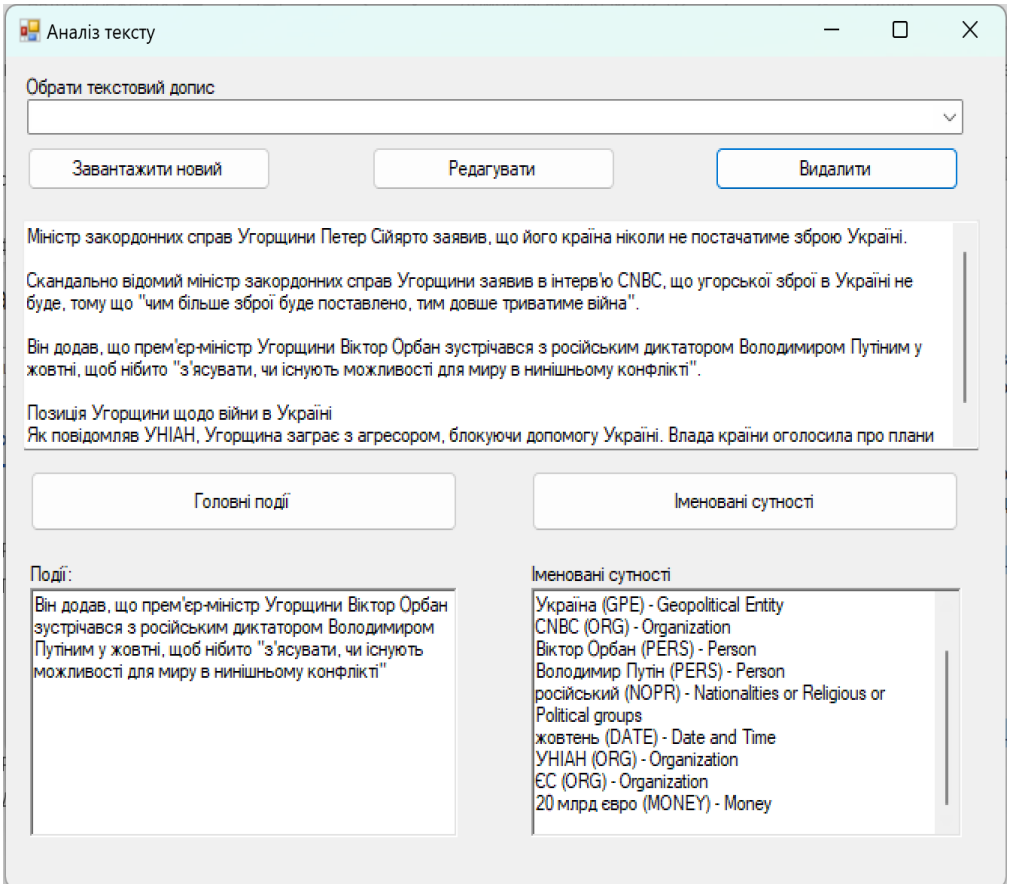


Рисунок 4.10 – Робота з формою «Аналіз тексту»

Інтерфейс форми для перевірки відповідності токенів до тексту наведено на рисунку 4.11. На даній формі наведено функціонал для:

– роботи з корпусами текстів: вибір тексту, редагування та видалення при необхідності;

– визначення та перегляд подій в тексті;

– визначення відповідності вхідного токену та висновку.

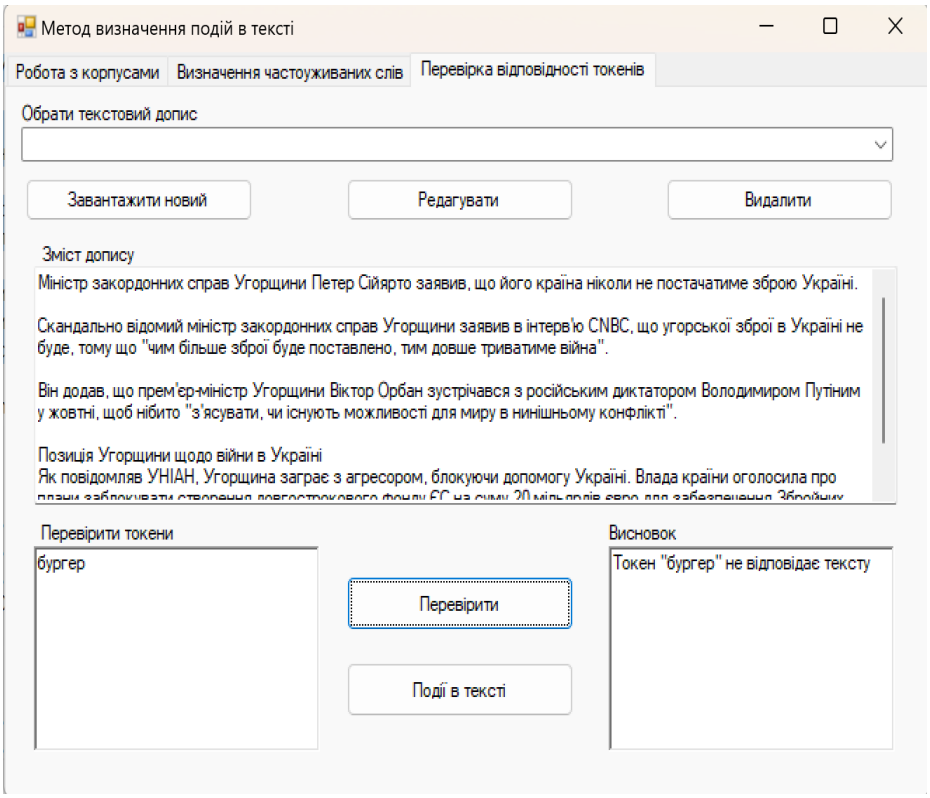


Рисунок 4.11 – Робота вкладки «Перевірка відповідності токенів»

Таким чином, було реалізовано програмний застосунок на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови. Програмний застосунок містить форми та вкладки, що повною мірою відповідають архітектурі інформаційної системи.

4.5 Дослідження ефективності методу ідентифікації подій в україномовних текстах засобами обробки природної мови

Дослідження ефективності інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови було спрямоване на оцінку точності та надійності розробленого програмного продукту. В основу дослідження було взято 30 текстів, які були отримані з популярних українських інформаційних ресурсів, таких як ukr.net [39] та hromadske [40]. Ці тексти охоплювали різні теми та були актуальними для аналізу сучасних новинних потоків.

Для оцінки ефективності розробленого програмного застосунку було використано метод порівняння результатів, отриманих програмним продуктом, з оцінками експерта та ChatGPT (версія 3.5) [41]. Експерт та ChatGPT, переглядав кожен текст, оцінюючи, наскільки подія, ідентифікована реалізованим програмним застосунком, відповідає дійсному змісту події в тексті. Проаналізувавши текст, експерт визначає оцінку у відсотках, наскільки результат роботи програми відповідає дійсному змісту подій в тексті.

На рисунку 4.12 наведено алгоритм оцінювання ефективності інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Першим кроком є завантаження текстового допису для оцінки експертом у реалізовану інформаційну систему на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови для подальшої оцінки експертом відповідності отриманого результату роботи застосунку.

Крок 2 передбачає виведення отриманого результату на екран для оцінки експертом.

Крок 3 – оцінка експертом, наскільки результат роботи програми відповідає реальним подіям, що описуються в тексті. На цьому кроці користувач вводить відсоткове значення, наскільки події, визначені в тексті відповідають реальним. Оцінка вводиться у відсотках.

Заключний крок – формування висновку щодо коректності роботи програмного застосунку за шкалою від 0 до 100 відсотків, де значення, наближені до 0 означають невідповідність отриманого результату до події, що описується в тексті, а значення, наближені до 100, відповідно, означають відповідність отриманого результату до події, що описується в тексті.

Після завантаження тексту, експерт порівняв результати програмного застосунку зі своїми висновками та визначив відсоток відповідності. Цей показник відображає, наскільки точно програмний продукт визначив події та тональність у порівнянні з експертною оцінкою.

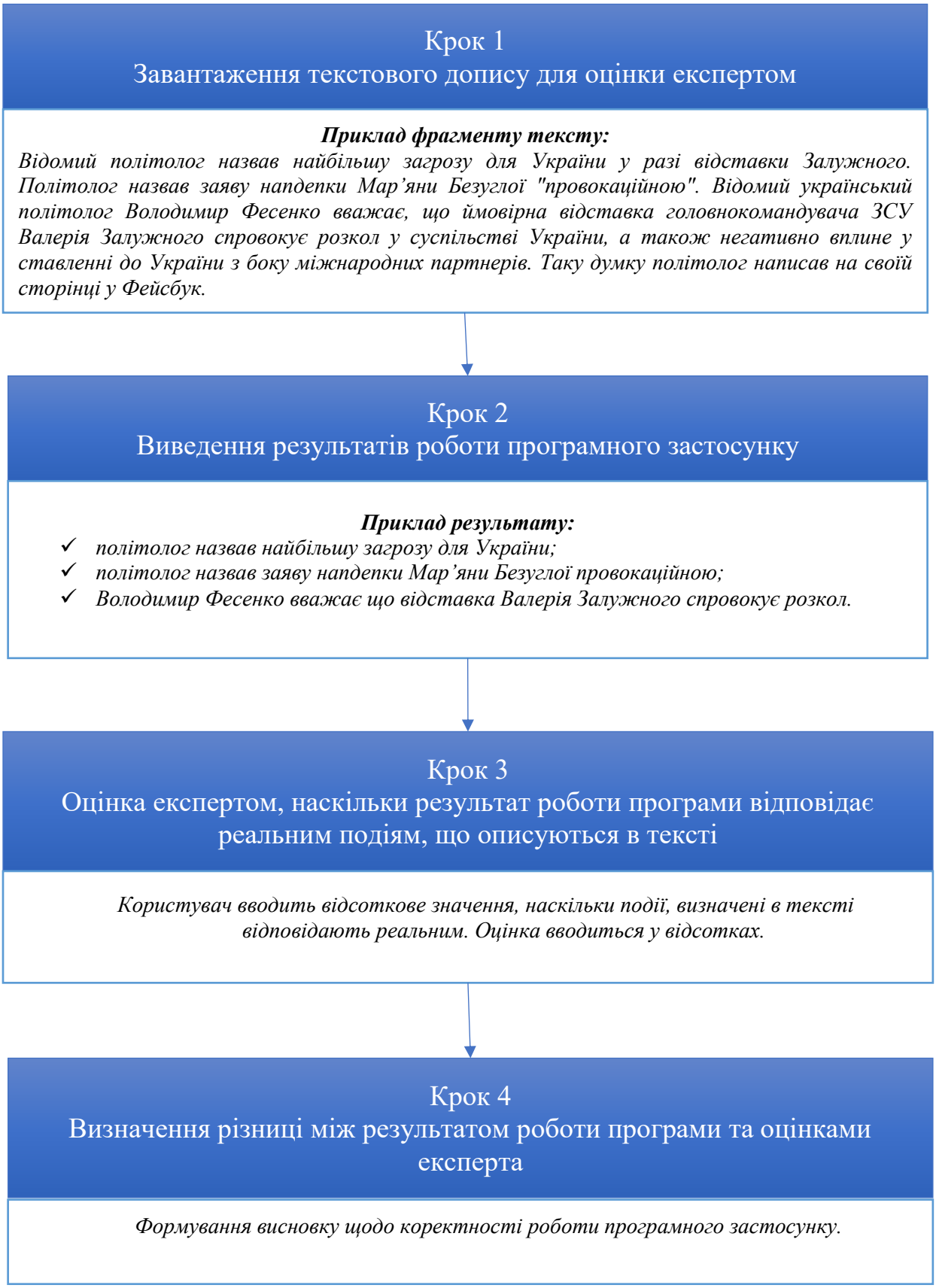


Рисунок 4.12 – Алгоритм оцінювання ефективності інформаційної системи на базі методу ідентифікації подій в україномовних текстах

Високий відсоток відповідності свідчить про ефективність розробленого методу та його придатність для аналізу текстової інформації. На рисунку 4.13 наведено результати, отримані за допомогою реалізованого програмного застосунку щодо подій в тексті.

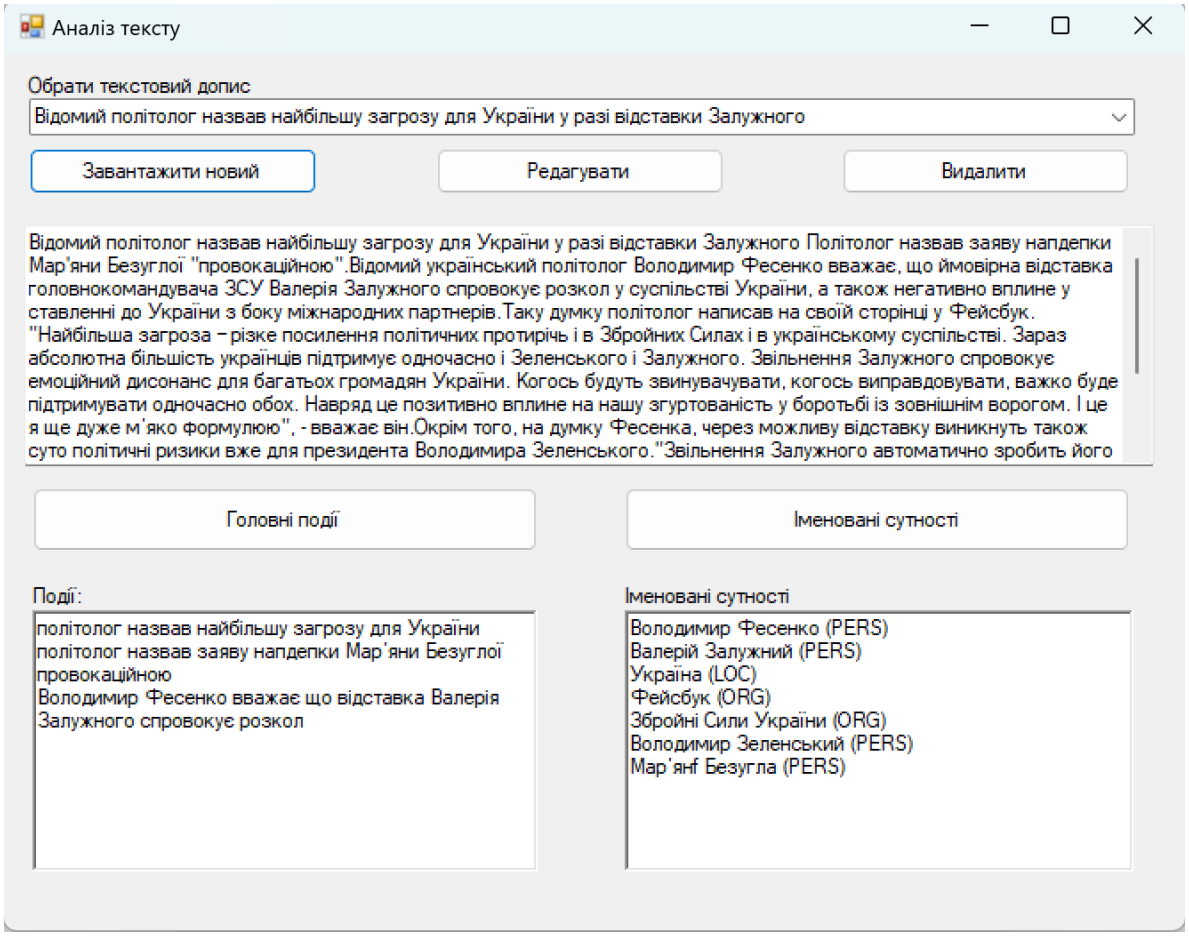


Рисунок 4.13 – Отримані допомогою реалізованого програмного застосунку щодо подій в тексті

На рисунку 4.14 наведено діаграму із результатами оцінювання ефективності інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови експертом. На графіку представлено оцінки, які виставив експерт при оцінці результатів роботи програми, порівнюючи їх та дійсні події, про які написано в тексті. Було проаналізовано 30 текстів українською мовою, отриманих з ресурсів новин.

Рисунок 4.14 – Діаграма із результатами оцінювання ефективності інформаційної системи на базі методу ідентифікації подій за участі експерта

Також відповідне тестування проводилось за допомогою ChatGPT, версії 3.5 [41]. Результати проведення дослідження за допомогою ChatGPT наведено на рисунку 4.15.

Рисунок 4.15 – Діаграма із результатами оцінювання ефективності інформаційної системи на базі методу ідентифікації подій за участі ChatGPT

Також було сформовано таблицю (таблиця 4.4), в якій продемонстровано отримані результати та різницю значень між оцінкою, що визначив експерт та GPT-3.5. З результатів та відповідного графіка (рисунок 4.16) можна побачити, що більша частина подій в текстових дописах, була ідентифікована із високою точністю за допомогою реалізованого програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Таблиця 4.4 – Результати та різниці значень між експертними оцінками

Текст, №	**Оцінка експерта, %**	**Оцінка ChatGPT, %**	**Різниця значень, модуль**	**Текст, №**	**Оцінка експерта, %**	**Оцінка ChatGPT, %**	**Різниця значень, модуль, %**
1	60	65	5	**16**	85	90	5
2	65	70	5	**17**	70	80	10
3	70	75	5	**18**	75	80	5
4	65	75	10	**19**	80	80	0
5	75	80	5	**20**	65	70	5
6	80	85	5	**21**	70	70	0
7	85	80	5	**22**	75	75	0
8	75	70	5	**23**	80	80	0
9	70	65	5	**24**	80	80	0
10	100	100	0	**25**	75	80	5
11	75	80	5	**26**	65	60	5
12	85	80	5	**27**	70	65	5
13	80	85	5	**28**	75	80	5
14	75	80	5	**29**	80	85	5
15	70	75	5	**30**	80	85	5

Візуалізація даних таблиці 4.4 наведена на рисунку 4.17, де продемонстровано, що значна частина дописів була коректно ідентифікована програмним застосунком, а різниця між оцінками експерта та GPT-3.5 незначна.

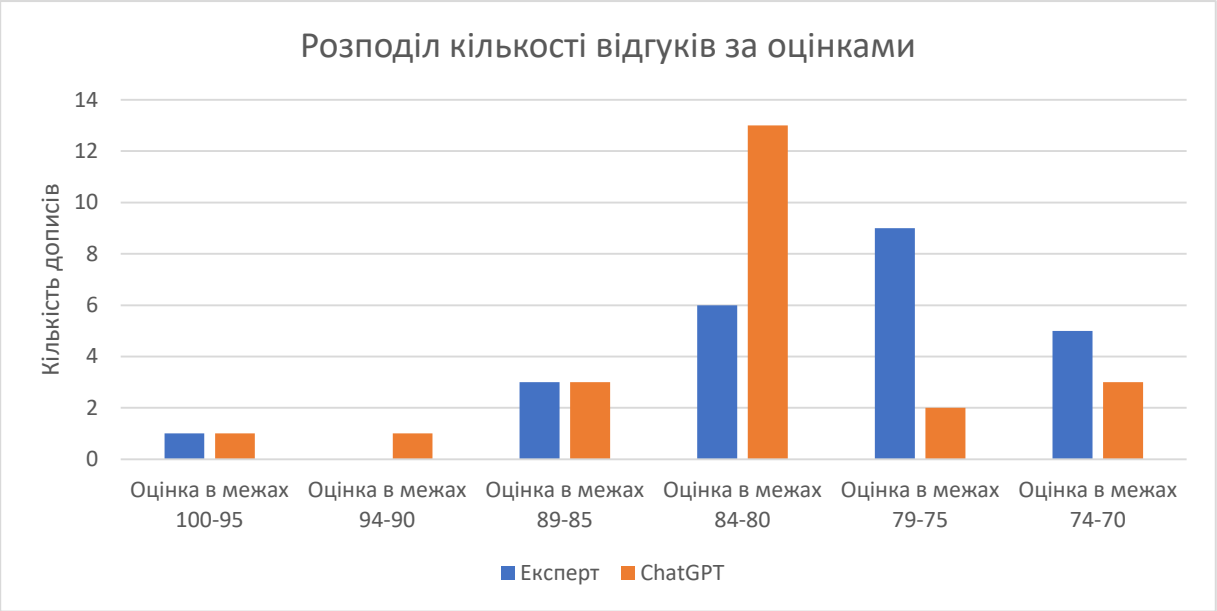


Рисунок 4.16 – Розподіл кількості правильно ідентифікованих подій в дописах за оцінками

Рисунок 4.17 – Результати проведення дослідження ефективності

Отже, розроблена інформаційна система автоматизованого визначення подій в україномовних текстах засобами обробки природної мови є ефективною та допомагає коректно визначати в текстових дописав події та виводити відповідний текст користувачеві. Система спроможна коректно аналізувати та

ідентифікувати ключові події, виявляючи іменовані сутності, тематичні ключові слова та контекстуальні зв’язки в текстах.

Завдяки своїй гнучкості та масштабованості, розроблена інформаційна система може бути використана у широкому спектрі застосувань, від журналістики та досліджень у сфері соціальних наук до моніторингу громадської думки та аналізу соціальних тенденцій. Це відкриває нові можливості для збору та аналізу даних, що може впливати на прийняття рішень у багатьох сферах.

Висновки до розділу 4

В результаті виконання розділу було проведено дослідження ефективності інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови, що за допомогою визначення ключових слів шляхом розрахунку дисперсійної оцінки, визначення іменованих сутностей за допомогою нейронної мережі Stanza та використання словника термінів подій визначає ключові події в тексті та виводить відповідний текст користувачеві.

В рамках виконання розділу було описано програмну архітектуру інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови, описано основні компоненти для реалізації програмного продукту.

Було проведено дослідження із залученням експерта та GPT-3.5 в ролі експерта для оцінки якості отриманих результатів. Розроблена інформаційна система автоматизованого визначення подій в україномовних текстах засобами обробки природної мови є ефективною та допомагає коректно визначати в текстових дописав події та виводити відповідний текст користувачеві. Система спроможна коректно аналізувати та ідентифікувати ключові події, виявляючи іменовані сутності, тематичні ключові слова та контекстуальні зв’язки в текстах.

Загальні висновки

Кваліфікаційна робота магістра розв’язує задачу ідентифікації подій в україномовних текстах засобами обробки природної мови, що дає можливість за вхідними даними у вигляді україномовного тексту одержувати вихідні дані у вигляді переліку виокремлених іменованих сутностей подій та сформованих текстових описів подій. Для досягнення мети використовується дисперсійне оцінювання важливості слів і нейромережева модель Stanza для виокремлення іменованих сутностей та формування словника термінів подій і формування речень з опису події. Також було створено відповідну програмну реалізацію для апробації методу.

В результаті виконання роботи було розроблено інформаційну систему ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового допису перетворити у вихідні дані у вигляді тексту, що найточніше відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій.

Проведені дослідження ефективності розробленого методу ідентифікації подій в україномовних текстах засобами обробки природної мови з використанням розробленої відповідної інформаційної системи. Розроблена інформаційна система автоматизованого визначення подій в україномовних текстах засобами обробки природної мови є ефективною та допомагає коректно визначати в текстових дописав події та виводити відповідний текст користувачеві. Система спроможна коректно аналізувати та ідентифікувати ключові події, виявляючи іменовані сутності, тематичні ключові слова та контекстуальні зв’язки в текстах.

У результаті виконання роботи поставлено та *вирішено наступні завдання*:

1. Досліджено предметну область ідентифікації подій в україномовних текстах засобами обробки природної мови.

2. Досліджено існуючі методи та засоби ідентифікації подій в україномовних текстах засобами обробки природної мови.

3. Створено метод ідентифікації подій в україномовних текстах засобами обробки природної мови.

4. Спроєктувано інформаційну систему на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови..

5. Обрано засоби розробки для спроєктованої архітектури інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

6. Розроблено відповідну програмну реалізацію методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

7. Досліджено практичну ефективність застосування методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

Під час виконання роботи були одержано результати, що містять інновації й наукову новизну, зокрема було реалізовано метод ідентифікації подій в україномовних текстах засобами обробки природної мови.

Особливістю реалізованого методу є його здатність повертати множину ідентифікованих в тексті іменованих сутностей, а за допомогою поєднання дисперсійної оцінки для визначення ключових термінів та сформованого словника термінів подій, програмний застосунок на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови, повертає текст, що містить подій, про які описано в тексті.

Для дослідження практичної ефективності застосування методу ідентифікації подій в україномовних текстах засобами обробки природної мови, було реалізовано прикладну реалізацію методу ідентифікації подій в україномовних текстах засобами обробки природної мови. Система дозволяє аналізувати та ідентифікувати ключові події, виявляючи іменовані сутності, тематичні ключові слова та контекстуальні зв’язки в текстах.

Розроблена інформаційна система дозволяє для визначеного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового допису перетворити у вихідні дані у вигляді тексту, що найточніше відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій.

Напрямком практичного використання розробленого методу й інформаційної системи є автоматизація визначення подій в україномовних текстах з використанням технологій обробки природної мови, знаходить своє застосування в широкому спектрі областей, від медіа-аналітики та журналістики до наукових досліджень та моніторингу громадської думки. Реалізований метод може бути використаний для автоматичного збору та аналізу новин, соціальних медіа, блогів, забезпечуючи оперативне виявлення та відслідковування соціально значущих подій, тенденцій та настроїв у суспільстві.

Основні наукові й практичні результати кваліфікаційної роботи магістра було опубліковано в тезах XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023» [42].

Перелік посилань

1. Shankar Venkatesh, Sohil Parsana. «An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing.» Journal of the Academy of Marketing Science 50.6, 2022, p. 1324-1350

2. Hupkes Dieuwke, «State-of-the-art generalisation research in NLP: a taxonomy and review». arXiv preprint arXiv:2210.03050,2022

3. Ribeiro Marco Tulio, Scott Lundberg. «Adaptive testing and debugging of NLP models». Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1, 2022

4. Vivek Will. «A Novel Technique for User Decision Prediction and Assistance Using Machine Learning and NLP: A Model to Transform the E-commerce System». Big Data Management in Sensing; River Publishers: Aalborg, Denmark, 2022

5. Ganesh Ananya. «Response Construct Tagging: NLP-Aided Assessment for Engineering Education». Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications, BEA-2022, 2022

6. Liu Zhengliang. «Survey on natural language processing in medical image analysis». Journal of Central South University. Medical Sciences,202

7. Camacho-Collados Jose. «Tweetnlp: Cutting-edge natural language processing for social media». arXiv preprint arXiv:2206.14774, 2022

8. Пальчевська О.С., Лучик А.А. Мовна різноманітність в українському національному лінгвістичному корпусі, 2022

9. Hupkes Dieuwke. «State-of-the-art generalisation research in NLP: a taxonomy and review». arXiv preprint arXiv:2210.03050, 2022

10. Liu Shengyua. «Data-driven event detection of power systems based on unequal-interval reduction of PMU data and local outlier factor». IEEE Transactions on Smart Grid 11.2, 2019

11. Guzman-Nateras Luis, Minh Van Nguyen, Thien Nguyen. «Cross-lingual event detection via optimized adversarial training». IEEE Transactions on Smart Grid 11.2, 2019

12. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022

13. Kim Sang-Woon, Joon-Min Gil. «Research paper classification systems based on TF-IDF and LDA schemes». Human-centric Computing and Information Sciences 9, 2019

14. Pulis Michael, Joel Azzopardi, Jeffrey Micallef. «Intelligent Artificial Agent for Information Retrieval». International Conference on Practical Applications of Agents and Multi-Agent Systems. Cham: Springer International Publishing, 2022

15. Karita Shigeki. «A comparative study on transformer vs rnn in speech applications». 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019

16. Kenton Jacob Devlin, Ming-Wei Chang, Lee Kristina Toutanova. «Bert: Pre-training of deep bidirectional transformers for language understanding». Proceedings of naacL-HLT. Vol. 1. 2019

17. Scopus. Scopus Preview. URL: https://www.scopus.com/home.uri

18. Google Scholar. Google Scholar Search. URL: https://scholar.google.com

19. Grabar Natalia; Hamon Thierry Automatic detection of temporal information in Ukrainian general-language texts. Automatic detection of temporal information in Ukrainian general-language texts, 2018

20. Vysotska Victoria. «Intelligent Analysis of Ukrainian-language Tweets for Public Opinion Research based on NLP Methods and Machine Learning Technology»

21. Prokipchuk O., Vysotska V., Pukach P. «Intelligent Analysis of Ukrainian-language Tweets for Public Opinion Research based on NLP Methods and Machine Learning Technology». International Journal of Modern Education and Computer Science, 15(3), 70-93. doi: 10.5815/ijmecs.2023.03.06

22. Hume AI. Getting started. URL: https://hume.ai/

23. Komprehend.io. Downloads. URL: https://komprehend.io/

24. GitHub. Personal Events in Dialogue Corpus Dataset. URL: https://metatext.io/datasets/personal-events-in-dialogue-corpus

25. Dong Hao-Wen. «Achromatic metasurfaces by dispersion customization for ultra-broadband acoustic beam engineering». National Science Review 9.12, 2022

26. Stanza Online. Test the model. URL: http://stanza.run/

27. GitHub. Браунський корпус української мови. URL: https://github.com/brown-uk/corpus

28. Stanza Stanford NLP. Stanza Documentation URL: https://github.com/stanfordnlp/stanza

29. Python TM. Python Downloads. URL: https://www.python.org/doc

30. PyTorch.PyTorch Get Started. URL: https://pytorch.org/get-started/locally/

31. Numpy.org. Numpy downloads. URL: https://numpy.org/

32. NET Platform. NET functions overview. URL: https://openslime.it/2020/03/net-5-preview

33. Medium.com. Visual Studio 2019 goes live with C++, Python shared editing. URL: https://arstechnica.com/gadgets/2019/04/visual-studio-2019-goes-live-with-c-python-shared-editing/

34. IronPython. IronPython overview. URL:https://github.com/cjhutto/vaderSentiment

35. GitHub. Implementation of Python 3.x for .NET Framework that is built on top of the Dynamic Language Runtime. URL: https://github.com/IronLanguages/ironpython3

36. GitHub. IronPython 2.7.12. URL: https://github.com/IronLanguages/ironpython2/releases

37. Rose Anthony, Scott Graham, Jacob Krasnov. «IronNetInjector: Weaponizing. NET Dynamic Language Runtime Engines». Digital Threats: Research and Practice, 2023

38. Srivaranon Suchart. «Autonomous Parametric and Machine Learning Estimation Models Revolutionize Capital Expenditure Process of Engineering and Construction Project in Pttep». Abu Dhabi International Petroleum Exhibition and Conference. SPE, 2023

39. UKR.NET. Новини. URL: https://www.ukr.net/

40. Hromadske.com. Головна. URL: https://hromadske.ua/

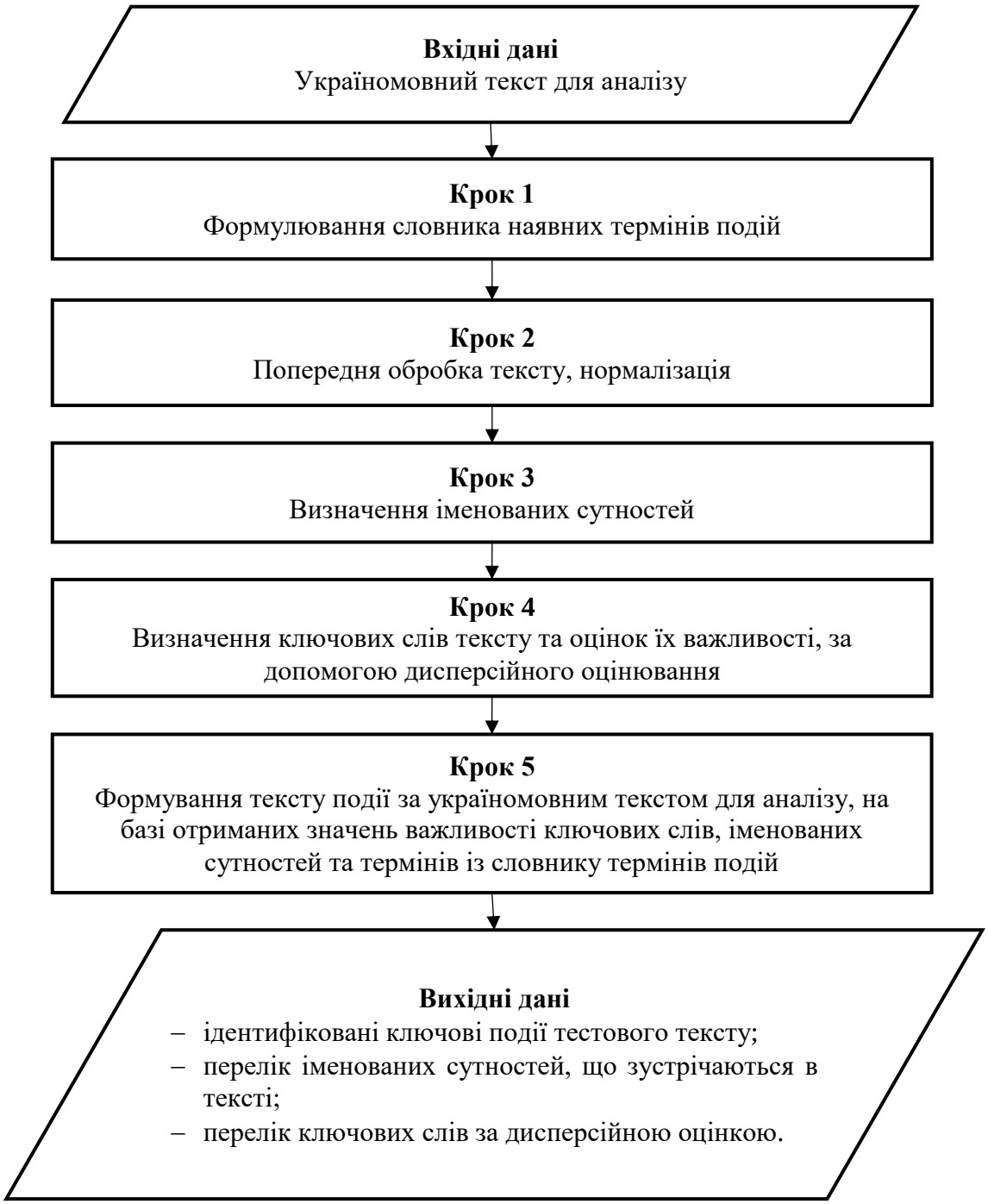
41. OpenAI. ChatGPT-3.5. URL: https://chat.openai.com/

42. Домбровський Н.С., Скрипник Т.К., Вознюк Л.О. Метод ідентифікації подій в україномовних текстах засобами обробки природної мови. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 80-83.

ДОДАТКИ

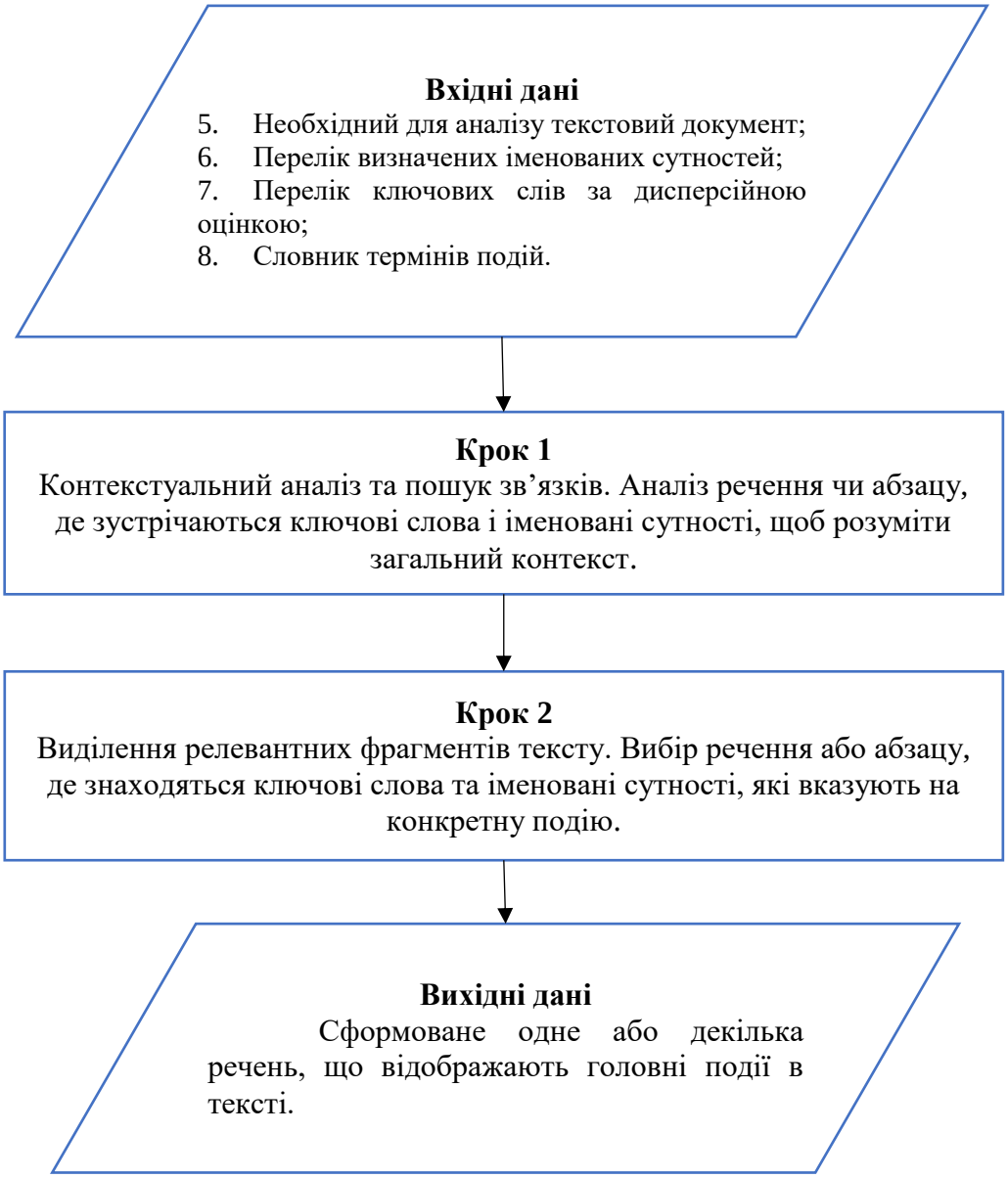
Додаток А

Схема методу ідентифікації подій в україномовних текстах засобами обробки природної мови



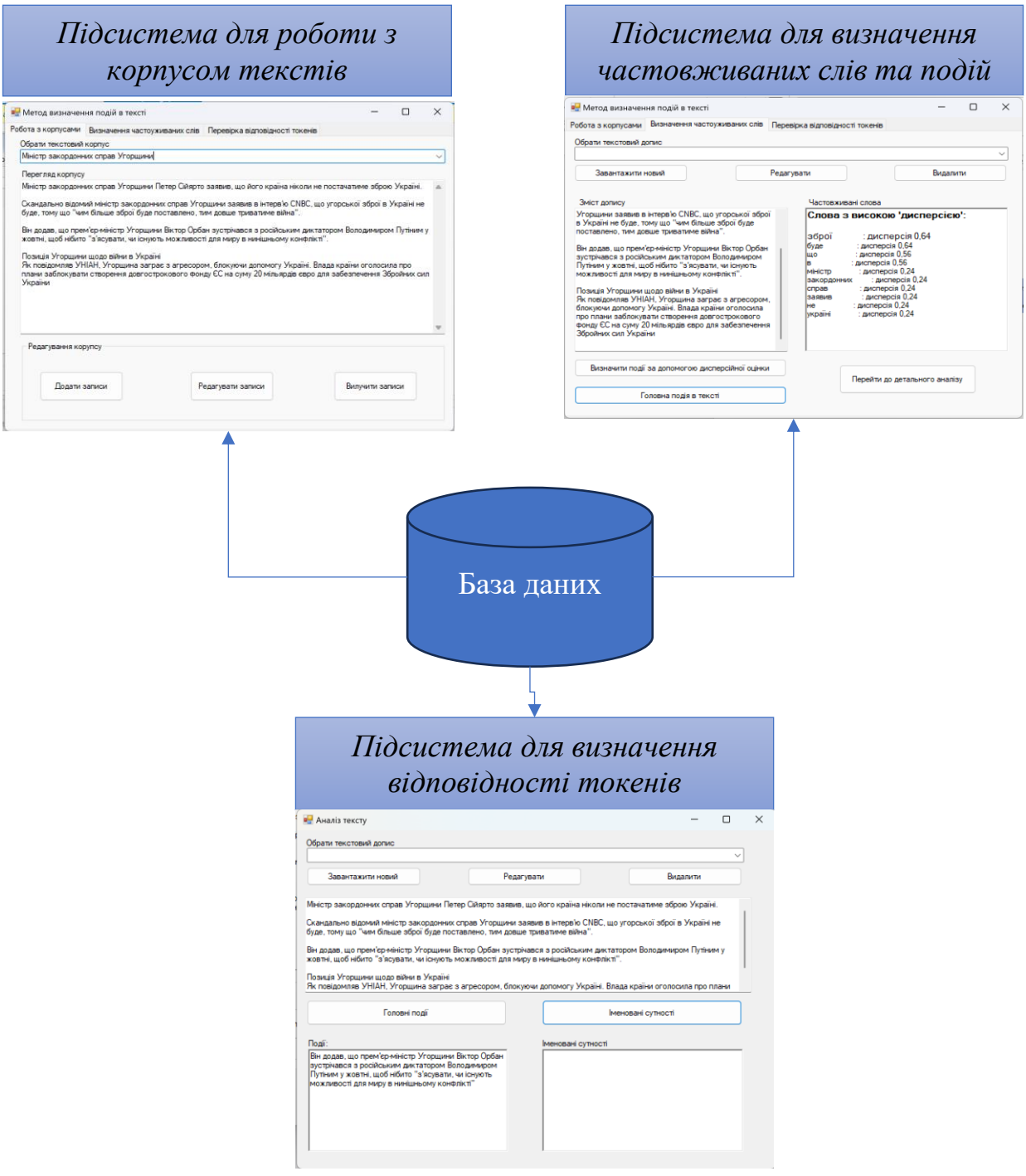
Додаток Б

Схема процесу формування тексту ключових подій



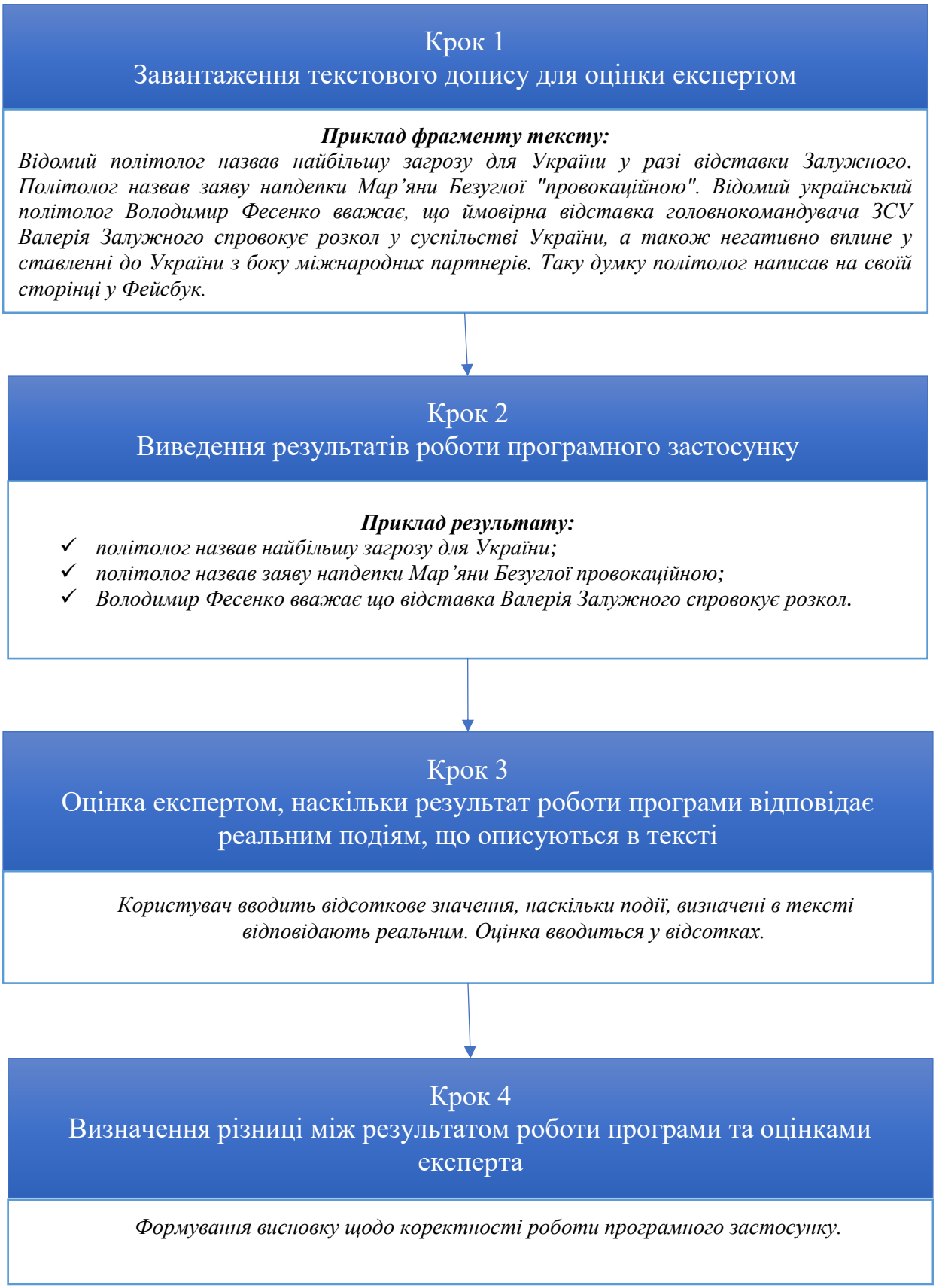
Додаток В

Схема інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови



Додаток Г

Алгоритм оцінювання ефективності інформаційної системи ідентифікації подій в україномовних текстах



Додаток Д

Результати та різниці значень між експертними оцінками

Текст, №	Оцінка експерта, %	Оцінка ChatGPT, %	Різниця значень, модуль	Текст, №	Оцінка експерта, %	Оцінка ChatGPT, %	Різниця значень, модуль, %
1	60	65	5	**16**	85	90	5
2	65	70	5	**17**	70	80	10
3	70	75	5	**18**	75	80	5
4	65	75	10	**19**	80	80	0
5	75	80	5	**20**	65	70	5
6	80	85	5	**21**	70	70	0
7	85	80	5	**22**	75	75	0
8	75	70	5	**23**	80	80	0
9	70	65	5	**24**	80	80	0
10	100	100	0	**25**	75	80	5
11	75	80	5	**26**	65	60	5
12	85	80	5	**27**	70	65	5
13	80	85	5	**28**	75	80	5
14	75	80	5	**29**	80	85	5
15	70	75	5	**30**	80	85	5

Додаток Е

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

(ксерокопії титульної сторінки, сторінки змісту та всіх сторінок із публікацією)

Наукова публікація:

Домбровський Н.С., Скрипник Т.К., Вознюк Л.О. Метод ідентифікації подій в україномовних текстах засобами обробки природної мови. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». Хмельницький, 2023. с. 80-83.

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ
за матеріалами XV Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2023»

17-18 листопада 2023

Хмельницький 2023

УДК 004.4

Домбровський Н.С., Скрипник Т.К., Вознюк Л.О.

Хмельницький національний університет

МЕТОД ІДЕНТИФІКАЦІЇ ПОДІЙ В УКРАЇНОМОВНИХ ТЕКСТАХ ЗАСОБАМИ ОБРОБКИ ПРИРОДНОЇ МОВИ

В даному дослідженні проводиться оцінка, наскільки текст тестового документа відповідає введеним користувачем ключовим токенам з урахуванням оцінки важливості слів у тексті. Для визначення цієї важливості можуть використовуватися два підходи: дисперсійна оцінка та метод BM-25. Передбачається автоматизований процес формування синонімічних рядів до введених користувачем токенів з метою розширення можливостей методу.

In this study, we evaluate the extent to which the text of the test document corresponds to the key tokens entered by the user, taking into account the importance of words in the text. To determine this importance, two approaches can be used: variance estimation and the VM-25 method. An automated process of forming synonymous series to the user-entered tokens is envisaged to expand the capabilities of the method.

Обробка природної мови (Natural Language Processing, NLP) – це міждисциплінарна галузь інформатики та мовознавства, що вивчає та розробляє методи та технології для взаємодії між комп'ютерами та людьми через природну мову. Вона включає в себе комплексний аналіз та обробку тексту та мови з метою автоматизації розуміння та генерації текстової інформації [1].

Метод ідентифікації подій в україномовних текстах засобами обробки природної мови є актуальною задачею в галузі обчислювальної лінгвістики та штучного інтелекту. Цей метод спрямований на автоматичне визначення подій, які відбуваються в тексті, зокрема, діяльності, подій, процесів та їхніх атрибутів. Ідентифікація подій є важливою для багатьох застосувань, включаючи аналіз новин, моніторинг соціальних мереж, розробку систем автоматичного розуміння тексту, аналізу семантики та багато інших областей [2].

Для специфічних задач NLP іноді використовуються поєднання декількох методів та моделей, а також аналіз контексту та зв'язків між словами та фразами. Застосування цих математичних методів дозволяє здійснювати складний аналіз текстів і виявляти події та інформацію, пов'язану з ними, що може бути корисним в різних додатках, включаючи аналітику соціальних медіа, пошукові системи, моніторинг новин тощо [3].

Для поставленого завдання було обрано метод дисперсійної оцінки, це статистичний метод для визначення різноманітності та варіабельності даних у

вибірці. Дисперсія грає важливу роль у багатьох аналітичних задачах, включаючи оцінку важливості слів у текстових документах для подальшого використання в методах обробки природної мови, таких як ідентифікація подій.

Дисперсійна оцінка базується на обчисленні середнього квадратичного відхилення (стандартного відхилення) даних від їхнього середнього значення. Вона вимірює, наскільки дані розподілені відносно середнього значення. Вища дисперсія вказує на більшу розкиданість даних, тоді як нижча дисперсія вказує на менший розкид [4].

Формула для обчислення дисперсії наведено у формулі:

$$\sigma^2 = \frac{\sum(xi - \mu)^2}{N}$$

де σ^2 – дисперсія, xi – кожне окреме значення вибірки, μ – середнє значення (середнє арифметичне) вибірки, N – кількість значень вибірки.

Основна ідея полягає в тому, що дисперсія вимірює, наскільки середнє значення відхиляється від кожного окремого значення вибірки. Велика дисперсія вказує на те, що дані мають великий розкид, тоді як мала дисперсія означає, що значення вибірки майже однакові.

Дисперсійна оцінка може бути важливим інструментом в контексті ідентифікації подій в текстах. Вона може використовуватися для визначення важливості слів у текстових документах, які можуть вказувати на ключові терміни або інформацію, що вказує на певні події чи теми [5]. Розрахунок дисперсійної оцінки може допомогти виокремити найбільш значущі слова, які слід аналізувати подальше в контексті ідентифікації подій.

При розробці методу ідентифікації подій в україномовних текстах за допомогою обробки природної мови, одним з важливих виборів є вибір підходу для оцінки важливості слів та токенів у текстах. У контексті поставленої задачі, дисперсійна оцінка виявляється більш вдалою стратегією. Дисперсійна оцінка більше ураховує різноманітність слів та токенів у тексті. Це означає, що вона надає важливість словам, які вносять різноманітність та багатогранність в текст, що є важливим аспектом ідентифікації подій. Таким чином, дисперсійна оцінка відображає різноманітність тексту, що може допомогти виокремити ключові слова та фрази, пов'язані з подіями.

Для реалізації методу ідентифікації подій в україномовних текстах необхідно чітко окреслити послідовність його виконання. Схема роботи методу наведена на рисунку 1.

Вхідні дані. На цьому кроці вхідні дані включають текстовий документ, який піддається обробці, і список токенів, які вводить користувач. Текстовий документ представляє собою послідовність слів або фраз, які містять інформацію про події.

Токенізація тексту. Токенізація – це процес розбиття тексту на окремі токени (слова, фрази, символи тощо). Цей процес виконується для підготовки

тексту до подальшого аналізу. Токени представлені як послідовність символів, і кожен токен має свій внутрішній ідентифікатор.

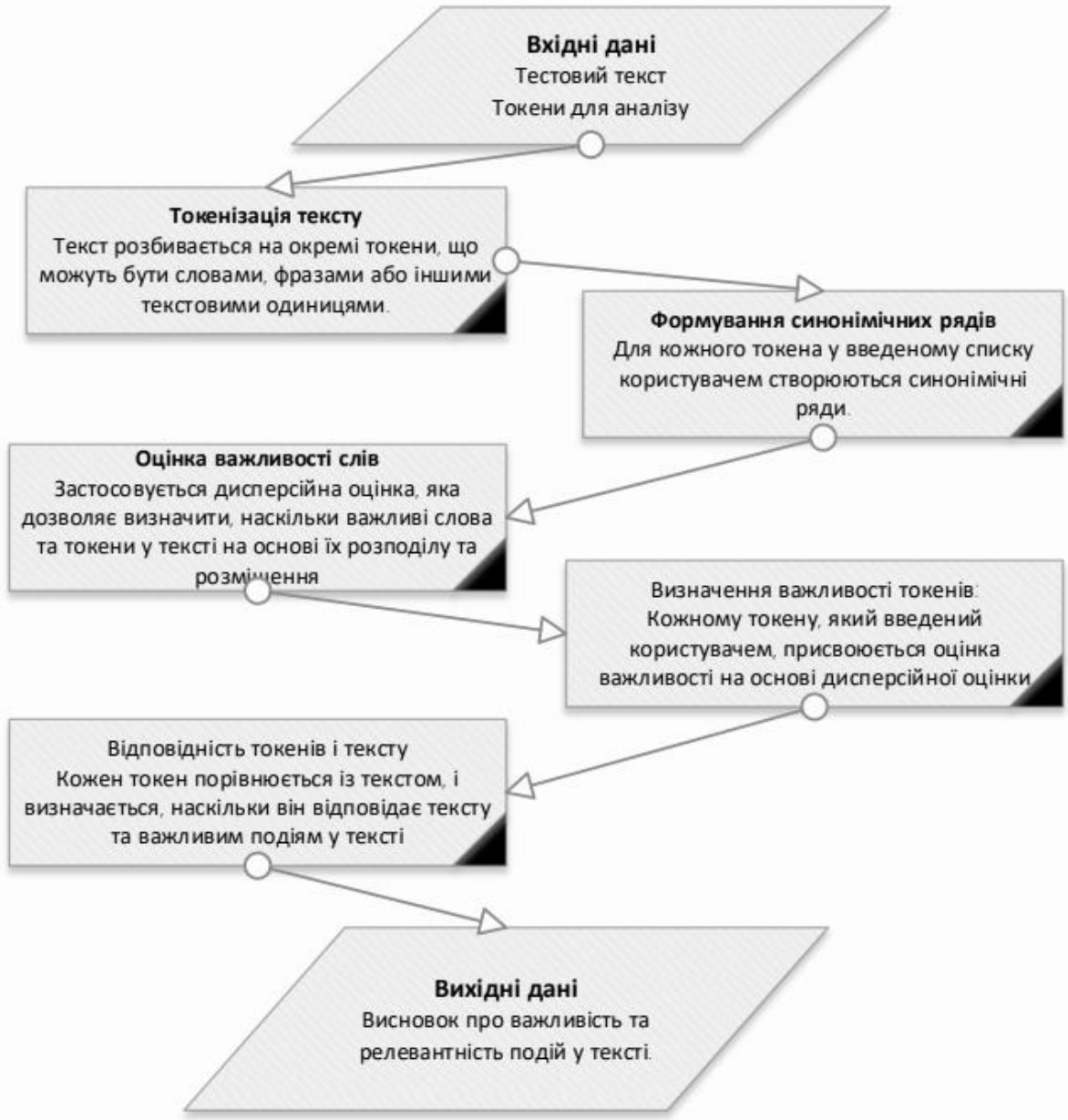


Рисунок 1 – Схема методу ідентифікації подій в україномовних текстах засобами обробки природної мови

Формування синонімічних рядів. Після токенізації для кожного токена, введеного користувачем, формуються синонімічні ряди. Синонімічний ряд - це набір інших слів або фраз, які мають схожий семантичний зміст або пов'язані за контекстом. Формування синонімічних рядів допомагає розширити можливості аналізу тексту.

Оцінка важливості слів. Для оцінки важливості слів у тексті використовується дисперсійна оцінка. Дисперсійна оцінка визначає, наскільки

слова розподілені у тексті та як вони пов'язані між собою. Ця оцінка допомагає визначити, які слова є ключовими для подій у тексті.

Визначення важливості токенів. Кожному токену, введеному користувачем, присвоюється оцінка важливості на основі дисперсійної оцінки. Оцінка важливості враховує розташування токену в тексті та його семантичний зв'язок з іншими словами. Токени, які мають високий рівень важливості, вважаються ключовими для подій у тексті.

Відповідність токенів і тексту. Кожен токен порівнюється із текстом з використанням отриманих оцінок важливості та дисперсійної оцінки. Це допомагає визначити, наскільки токен відповідає тексту і наскільки важливий для подій, які відображені в тексті.

Результати та висновок. На основі оцінок важливості та відповідності токенів тексту введеним користувачем токенам генерується висновок про важливість та релевантність подій у тексті. Ця інформація допомагає ідентифікувати та аналізувати події та їх контекст у тексті.

Отже, метод ідентифікації подій в україномовних текстах засобами обробки природної мови полягає в складному аналізі тексту та оцінці важливості токенів на основі дисперсійної оцінки. Цей метод дозволяє точно ідентифікувати та аналізувати події в тексті.

Таким чином, було проведено дослідження в галузі аналізу тексту, зокрема ідентифікації подій в україномовних текстах, запропоновано структуру методу, що може бути втілено в програмних реалізаціях.

Перелік посилань

1. An improved text mining approach to extract safety risk factors from construction accident reports
2. AL-NASSERI, Alya; ALI, Faek Menla; TUCKER, Allan. Investor sentiment and the dispersion of stock returns: Evidence based on the social network of investors. International Review of Financial Analysis, 2021, 78: 101910.
3. Text preprocessing for text mining in organizational research: Review and recommendations HICKMAN, Louis, et al. Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 2022, 25.1: 114-146.
4. KONONOVA, Olga, et al. Opportunities and challenges of text mining in materials research. Iscience, 2021, 24.3. NA, X. U., et al. An improved text mining approach to extract safety risk factors from construction accident reports. Safety science, 2021, 138: 105216.
5. GRIES, Stefan Th. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. Research in Corpus Linguistics, 2021, 9.2: 1-33.

Додаток Ж

Презентаційний матеріал

Кваліфікаційна робота магістра

Метод ідентифікації подій в україномовних текстах засобами обробки природної мови

Виконав
студент групи КНм-22-1
Домбровський Назарій Сергійович

Керівник
старший викладач кафедри КН
Скрипник Тетяна Казимирівна

Мета роботи

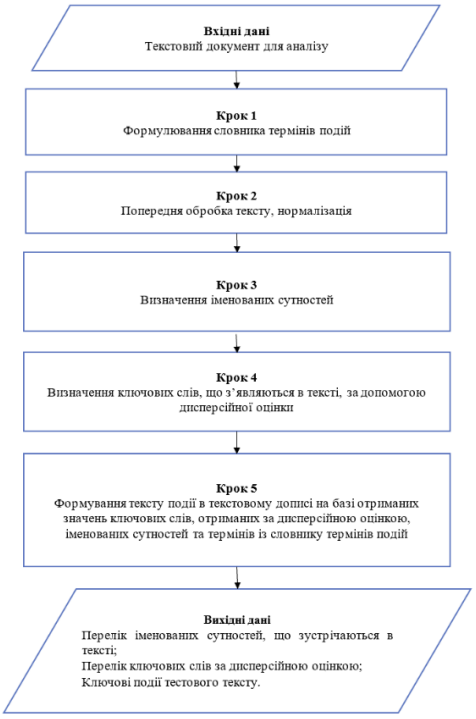
Мета кваліфікаційної роботи магістра – вирішення задачі інтелектуальної ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для вхідного текстового контенту визначити ключові події, та виводити відповідний текст на основі визначення іменованих сутностей, що зустрічаються в тексті, множини ключових термінів, отриманих шляхом обчислення дисперсійної оцінки та сформованого словника термінів подій. Також необхідно створити відповідну програмну реалізацію для апробації методу.

Завдання роботи

Вирішення задачі ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для вхідного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового контенту, перетворити у вихідні дані у вигляді тексту, що стисло відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій для формування речень з опису події.

Також необхідно створити відповідну програмну реалізацію для апробації методу.

Схема методу ідентифікації подій в україномовних текстах засобами обробки природної мови



Формування словника термінів подій для їх ідентифікації в україномовних текстах засобами обробки природної мови

Алан: З застережень безпеки [застерігати], з міркувань [міркувати] безпеки... Але я можу надрукувати [надрукувати] дещо з того, що ви хочете [хотіти] їм сказати [сказати], передати притулку [передати], і я можу переконатися [переконатися], що вони отримають [отримати] це повідомлення, якщо це вам допоможе [допомагати].

Підхід до визначення ключових термінів із використанням дисперсійної оцінки

Дисперсійна оцінка зберігає смисловий зміст тексту, оскільки вона виділяє ключові слова та фрази, які мають семантичний зв'язок із змістом подій. Це важливо для правильної ідентифікації та розуміння подій в текстах. Також дисперсійна оцінка може бути легко та ефективно реалізована та розрахована, що робить її застосування доцільним в методах обробки природної мови.

$$\sigma^2 = \frac{\sum(xi - \mu)^2}{N}$$

Нейромережева архітектура моделі Stanza для обробки природної мови

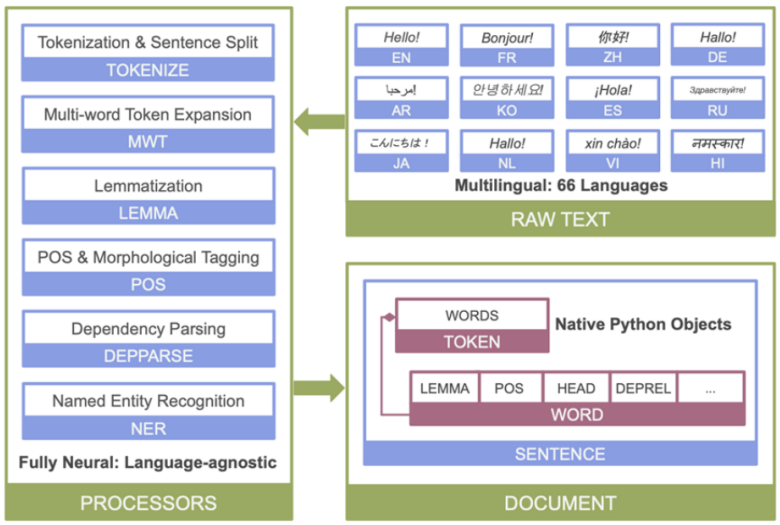


Схема прикладу застосування NER та дисперсійної оцінки

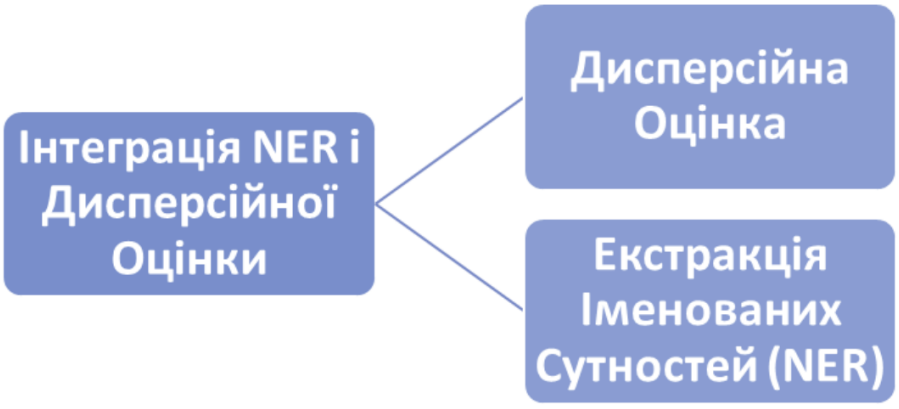


Схема процесу формування тексту ключових подій

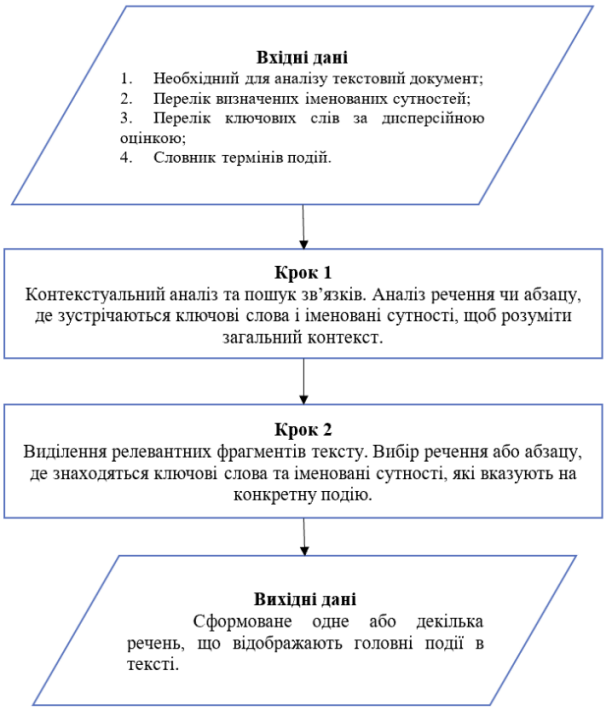
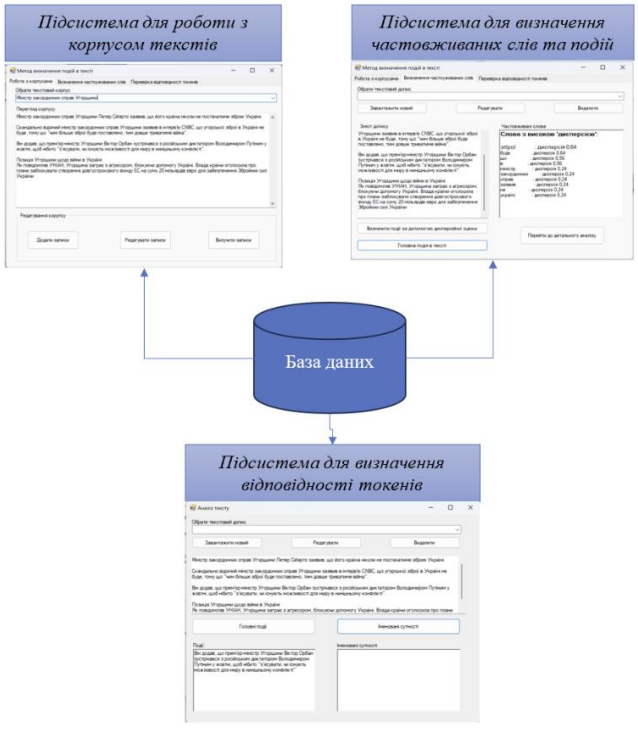


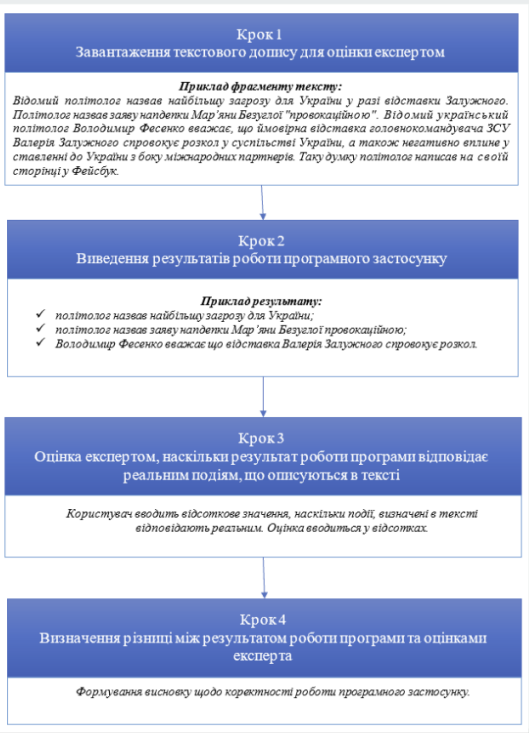
Схема інформаційної системи ідентифікації подій в україномовних текстах засобами обробки природної мови



Дослідження ефективності інформаційної системи на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови

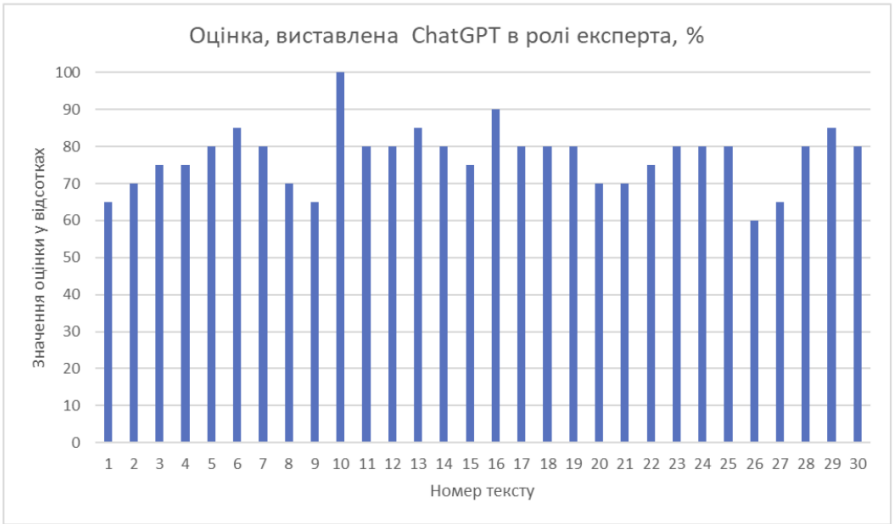
Для оцінки ефективності розробленого програмного застосунку було використано метод порівняння результатів, отриманих програмним продуктом, з оцінками експерта та Chat GPT(версія 3.5). Експерт та Chat GPT, переглядав кожен текст, оцінюючи, наскільки подія, ідентифікована реалізованим програмним застосунком, відповідає дійсному змісту події в тексті. Проаналізувавши текст, експерт визначає оцінку у відсотках, наскільки результат роботи програми відповідає дійсному змісту подій в тексті.

Алгоритм оцінювання ефективності інформаційної системи на базі методу ідентифікації подій в україномовних текстах

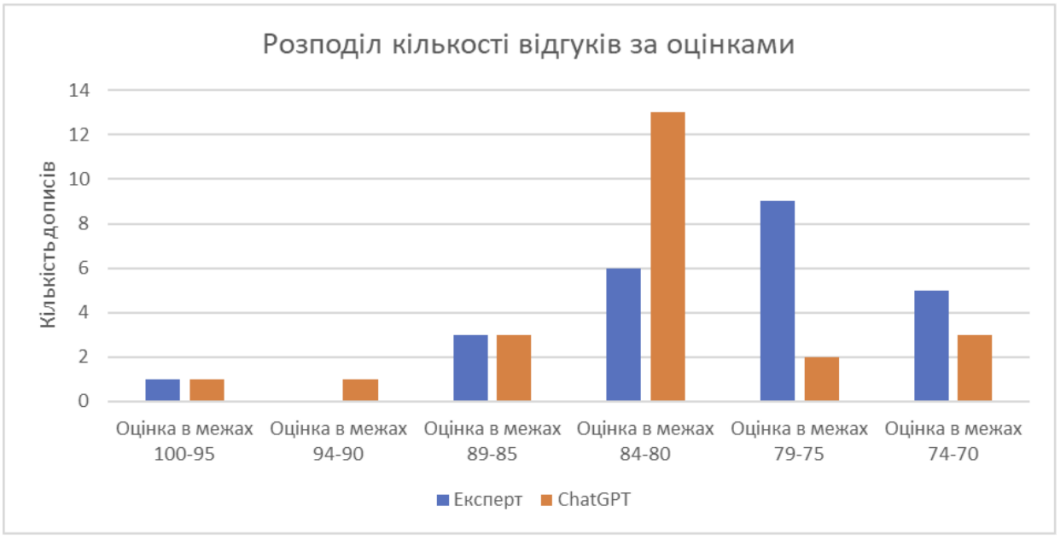


Діаграма із результатами оцінювання ефективності інформаційної системи на базі методу ідентифікації подій за участі експерта

Діаграма із результатами оцінювання ефективності інформаційної системи на базі методу ідентифікації подій за участі Chat GPT



Розподіл кількості правильно ідентифікованих подій в дописах за оцінками



Висновки

Кваліфікаційна робота магістра розв'язує задачу ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового допису перетворити у вихідні дані у вигляді тексту, що найточніше відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій. Також необхідно створити відповідну програмну реалізацію для апробації методу.

Висновки

В результаті виконання роботи було розроблено інформаційну систему ідентифікації подій в україномовних текстах засобами обробки природної мови, що дозволяє для визначеного текстового контенту визначити ключові події, що згадуються в тексті за вхідними даними у вигляді тестового текстового допису перетворити у вихідні дані у вигляді тексту, що найточніше відображає подію за допомогою використання дисперсійної оцінки для визначення ключових слів, нейромережевої моделі Stanza для виокремлення іменованих сутностей та сформованого словника термінів подій.

Mon Dec 18 12:20:49 EET 2023, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 2.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 9%**

	Документ		Сумарний збіг по Базі Даних	
ID: 123671 Назва: КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА на тему Метод ідентифікації подій в україномовних текстах засобами обробки природної мови Додано в БД: 2023-12-18 Автора: Н.С. Домбровський Керівники: Т.К. Скрипник Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	91571	1301	3276 (4%)	51 (4%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016017332

Дата перевірки:
18.12.2023 13:22:47 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
18.12.2023 13:44:05 EET

ID користувача:
100005671

Назва документа: **КНм-22-1 Домбровський**

Кількість сторінок: **80** **Кількість слів:** **14335** **Кількість символів:** **108963** **Розмір файлу:** **2.40 MB** **ID файлу:** **1015704596**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

9.05%
Схожість

Найбільша схожість: **2.08%** **з джерелом з Бібліотеки** (**ID файлу:** **1013021999**)

7.79% Джерела з Інтернету	**739**	**Сторінка** 82
5.29% Джерела з Бібліотеки	**190**	**Сторінка** 87

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Підозріле форматування	**14 сторінок**

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ
КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ
ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод ідентифікації подій в україномовних текстах засобами обробки природної мови*

Автор: *Домбровський Назарій Сергійович*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *старший викладач кафедри КН Скрипник Тетяна Казимирівна*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	–
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	–
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	–

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) За програмою Anti-Plagiarism виявлені 2%, які є фрагментарними, загальновідомі терміни та визначення.
2) За програмою UNICHECK виявлені 9,05%, які є фрагментарними, загальновідомі терміни та визначення.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 2% і 9,05% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Тетяна Скрипник

Гарант ОП _____ Руслан Багрій

Завідувач кафедри КН _____ Олександр Бармак

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Домбровського Назарія Сергійовича за темою: *Метод ідентифікації подій в україномовних текстах засобами обробки природної мови*

1. Актуальність обраної теми

У наш час, коли обсяги текстових даних, які включають новини, соціальні медіа, блоги та наукові публікації, невпинно ростуть, з'являється нагальна потреба в ефективних методах їх аналізу. Це особливо важливо для мов, які традиційно мають обмежені ресурси, наприклад, української мови. Швидке і точне виявлення подій, особливо у сфері новин та соціальних медіа, відіграє вирішальну роль у інформуванні громадськості та прийнятті критично важливих рішень. Тому тема «Метод ідентифікації подій в україномовних текстах засобами обробки природної мови», є важливою і перспективною у контексті сучасного інформаційного потоку.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Поставлена у кваліфікаційній роботи магістра мета, пов'язана з створенням методу ідентифікації подій в україномовних текстах засобами обробки природної мови, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи.

3. Професійні та особистісні якості магістранта

Виконуючи кваліфікаційну роботу магістра Домбровський Назарій Сергійович проявив себе як кваліфікований та сумлінний студент, поставлені задачі виконував якісно, вчасно та старанно. Проявив достатні знання та навички для одержання успішного результату компетентності знання та навички.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Результати, отримані у рамках виконання магістерської кваліфікаційної роботи, відображають самостійну діяльність та дослідницькі зусилля студента. Отримані положення наукової новизни та інновації, описані в роботі, дозволили покращити існуючі методи в галузі ідентифікації подій в україномовних текстах засобами обробки природної мови.

5. Наукова новизна та оригінальність запропонованих підходів

У магістерській кваліфікаційній роботі була представлена наукова новизна та інноваційні підходи, які відповідають вимогам спеціальності 122 «Комп'ютерні науки»,

особливо у контексті розробленої теми «Метод ідентифікації подій в україномовних текстах засобами обробки природної мови». Основною особливістю цього підходу є його спроможність аналізувати текстові дані українською мовою, що дає змогу отримати більш глибоке розуміння впливу певних подій чи особистостей на громадську думку. Результати цієї роботи були успішно представлені на науково-практичній конференції.

6. Ступінь оволодіння методами дослідження

Магістрант виявив високий ступінь оволодіння необхідними методами дослідження.

7. Повнота та якість розкриття теми роботи

Тема роботи в повній мірі обґрунтована й розкрита, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання у роботі виконані, а також проведено аналіз результатів прикладного застосування запропонованих засобів ідентифікації подій в україномовних текстах засобами обробки природної мови.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Структура роботи й послідовність викладення логічні та відповідні поставленій меті. Викладення матеріалу грамотне та виявляє високий ступінь відповідності стилю.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Було створено інформаційну систему на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови, яка є прикладною програмною реалізацією відповідного методу й призначена для проведення експериментів із метою дослідження його ефективності. Проведені дослідження ефективності запропонованого в роботі методу ідентифікації подій в україномовних текстах засобами обробки природної мови виконувалось з використанням розробленої відповідної інформаційної системи.

10. Висновок про можливість допуску кваліфікаційної роботи до захисту, на яку оцінку заслуговує робота

Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка *«добре»*.

Науковий керівник _____ старший викладач каф. КН Скрипник Тетяна Казимирівна

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Домбровського Назарія Сергійовича за темою: *Метод ідентифікації подій в україномовних текстах засобами обробки природної мови*

1. Актуальність обраної теми

У сучасному цифровому світі обсяги текстової інформації, включаючи новини, соціальні медіа, блоги, наукові публікації, постійно зростають. Це створює потребу в ефективних інструментах для їх аналізу, особливо для мов, які мають менше ресурсів, таких як українська. У контексті новин та соціальних медіа швидке та точне розпізнавання подій є ключовим для інформування громадськості та прийняття важливих рішень.

Тема "Метод ідентифікації подій в україномовних текстах засобами обробки природної мови" набуває особливої актуальності в контексті сучасного інформаційного простору, де обсяги текстових даних стрімко зростають. Тому робота, виконана автором є актуальною та перспективною.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Обрана тема ідентифікації подій в україномовних текстах засобами обробки природної мови, в межах якої виконані поставлені задачі, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи магістра.

3. Повнота розкриття мети та завдань дослідження

В роботі автор повністю розкриває мету дослідження та поставленні в межах теми завдання.

4. Наявність наукової новизни

В кваліфікаційній роботі представлена наукова новизна та інновації, відповідна спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження. Продемонстровано й обґрунтовано результати, які мають наукове значення. Результати дослідження оприлюдненні у збірнику наукових праць АПКН-2023.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. У першому розділі виконано аналіз сучасного стану області ідентифікації подій в україномовних текстах засобами обробки природної мови.

Другий розділ присвячено розробці методу ідентифікації подій в україномовних текстах засобами обробки природної мови. У третьому розділі виконано розробку прикладного програмного застосунку на базі методу ідентифікації подій в україномовних текстах засобами обробки природної мови. У четвертому розділі виконано дослідження ефективності методу ідентифікації подій в україномовних текстах засобами обробки природної мови.

6. Ступінь розкриття теми роботи

Тема кваліфікаційної роботи повною мірою розкрита та обґрунтована, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, та проведено аналіз результатів прикладного застосування запропонованих методу і засобів.

7. Якість оформлення кваліфікаційної роботи

Оформлення роботи відповідає необхідним нормам та вимогам, які ставляться до оформлення кваліфікаційних робіт.

8. Недоліки кваліфікаційної роботи

Було б корисно привести приклади випробувань запропонованого підходу, враховуючи наявність численних граматичних та лексичних недоліків у тексті.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота

Враховуючи рівень виконання кваліфікаційної роботи магістра та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка *«добре».*

Опонент: Ярецька Н.О., к.ф.-м.н., доцент кафедри ВМКЗ