Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерної інженерії та інформаційних систем

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

Галузь знань _____12 – Інформаційні технології_____

Спеціальність _____123 –Комп'ютерна інженерія _____

на тему:
«Метод автоматичної класифікації фразеологічних одиниць англомовних текстів»»

КвРКІ. 170154.21.01.26 ПЗ

Виконав: студент 2 курсу, група КІ2м-21-1 _____ Є.Р. Зінюк
Підпис Ініціали, прізвище

Керівник доктор техн. наук, професор _____ О.В. Боровик
Науковий ступінь, вчене звання Підпис Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КІІС, д.т.н, проф.
Т.О. Говорущенко _____
19 травня 2023 р.

Хмельницький, 2023

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОРМАЦІЙНИХ СИСТЕМ

Освітній рівень МАГІСТР

Галузь знань 12 ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Спеціальність 123 КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Освітня програма ОСВІТНЬО-НАУКОВА ПРОГРАМА «КОМП'ЮТЕРНА ІНЖЕНЕРІЯ ТА ПРОГРАМУВАННЯ»

ЗАТВЕРДЖУЮ
Зав. кафедри Т.О.Говорущенко
" 01 " 09 2022 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

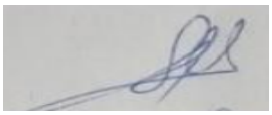
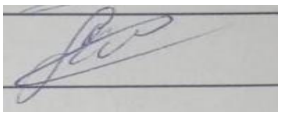
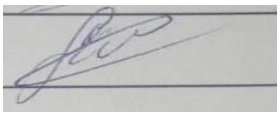
Зінюк Євгеній Ростиславович
Прізвище, ім'я, по батькові студента

1. Тема проекту (роботи) Метод автоматичної класифікації фразеологічних одиниць англомовних текстів
Керівник проекту (роботи) Боровик О.В. д.т.н, професор
Прізвище, ім'я, по батькові, науковий ступінь, вчене звання

Затверджена наказом ректора університету від 09.01.2023 р. № 1
2. Строк подання студентом проекту (роботи) на кафедру 01.05.2023 р.
3. Вихідні дані до проекту (роботи) Завдання на дипломне проектування
4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)
Аналіз існуючих методів автоматичної класифікації фразеологізмів
Розробка алгоритму роботи системи класифікації фразеологічних одиниць;
Програмна реалізація методу автоматичної класифікації фразеологічних одиниць англомовних текстів

5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень)

6. Консультанти розділів кваліфікаційної роботи магістра

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Нормоконтроль	Лисенко С.М, професор кафедри КІІС		
Антиплагіат	Нічепорук А.О, доцент кафедри КІІС		

7. Дата видачі завдання «_06_»_____09_____2022р.

КАЛЕНДАРНИЙ ПЛАН

№з/п	Назва етапів (розділів) кваліфікаційної роботи магістра	Термін виконання етапів проекту (роботи)	Примітка
1	Вибір напряму дослідження та узгодження тематики КвРМ з керівником	05.09.2022	виконано
2	Ознайомлення з предметною областю; формулювання мети та задач дослідження; визначення об’єкта та предмета дослідження	05.10.2022	виконано
3	Робота над розділом 1 – аналіз відомих моделей, методів за темою; постановка задачі	05.11.2022	виконано
4	Робота над розділом 2 – розробка моделей для вирішення поставленої задачі	05.12.2022	виконано
5	Робота над науковою статтею	05.01.2023	виконано
6	Робота над розділом 3 – розробка методів для вирішення поставленої задачі	15.02.2022	виконано
7	Робота над розділом 4 – проектування та розробка ПЗ для вирішення поставленої задачі, експериментальна частина	05.04.2023	виконано
8	Оформлення пояснювальної записки згідно вимог	15.04.2023	виконано
9	Попередній захист ДРМ	18.04.2023	виконано
10	Захист ДРМ на засіданні ЕК	До 10.05.2023	

Студент  Підпис Є.Р. Зінюк Ініціали, прізвище

Керівник роботи  Підпис О.В. Боровик Ініціали, прізвище

РЕФЕРАТ

Тема дипломної роботи: Метод автоматичної класифікації фразеологічних одиниць англомовних текстів

Автор роботи: Є.Р. Зінюк

Керівник роботи: О.В. Боровик

Пояснювальна записка: 91 с, 8 рис, 10 дод, 80 джерел.

ПЕРЕЛІК КЛЮЧОВИХ СЛІВ: фразеологічна одиниця, програмне забезпечення, штучний інтелект, автоматична класифікація.

Об’єктом дослідження є класифікація фразеологічних одиниць в англомовних текстах.

Предметом дослідження є науково-методичний апарат автоматичної класифікації фразеологічних одиниць в англомовних текстах.

Метою дипломної роботи є підвищення ефективності автоматичної класифікації фразеологічних одиниць англомовних текстів та зменшення кількості помилкових тлумачень. Для розв’язання поставлених задач використовувалися методи:

1) Аналіз відомих методів та рішень для автоматичної класифікації фразеологічних одиниць.

2) Моделювання методу удосконалення автоматичнох класифікації фразеологічних одиниц.

Наукова новизна отриманих результатів: удосконалено метод автоматичної класифікації фразеологічних одиниць англомовних текстів на основі інтеграції алгоритмів методу на основі правил і машинного навчання.

У результаті виконання наукового дослідження опрацьовано програмне забезпечення автоматичної класифікації фразеологічних одиниць на основі реалізації алгоритму гібридного методу, що поєднує метод на основі правил і метод на основі машинного навчання. Це дозволило збільшити ефективність класифікації ФО англомовних текстів, зменшило час класифікації та збільшило кількість ознак для класифікації.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ 8
ВСТУП..... 9
1. ХАРАКТЕРИСТИКА ПРЕДМЕТНОЇ ОБЛАСТІ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ 12
1.1 Аналіз предметної області автоматичної класифікації фразеологізмів 12
1.2 Аналіз існуючого науково-методичного апарату автоматичної класифікації фразеологізмів 16
1.2.1 Аналіз існуючих методів автоматичної класифікації фразеологізмів 22
1.2.2 Аналіз існуючого програмного забезпечення автоматичної класифікації фразеологізмів 28
1.3 Постановка задачі..... 30
1.4 Висновки 31
2. ПРОЕКТУВАННЯ СТРУКТУРИ ІНФОРМАЦІЙНОЇ СИСТЕМИ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ..... 33
2.1 Метод автоматичної класифікації фразеологізмів англомовних текстів на основі інтеграції методу на основі правил з алгоритмами машинного навчання33
2.2 Розробка структури інформаційної системи автоматичної класифікації фразеологізмів 40
2.3 Вибір апаратних засобів для автоматичної класифікації фразеологізмів..... 48
2.4 Висновки 54
3. АЛГОРИТМИ ТА ТЕХНОЛОГІЯ ОБРОБКИ ІНФОРМАЦІЙНИХ ПОТОКІВ У СИСТЕМІ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ..... 56
3.1 Алгоритми автоматичної класифікації фразеологізмів 56
3.2 Технологія обробки інформаційних потоків у системі автоматичної класифікації фразеологізмів..... 64
3.3 Проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів..... 67

3.4 Висновки 74
4. ПРОГРАМНО-ТЕХНІЧНА СИСТЕМА РЕАЛІЗАЦІЇ МЕТОДУ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ 76
4.1 Опис середовища розробки програмно-технічної системи реалізації методу автоматичної класифікації фразеологізмів 76
4.2 Програмна реалізація методу автоматичної класифікації фразеологічних одиниць англомовних текстів 81
4.3 Висновки 89
ВИСНОВОК 91
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ 93
ДОДАТОК А Копія наукової публікації 5
ДОДАТОК Б Презентація дипломної роботи 6
ДОДАТОК В Опис підходів автоматичної класифікації ФО 22
ДОДАТОК Г Можливості ПЗ автоматичної класифікації ФО 24
ДОДАТОК Ґ Характеристика методу в порівнянні з іншими та прогнозований ефект від удосконалення 27
ДОДАТОК Д Автоматична класифікація ФО на основі застосування Sketch Engine 29
ДОДАТОК Е Автоматична класифікація ФО на основі застосування гібридного ПЗ Hybrid Soft 41
ДОДАТОК Є Приклад фрагменту коду, який демонструє, як використовувати модуль оцінки в Python 52
ДОДАТОК Ж Приклад фрагменту коду, який демонструє, як використовувати модуль оптимізації в Python 53
ДОДАТОК З Приклад реалізації гібридного алгоритму з використанням алгоритму NER і алгоритму Naïve Bayes 54
ДОДАТОК И UML діаграма, яка показує основні компоненти та їх звязки в гібридному алгоритмі 55

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

ПЗ - програмне забезпечення

БД - база даних

ОС - операційна система

ФО - фразеологічна одиниця

HS - Hybrid Soft

NLP - Neuro-linguistic programming

NLTK - Natural Language Toolkit

ВСТУП

У зв'язку з бурхливим розвитком комп'ютерних технологій в останній період дедалі більше дослідників приділяють свою увагу проблемам автоматичної обробки текстів. Одним із актуальних завдань, яке стосується автоматичної обробки текстів, є автоматична класифікація фразеологічних одиниць англомовних текстів.

Це пояснюється наступними причинами. Величезний обсяг текстових даних, що доступні сьогодні, унеможливлює «вручну» аналізувати та класифікувати всі фразеологічні одиниці, які присутні в англомовних текстах. Фразеологічні одиниці часто відіграють вирішальну роль у розумінні та створенні мови. Автоматична класифікація фразеологічних одиниць має численні застосування в різних сферах, включаючи навчання мови, переклад, обробку природної мови та комп’ютерну лінгвістику. Фразеологічні одиниці можуть становити серйозну проблему для систем машинного перекладу, яким часто важко точно перекладати ідіоматичні вирази та усталені фрази.

На даний час існують різні підходи до класифікації фразеологічних одиниць англомовних текстів. Серед найбільш поширених можна виокремити такі: корпусний підхід; когнітивний лінгвістичний підхід; граматичний підхід; стилістичний підхід; міжлінгвістичний підхід.

Складність задачі класифікації фразеологічних одиниць обумовлює застосування для її вирішення різних програмних засобів. Хоча використання цих засобів для автоматичної класифікації фразеологічних одиниць в англомовних текстах має багато переваг, є також деякі недоліки, які слід враховувати. До числа таких можна віднести обмежену точність, труднощі з контекстом, обмежене охоплення, необхідність налаштування, вартість.

Одним із можливих шляхів підвищення точності та ефективності програмних засобів автоматичної класифікації фразеологічних одиниць є їх інтеграція з алгоритмами машинного навчання. Це може включати навчання алгоритмів на

великих наборах даних анотованого тексту з перевіркою людиною для підтвердження точності класифікацій. Постійно вдосконалюючи точність алгоритмів за допомогою машинного навчання, програмні інструменти можуть стати більш ефективними в ідентифікації та класифікації ширшого діапазону фразеологічних одиниць, у тому числі тих, які є складними або неоднозначними.

Метою дипломної роботи є підвищення ефективності автоматичної класифікації фразеологічних одиниць англомовних текстів та зменшення кількості помилкових тлумачень.

Поставлена мета досягається розв’язанням таких основних задач:

1) Аналіз існуючих методів автоматичної класифікації фразеологізмів;

2) Аналіз існуючого програмного забезпечення автоматичної класифікації фразеологізмів;

3) Обгрунтування методу автоматичної класифікації фразеологізмів англомовних текстів на основі інтеграції алгоритмів машинного навчання;

4) Розробка структури інформаційної системи автоматичної класифікації фразеологізмів;

5) Проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів;

6) Розробка алгоритму роботи системи класифікації фразеологічних одиниць;

7) Програмна реалізація методу автоматичної класифікації фразеологічних одиниць англомовних текстів.

Об’єктом дослідження є класифікація фразеологічних одиниць в англомовних текстах

Предметом дослідження є науково-методичний апарат автоматичної класифікації фразеологічних одиниць в англомовних текстах.

Наукова новизна отриманих результатів:

1) Удосконалено метод автоматичної класифікації фразеологічних одиниць англомовних текстів на основі інтеграції алгоритмів методу на основі правил і машинного навчання.

2) Удосконалено програмно-технічну систему реалізації методу автоматичної класифікації фразеологізмів.

Практична цінність отриманих результатів. У результаті виконання наукового дослідження опрацьовано програмне забезпечення автоматичної класифікації фразеологічних одиниць на основі реалізації алгоритму гібридного методу, що поєднує метод на основі правил і метод на основі машинного навчання. Це дозволило збільшити ефективність класифікації ФО англомовних текстів, зменшило час класифікації та збільшило кількість ознак для класифікації. Матеріали дипломної роботи апробовані на конференції "Автоматизація та комп’ютерно-інтегровані технології у виробництві та освіті: стан, досягнення, перспективи розвитку: матеріали Всеукраїнської науково-практичної Internet-конференції. – Черкаси, 2023. - 196 с.".

1. ХАРАКТЕРИСТИКА ПРЕДМЕТНОЇ ОБЛАСТІ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ

1.1 Аналіз предметної області автоматичної класифікації фразеологізмів

Фразеологічні одиниці є фундаментальним компонентом будь-якої мови, відіграючи вирішальну роль у процесі спілкування.

Мета цього параграфа заключається у визначенні таких понять, як фразеологічна одиниця, типи фразеологічних одиниць та чому є важливим дослідження в області автоматичної класифікації фразеологічних одиниць.

Фразеологічні одиниці – це багатослівні лексичні одиниці, що характеризуються певним ступенем фіксації чи ідіоматичністю своїх компонентів. Іншими словами, фразеологізми є поєднанням слів, значення яких не обов'язково виводиться зі значення його компонентів, тобто слова разом можуть означати більше, ніж їх сума частин [1].

Ці лінгвістичні структури також відомі у літературі як фраземи, стійкі висловлювання та багатослівні висловлювання. Незважаючи на те, що носії мови легко засвоюють такі висловлювання, їх інтерпретація представляє серйозну проблему для обчислювальних систем через їхній гнучкий і гетерогенний характер. Крім того, фразеологічні одиниці не так часто зустрічаються в лексичних ресурсах, як у реальному тексті, і ця проблема охоплення може вплинути на виконання багатьох завдань обробки природної мови.

Фразеологічні одиниці широко використовуються людиною. Кількість фразеологічних одиниць, виражених багатослівними виразами, має той же порядок, що і кількість простих чи відокремлених слів [2].

Ще одним доцільним визначенням фразеологічних одиниць є те, що ФО - це вислови, що складаються з двох або більше слів, які функціонують як єдина лексична одиниця. Вони мають фіксовану структуру, і їх значення не можна вивести зі значень їхніх окремих компонентів. Приклади фразеологічних одиниць

в англійській мові включають «кикнути відро», «тягнути когось за ногу», «тримати коней» тощо. Фразеологічні одиниці є невід’ємним компонентом мови, який широко використовується в повсякденному спілкуванні, літературі та інших формах дискурсу. Вони забезпечують стислий і часто яскравий спосіб вираження складних ідей, емоцій і дій. Наприклад, фраза «стримай своїх коней» означає бути терплячим і чекати, тоді як буквальне значення слів не передає цієї ідеї. Фразеологізми можна розділити на кілька категорій залежно від їхньої структури та значення. Найпоширеніша класифікація базується на їхньому лексичному складі, який включає ідіоми, словосполучення, фразові дієслова та прислів'я.

Ідіоми - це стійкі вирази, значення яких не може бути виведене з окремих слів, що входять до складу виразу. Словосполучення - це словосполучення, які часто зустрічаються і мають сильну лексичну асоціацію. Фразові дієслова - це дієслівні словосполучення, які складаються з дієслова та однієї або кількох часток, які разом передають певне значення. Прислів'я - це вирази, які передають загальну істину або пораду. Вербальна фразеологічна одиниця - це фразеологічна одиниця, яка містить одне дієслово як граматичний центр. Наприклад, фразеологізм "прийти до тями" означає змінити свою думку. Дієслівні фразеологізми чудово ілюструють загальну насиченість. Зважаючи на цю особливість, а також на те, що дієслівні словосполучення мають парадигматичний розрив, змушують нас зосередити свою увагу саме на цьому типі фразеологізмів, що передбачає дуже складну дослідницьку лінію в плані семантичної ідентифікації та класифікації фразеологізмів.

Таким чином, важливо вивчити природу таких мовних структур, щоб ми могли сконструювати автоматичні методи роботи з цими одиницями.

Автоматична класифікація фразеологізмів в англійських текстах є важливим завданням в обробці природної мови (NLP), що передбачає ідентифікацію та категоризацію груп слів або словосполучень на основі їхніх семантичних та синтаксичних властивостей. Одним з ефективних підходів до автоматичної класифікації фразеологізмів є інтеграція алгоритмів машинного навчання.

Машинне навчання - це підгалузь штучного інтелекту (ШІ), що передбачає навчання алгоритмів, які навчаються на основі даних і роблять прогнози або рішення без явного програмування. Інтеграція алгоритмів машинного навчання в автоматичну класифікацію фразеологічних одиниць передбачає навчання алгоритмів на маркованому наборі даних фразеологічних одиниць та їхніх відповідних категорій, що дозволяє алгоритмам вивчати характеристики та властивості різних типів фразеологічних одиниць і точно класифікувати нові екземпляри. Автоматична класифікація фразеологізмів складається з двох основних етапів: вилучення ознак і класифікація. Виділення ознак передбачає ідентифікацію релевантних ознак фразеологічних одиниць, які можуть бути використані для створення векторного представлення. Вибір методу вилучення ознак залежить від характеристик набору даних і наявних обчислювальних ресурсів. Інший підхід до інтеграції алгоритмів ШІ з текстом англійською мовою полягає у використанні нейронних мереж.

Нейронні мережі – це тип алгоритму машинного навчання, який моделюється за структурою людського мозку. Вони складаються із взаємопов’язаних вузлів, які обробляють інформацію та вивчають досвід. Нейронні мережі можна навчити розпізнавати шаблони в англійському тексті, наприклад структуру та значення фразеологічних одиниць, і використовувати цю інформацію для класифікації нових виразів.

Предметна область автоматичної класифікації фразеологічних одиниць лежить на стику комп’ютерної лінгвістики та обробки природної мови. Це передбачає розробку алгоритмів і методів, які можуть автоматично ідентифікувати та класифікувати фразеологічні одиниці, які є сталими або напів фіксованими виразами в мові, які мають переносне чи ідіоматичне значення, яке не можна легко вивести зі значень їхніх окремих слів. Філологи провели велику кількість досліджень та задавались питаннями, як визначити фразеологічні одиниці в тексті та визначили головні гіпотези, які допомагають ідентифікувати дані мовні одиниці в тексті. Серед основних можна виділити такі як:

1) Гіпотеза фіксації. Чим чіткіше є дієслівна фраза, тим вище її ймовірність бути дієслівною фразеологічною одиницею. Ми замінимо кожен компонент цільової вербальної фрази їхніми близькими синонімами, щоб перевірити, чи не втрачає сенс нова вербальна фраза. Щоб перевірити значення нової словесної фрази, ми розглядаємо можливість використання довідкового корпусу, де ми можемо шукати докази такої фрази.

2) Трансляційна гіпотеза. Чим дослівніше переклад словесного словосполучення, тим нижчою є можливість бути словесним фразеологізмом. Ми перекладемо словесну фразу з однієї мови на іншу. Після цього ми шукатимемо докази такого перекладу у довідковому корпусі, написаному цільовою мовою.

3) Гіпотеза внутрішньої привабливості та контекстуальної кореференції. Чим більше внутрішнє тяжіння і що нижча контекстуальна співвіднесеність у словесному словосполученні, то вище ймовірність того, що словесне словосполучення є дієслівною фразеологічною одиницею. Ми будемо використовувати статистичні методи для визначення рівня внутрішньої привабливості та контекстуальної кореляції між термінами вербальної фрази та термінами їхнього контексту.

4) Гіпотеза термінологічної галузі. Чим більша кількість словникових термінів поза поточним доменом для дієслівної фрази, тим вища ймовірність того, що вона є дієслівною фразеологічною одиницею. Використання термінів поза поточним доменом у реальному ДФО досить поширене, тому ми будемо визначати термінологію поза поточним доменом, щоб визначити, чи є така словесна фраза реальним ДФО. Автоматична класифікація фразеологічних одиниць може бути корисною в різноманітних програмах, таких як машинний переклад, аналіз тексту та генерація природної мови. Він включає такі методи, як розпізнавання образів, машинне навчання та статистичний аналіз для ідентифікації та класифікації цих виразів на основі їхніх синтаксичних і семантичних властивостей. Деякі концепти виражаються у мові через набір слів чи словосполучень, які інтуїтивно вживаються носіями мови, характеризуючи цим різні культурні спільності. Фразеологія, що

вважається культурною спадщиною мовної спільноти, спрямована на вивчення цих блоків слів, які прийнято відносити до фразеологізмів [3]. Вивчення фразеологічних одиниць набуває все більшого значення в останні роки, частково тому, що лінгвістична та обчислювальна лінгвістична спільнота усвідомила, що це явище охоплює всі компоненти речення, що включає різні аспекти природної мови: лінгвістика [4]. Фактично, робота покликана представити дослідження, які в основному зосереджені на підвищенні ефективності автоматичної класифікації фразеологічних одиниць в англомовних текстах.

1.2 Аналіз існуючого науково-методичного апарату автоматичної класифікації фразеологізмів

Науково-методичний апарат автоматичної класифікації фразеологічних одиниць став предметом дослідження в галузі комп'ютерної лінгвістики та обробки природної мови. Було запропоновано кілька підходів для автоматичної класифікації фразеологічних одиниць, які можна загалом класифікувати на три категорії: підхід на основі правил, підхід на основі статистики та підхід на основі машинного навчання.

В додатку В, наведено таблицю, що містить огляд різних підходів до автоматичної класифікації фразеологічних одиниць. Ця таблиця містить короткий опис кожного підходу та математичних методів, що використовуються в кожному з них.

Підходи, засновані на правилах, покладаються на створенні вручну правил ідентифікації та класифікації фразеологічних одиниць. Хоча ці підходи можуть бути ефективними в певних випадках, вони обмежені складністю необхідних правил і їх не здатністю обробляти нові або невидимі вирази.

Переваги: гнучкість: підхід, заснований на правилах, дозволяє легко налаштувати класифікацію фразеологічних одиниць відповідно до конкретних потреб програми. Зрозумілість: підхід, заснований на правилах, легко

інтерпретувати, оскільки правила чітко визначені та можуть бути легко зрозумілі людям. Висока точність: коли правила ретельно розроблені та перевірені, підхід, заснований на правилах, може досягти високої точності в класифікації фразеологічних одиниць.

Недоліки: обмежена масштабованість: підхід на основі правил обмежений можливістю розробляти правила вручну. Схильність до помилок: підхід, заснований на правилах, може бути схильний до помилок, особливо коли мова йде про складні або неоднозначні правила. Накладні витрати на технічне обслуговування: підхід, заснований на правилах, вимагає обслуговування вручну, що може бути трудомістким і дорогим. Відсутність адаптивності: підхід, заснований на правилах, може бути непридатним для динамічних або мінливих середовищ, де правила класифікації потрібно часто оновлювати. Статистичні підходи спираються на аналіз великих корпусів тексту для ідентифікації та класифікації фразеологічних одиниць. Ці підходи передбачають ідентифікацію слів і фраз, які одночасно зустрічаються в мові, і використання статистичних показників, таких як частота та взаємна інформація, для ідентифікації та класифікації виразів. Наприклад, статистичні підходи можуть ідентифікувати фразеологізм «шматок пирога» шляхом аналізу частоти його появи у великому корпусі тексту. Ці підходи можуть бути ефективними для ідентифікації загальних фразеологічних одиниць, але можуть бути обмежені в їх здатності ідентифікувати рідкісні вирази або обробляти варіації у виразах. Переваги:

1) Мета – це підхід, який ґрунтується на статистичному аналізі даних, який забезпечує об’єктивний спосіб визначення закономірностей у використанні мови.

2) Масштабований: підхід може обробляти великі обсяги даних, що робить його масштабованим для аналізу тексту з різних джерел і доменів.

3) Швидко: статистичні методи можна застосовувати швидко та ефективно, що дає змогу обробляти великі обсяги даних за короткий період.

4) Узагальнений: статистичний підхід можна узагальнити для різних мов і областей, що робить його застосовним у різних контекстах.

Недоліки:

1) Обмежена точність: статистичні моделі спираються на закономірності в даних, які не завжди можуть охопити складність використання мови.

2) Відсутність розуміння контексту: статистичні моделі не розуміють контекстуального значення мови та можуть не вловити нюанси значення в певних ситуаціях.

3) Зміщення даних: на точність статистичних моделей може вплинути зміщення даних, коли навчальні дані можуть не відображати використання мови в реальному світі.

4) Доменно-залежні: статистичні моделі є предметно-специфічними, що означає, що вони можуть не працювати належним чином, якщо застосовувати їх до нових доменів, де використання мови суттєво відрізняється від навчальних даних.

Підходи, засновані на машинному навчанні, спираються на алгоритми навчання для визначення та класифікації фразеологічних одиниць на основі прикладів із корпусу тексту. Ці підходи передбачають використання таких методів, як глибоке навчання та нейронні мережі, для виявлення шаблонів і структур у мові та класифікації виразів на основі їхніх синтаксичних і семантичних властивостей. Наприклад, підходи, засновані на машинному навчанні, можуть ідентифікувати фразеологізм «гавкати не на те дерево», навчаючи нейронну мережу на великому корпусі тексту та ідентифікуючі шаблони слів і фраз, що зустрічаються одночасно. Ці підходи можуть бути дуже ефективними в ідентифікації та класифікації фразеологічних одиниць, але потребують великої кількості навчальних даних і обчислювальних ресурсів.

Переваги:

1. Висока точність: моделі машинного навчання можуть навчатися на основі даних і з часом підвищувати свою точність, що робить їх високоточними в ідентифікації фразеологічних одиниць у тексті.
2. Контекстуальне розуміння: моделі машинного навчання можуть вловлювати контекстуальні нюанси використання мови, роблячи їх більш точними у

визначенні правильного значення фразеологічної одиниці в певному контексті.

3. Надійність: моделі машинного навчання можуть справлятися з шумом і варіаціями даних, що робить їх більш надійними в реальних умовах.
4. Масштабованість: моделі машинного навчання можна навчити на великих обсягах даних, що робить їх масштабованими для аналізу тексту з різних джерел і доменів.

Недоліки:

1. Потрібні великі обсяги навчальних даних: моделі машинного навчання вимагають великих обсягів позначених навчальних даних для навчання, отримання яких може зайняти багато часу та бути дорогим.

2. Переобладнання: моделі машинного навчання можуть підлаштовуватися під навчальні дані, що означає, що вони можуть погано узагальнюватись для нових даних.

3. Можливість інтерпретації: моделі машинного навчання може бути важко інтерпретувати, що ускладнює розуміння того, як вони приймають рішення.

4. Відсутність прозорості: моделі машинного навчання можуть бути підходом «чорної скриньки», що ускладнює розуміння того, як алгоритм прийшов до своїх рішень.

Загалом існуючий науково-методичний апарат автоматичної класифікації фразеологічних одиниць передбачає низку підходів, кожен із яких має свої переваги та недоліки. Майбутні дослідження, ймовірно, будуть зосереджені на поєднанні цих підходів і розробці більш складних алгоритмів, які можуть впоратися зі складністю та варіабельністю мови. Найбільш базові алгоритми пошуку потрібних мовних структур все рівно розпочиналися з використанням паперових словників та фізичного пошуку та класифікації. Однак, якщо пошук за алфавітом у паперових словниках зручний, то зворотний пошук (тобто пошук фразеологічної одиниці за її визначенням) у класичному паперовому словнику буквально рівносильний пошуку голки в стогу сіна. Усвідомлення цього призвело

до неодноразових спроб створення ономасіологічних словників, призначених для пошуку слів за їхніми поняттями. Незважаючи на величезну кількість словників, які довели свою ефективність у цій галузі, як і раніше існує значна прогалина у знаннях про їх використання та принципи складання. З іншого боку, використати зворотні словники не так просто, як може здатися. Наприклад, якби потенційний користувач шукав усі записи, що містять лексемний алфавіт, його пошук у One Look Reverse Dictionary дав би більше сотні результатів, серед яких були б орфографія, алфавіт, мова тощо, що, проте, набагато простіше в обігу, ніж весь набір статей. Якщо пошук базових слів при написанні чи редагуванні тексту можна полегшити за допомогою зворотних словників, то обробка послідовностей фразеологічних одиниць може стати набагато складнішим завданням з непередбачуваним результатом, оскільки багато фразеологічних одиниць не фігурують у двомовних словниках. Більше того, навіть якщо вони це роблять, їх переклади часто мають базову семантику або конотацію, відмінну від оригінальної. Логічно припустити, що пошук відповідної фрази може стати трудомістким процесом, і перекладачеві, автору текстів чи журналісту іноді доводиться витрачати години, намагаючись встановити зв'язок між змістом, який має на увазі користувач, та фразами, які існують у цільовому запиті. Враховуючи ці труднощі, складання фразеологічних ономасіологічних словників зі зручними та зрозумілими ключовими словами, що виступають як точки входу для користувача, може оптимізувати пошук фрази із заданим значенням. Вони повинні виявитися корисними, коли письменник чи перекладач не знає потрібної фрази цільовою мовою.

Як зазначалося раніше, двомовні чи багатомовні фразеологічні словники необов'язково надають еквівалентні переклади фразеологічних одиниць з погляду семантики, базової мотивуючої структури (чи образів), конотації чи значення контексті, у якому фраза використовується тоді. Якщо перекладні еквіваленти іноді далекі від бажаних, то чи такі ж самі перешкоди можна віднести до фразеологічних одиниць у межах однієї мови? І як за такої невідповідності критеріїв визначати

фразеологічні одиниці? Хоча інтерес до цієї концепції різко зріс протягом останнього десятиліття, деякі нечасті визначення можна було знайти й раніше. За Куніном, фразеологічні одиниці є кореферентні одиниці мови, що належать до одного граматичного класу, або частково збігаються, або повністю незалежні один від одного за своєю лексичною будовою, що містять як загальні, так і диференціальні компоненти, що збігаються або відрізняються за своїми стилями [5]. Хоча поняття фразеологічних одиниць, мабуть, до цього часу викристалізувалось, його практичне значення для перекладів виразно враховувалося раніше, тому дослідник вводить поняття міжмовного фразеологічної одиниці як «ФО, що збігаються за морфологічним складом значимих компонентів, за типом всієї ФО в цілому, але позбавлені міжмовного лексичного інваріанту [6].

Очевидно, що практична необхідність використання фразеологічних одиниць може виникнути як під час перекладу, так і при одномовному спілкуванні. Але саме в галузі перекладу проблема стає очевидною, тоді як при одномовному листі чи розмові комунікатори з меншою ймовірністю усвідомлюють необхідність фразеологічної одиниці, якщо вони не є досвідченими професіоналами. Може здатися загальноприйнятим, що ФО повинні належати до того самого граматичного класу. Однак, як тільки розглядаються практичні потреби перекладу і враховуючи, що фразеологічний рівень дуже схильний до трансформацій у тексті перекладу, стає ясно, що транспозиції, модуляції та інші процедури, що широко використовуються в перекладі, часто можуть призвести до спотворення граматичної структури. зміни у тексті, який переклали. Водночас часто буває складно знайти більш відповідний переклад зазначеної фрази (типу all out of the blue або подібний) у відкритих джерелах, доступних через пошукову систему, наприклад Google. Однак цей приклад (як і багато інших можливих подібних ілюстрацій) дозволяє нам ігнорувати граматичну структуру як вимогу виключити фразеологічні синоніми.

Додаткові питання визначення фразеологічних одиниць були розглянуті Родрігесом-Піньєро, який поставив питання про те, чи можуть фразеологічні

одиниці бути варіантами тих самих ФО і чи слід розглядати ФО з різним розподілом і семантичним поєднанням [7]. Зі свого боку, Добровольський та Баранов зазначають, що невідповідність образів у кореферентних словосполученнях викликає питання про їхню синонімічність; крім того, питання про квазісинонімію має бути визначене і в галузі фразеології [8]. Таким чином, Піньєро робить висновок, що труднощами, з якими стикаються дослідники, є відсутність критеріїв для класифікації деяких ФО як таких (просто словосполучень або стійких словосполучень), їхня різна синтагматична комбінаторика, приналежність до різних областей вживання, а також їхня полісемія.

Отже, можна сказати, що науково методичний апарат автоматичної класифікації фразеологічних одиниць досліджувати досить важко і на момент нашого дослідження не існує ідеально працюючих методологій, які могли би без похибки та спотворень ідентифікувати та класифікувати фразеологічні одиниці в будь-якому англомовному тексті.

1.2.1 Аналіз існуючих методів автоматичної класифікації фразеологізмів

Існує щонайменше три різні методи класифікації фразеологізмів у необробленому тексті: застосування побудованих «локальних» граматик, використання словників та використання статистичних процесів.

Методологія, що використовується для організації синоптичної схеми, що служить точками входу для користувачів, заснована на спостереженні та класифікації. Спочатку кожній фразеологічній статті надається набір дескрипторів (тегів, ключових слів). Ці дескриптори призначаються спочатку інтуїтивно, але в подальших етапах відповідно до раніше використаними дескрипторами. Як тільки реляційна база даних готова, програма вибирає дескриптори, що найчастіше зустрічаються, щоб використовувати їх як ключові слова. Ці ключові слова є для користувача точкою входу на перший ієрархічний рівень. Однорівнева схема може

бути достатньою для деяких цілей і забезпечує доступ до всіх фразеологізмів за вибраним ключовим словом. Для організації ієрархії у дворівневу схему, тобто для забезпечення доступу до фразеологізмів через дескриптори та субдескриптори, програма враховує спільну появу дескрипторів. Для кожного дескриптора першого рівня програма вибирає найбільш часто зустрічаються з ним в одних і тих записах. Найчастіші зберігаються як піддескриптори. У міру поповнення словника новими фразеологізмами та відповідними дескрипторами програма реорганізовуватиме синоптичну схему на основі факторів частотності та збігу. Далі докладно пояснюємо кожен крок. Наше загальне припущення ґрунтується на тому, що універсальна синоптична схема, навіть якби вона була можлива, навряд чи була б придатною для використання в ономасіологічних словниках, орієнтованих на практичні цілі. Навпаки, синоптична схема, згенерована ad hoc (тобто динамічно скомпільована), могла б бути набагато більш практичною і працездатною і краще задовольняла би критерію інтерактивності, тоді як незмінні статичні схеми, що використовуються досі, позбавлені такої можливості. Як емпіричний матеріал для експерименту ми вибрали записи бази даних. Оскільки статті словника містять семантичні дескриптори, вони є основою для розрахунку семантичної дистанції та пошуку фразеологічних синонімів. Семантична відстань тут розглядається як величина, прямо пропорційна кількості дескрипторів, що збігаються. Відомо, що існує набір більш конкретних способів розрахунку семантичної дистанції, таких як методи Арапова-Рацевої або Шрейдера, і цей вибір аргументував Фокін у статті розширення функціональності електронних словників», як зручного для цього виду словника. Сучасні обчислювальні лексикографічні ресурси надають дослідникам постійно зростаючу кількість конкретних інструментів для розрахунку семантичних подібностей з різноманітними цілями. Зокрема Купер описує результати обчислення семантичної дистанції, виконані на основі порівняння структур у статтях двомовних словників. Незважаючи на значний прорив у галузі лексичної семантики в контексті комп'ютерної лексикографії, рідкісні роботи, в яких згадується проблема фразеології, та ще менше робіт, присвячених

ономасіологічним словникам. У таблиці 1 ми показуємо фрагмент із бази даних, що містить 7 записів:

Таблиця 1 - приклад перекладу ФО

English	Ukrainian
to know something on good authority	з перших рук
to get to the point, to get down to brass tracks	переходити до справи
sharp tongue, viper’s tongue	гострий язик
to lay cards on the table	відкривати карти

ФО у крайньому лівому стовпці табл. 2 були оброблені як фразеологічні статті, тобто приймаючи на віру їхнє метафоричне, а не буквальне значення. Дескриптори в останній колонці Таблиці 1 надано кожній фразеологічній статті вручну; спочатку інтуїтивно, відповідно до його семантики та контексту використання, а потім замінюється найбільш часто використовуваним при описі декількох записів. Виявилося, що деякі дескриптори не знайшли широкого застосування. Ця процедура здійснювалася за Методикою лексикографічного портрета: спочатку інтуїтивно виділяються лексикографічні компоненти, серед яких виділяються опорні категорії [10]. Число дескрипторів повинно бути обмежене певним набором (інакше було б неможливо встановити спільний знаменник серед описів ФО), щоб їх можна було багаторазово використовувати у базі даних . Інакше кажучи, цей набір діятиме як свого роду семантичний алфавіт чи визначальний словник.

Інтуїтивні критерії присвоєння дескрипторів фразеологізмам, засновані на найбільш уживаних лексемах, можуть здатися спрощуючими і збіднювальними, оскільки найбільш уживані лексеми не обов'язково входять до найбільш уживаних

категорій. Деякі дослідники попереджають про таку проблему, як полісемія лексем, які є новопридбаними значеннями, що використовуються при визначенні лексики їх поєднань або словосполучень. Крім того, визначальна лексика повинна також включати деякі абстрактні поняття, такі як властивість, явище, якість, які не входять до числа слів, що найчастіше використовуються. Інші пропозиції розрахунку семантичної дистанції засновані на вивченні бази даних лексичних відносин WordNet, оскільки семантичні відносини лексеми можуть відображати їх семантичні властивості. Наприклад, Кенетт, Леві, Анакі та Фауст зазначають, що семантичну відстань можна виміряти, обчисливши «кількість кроків, необхідних для переходу від одного слова до іншого». Аналогічний підхід, що включає використання зумовленої ієрархії слів, в якій слова, значення та стосунки з іншими словами зберігаються в деревоподібній структурі, було досліджено Паваром та Маго [9]. Простий аналіз списку опорних дескрипторів може допомогти виявити важливі особливості: деякі з дескрипторів, що найчастіше зустрічаються, є досить синонімічними (абсурдність і нісенітниця, скритність і обман), що спонукало нас до подальшого перегляду семантичного опису. З іншого боку, використання синонімів у синоптичній схемі має певний сенс, тому що деякі користувачі можуть захотіти шукати потрібну фразу за абсурдністю ідеї, а деякі – за абсурдністю. Таким чином ця дилема відкрита для обговорення. Варто зазначити, що деякі дескриптори становлять антиноміальні пари (наприклад, ризик-обережність), але більшість із них не є такими. Наприклад, хоча дескриптори бідність, смерть і критика є одними із самих частих просто зменшуючи кількість понять. В ідеалі кількість дескрипторів на кожному рівні має дорівнювати кореню n-го ступеня, тобто:

$$d = \sqrt[n]{N}, \tag{1.1}$$

де d – кількість дескрипторів на кожному рівні

n – кількість рівнів

N – загальна кількість дескрипторів.

Зазначимо, що ця формула має бажаний і приблизний характер, оскільки деякі дескриптори можуть належати до кількох груп одночасно, тобто d на практиці дасть більше, ніж теоретично розраховане. Словники також є поширеним методом локалізації фразеологізмів, але цього недостатньо, коли йдеться про виявлення нових стійких виразів, тобто тих, що не були збережені у словнику. Побудова локальних граматик — це метод, заснований на знаннях, який забезпечує широке охоплення, тому що він може знайти нові ФО, які мають аналогічні лінгвістичні структури, ніж ті, що розглядалися під час побудови граматик. Статистичні методи зазвичай використовують частотність термінів у документах для пошуку свідчень мовних явищ. Незважаючи на те, що ми використовуємо термін «фразеологічна одиниця» у нашій роботі, ми знаємо, що існує низка досліджень, в яких використовується термін «багатослівний вираз». Однак у літературі зустрічаються й інші роботи, які використовують іншу термінологію, що стосується фразеологізмів, тому вкажемо деякі з них таким чином. У [15] автори пропонують статистичну міру до розрахунку ступеня прийнятності легких фразеологічних конструкцій, засновану з їхніми лінгвістичними властивостями. Цей показник показує хорошу кореляцію з оцінками людини на невидимих тестових даних. Більш того, вони виявляють, що їхній захід корелює сильніше, коли потенційні доповнення конструкції розділені на семантично подібні класи. Їхній аналіз демонструє системний характер напів продуктивності цих конструкцій.

Пол Кук та ін. представляє набір даних VNC-Tokens, ресурс майже 3000 англійських поєднань дієслів та іменників, анотованих щодо того, чи є вони буквальними чи ідіоматичними [16]. Ці автори почали з набору даних, використаного Фазлі і Стівенсоном [17], який включає список ідіоматичних поєднань дієслово-іменник (VNCs), і вони виявили, що приблизно половина цих виразів досить часто засвідчена в їх буквальному сенсі в Британському національному Корпус (BNC)4. Їх дослідження ґрунтується на спостереженні, що

ідіоматичне значення має тенденцію виражатися в невеликій кількості кращих лексико-синтаксичних моделей, які називаються канонічними формами. У [17] автори досліджують лексичну та синтаксичну гнучкість класу ідіоматичних виразів. Вони розробляють заходи, що ґрунтуються на таких лінгвістичних властивостях, і демонструють, що ці заходи на основі статистичного корпусу можуть бути успішно використані для розрізнення ідіоматичних поєднань від не ідіоматичних. Вони також пропонують процес автоматичного визначення того, в яких синтаксичних формах може зустрічатися конкретна ідіома і, отже, повинна бути включена до її лексичного уявлення. Використовуваний підхід до ідентифікації ФО заснований на методах контрольованого машинного навчання, області штучного інтелекту, що стосується створення та вивчення обчислювальних систем, які можуть навчатися на контрольованих даних. Методи контрольованого машинного навчання здатні вивчати людський процес ідентифікації вербальних фразеологізмів на основі ознак, переданих класифікатором, за допомогою анотованого вручну корпусу. Щоб мати уявлення про тип класифікатора, який може найкраще вирішувати проблему автоматичного виявлення ФО, ми вибрали один алгоритм навчання з чотирьох різних типів класифікаторів: байесовський, лінивий, функції та дерева.

Методи контрольованого машинного навчання припускають, що ми маємо контрольовані дані, з яких вони можуть отримати знання. У цьому випадку потрібні корпуси, анотовані експертами вручну із зазначенням наявності чи відсутності у тому чи іншому тексті дієслівної фразеологічної одиниці.

Таким чином, ми побудували набір даних для експериментів, обравши низку новин, що містять і не містять фразеологічні вербальні одиниці.

1.2.2 Аналіз існуючого програмного забезпечення автоматичної класифікації фразеологізмів

Для автоматичної класифікації фразеологічних одиниць доступно кілька основних програмних засобів:

Sketch Engine: це веб-інструмент керування та аналізу корпусів, який надає різні функції для обробки мови, зокрема автоматичну класифікацію фразеологічних одиниць [11]. Він використовує статистичний аналіз і алгоритми машинного навчання для ідентифікації та класифікації фразеологічних одиниць у корпусі тексту. Лінгвістичне дослідження та підрахунок слів (LIWC): LIWC — це програмний інструмент, розроблений Джеймсом В. Пеннебейкером, який забезпечує аналіз тексту та можливості обробки мови. Він містить функцію ідентифікації та класифікації фразеологічних одиниць у корпусі тексту на основі створених правил і статистичного аналізу [12]. ConText: ConText — це програмний інструмент, розроблений клінікою Mayo, який надає можливості обробки природної мови, включаючи автоматичну класифікацію фразеологічних одиниць. Він використовує алгоритми машинного навчання для визначення та класифікації фразеологічних одиниць у клінічному тексті. Набір інструментів природної мови (NLTK) [13]: NLTK — це бібліотека Python, яка надає різні функціональні можливості для обробки природної мови, включаючи автоматичну класифікацію фразеологічних одиниць. Він містить модуль для ідентифікації та класифікації фразеологічних одиниць на основі рукотворних правил і статистичного аналізу [14]. PhraseDetective: PhraseDetective — це веб-програмний інструмент, який забезпечує автоматичну класифікацію фразеологічних одиниць у корпусі тексту. Він використовує комбінацію статистичного аналізу та алгоритмів машинного навчання для ідентифікації та класифікації фразеологічних одиниць.

Загалом, існуючі програмні інструменти для автоматичної класифікації фразеологічних одиниць надають низку функціональних можливостей і можливостей, від ручних підходів на основі правил до більш складних підходів на

основі машинного навчання. Проте все ще існує потреба в подальших дослідженнях і розробках у цій галузі, щоб підвищити точність і ефективність цих інструментів і впоратися зі складністю та мінливістю мови.

Під час аналізу даних часто використовують програмне забезпечення лінгвістичного аналізу, після чого проводяться статистичні розрахунки.

Що стосується програмного забезпечення, для обробки текстового вмісту корпусу використовуються версія 5, розроблена Скоттом лінгвістико-статистичного інструменту WordSmith Tools (WST).

Усі тексти спочатку перетворюються на формат «простий текст» з розширенням «txt», щоб спростити обробку даних програмним забезпеченням. Проте неможливо розглянути формули, схеми, цифри та графіки у цьому перетворенні. Проте субтитри відновлюються.

Виникла потреба скласти корпус з академічним змістом англійською мовою, який значною мірою представляв би академічний дискурс. Розглянутий корпус містить лише письмові форми восьми основних галузей знання. Для аналізу текстів можна використовувати статті, журнали та довідники, зібрані з Інтернету та доступні безкоштовно. Таке рішення можливо тоді, коли тексти не мають впливу інших іноземних мов. Спеціально створений для цього дослідження корпус отримав назву Academic Corpus of English (ACE).

Важливо відзначити, що нерівні розміри між областями та під областями не погіршують аналіз, якщо використовувані статистичні процедури враховують розбіжності та «виправляють» спотворення. Згідно з Sinclair [18], мають бути встановлені певні критерії, що лежать в основі вибору текстів, що становлять корпус. ACE можна описати так: текстовий режим (написаний та отриманий з електронного носія); тип тексту (довідники та наукові статті); знання тексту (академічний); різноманітність мови (спеціальна мова); регіон текстів (в основному США та Великобританія) та дата текстів (з 2000 по 2012 рік).

Спеціалізований корпус, складений для цього дослідження, містить 122464043 токена або входження. Виявлення найпоширеніших внутрішньо- та

міждисциплінарних фразеологізмів академічного спілкування вимагає проведення низки розрахунково-статистичних процедур. Першим кроком у створенні файлів потрібно створити таблиці кожної основної області. Таблиця являє собою файл, згенерований WST, в якому записується положення всіх слів в корпусі, що досліджується, що дозволяє запитувати n-грами або рядки відповідності. У додатку Г було розглянуто головне програмне забезпечення, яке відповідає кожному методу автоматичної класифікації фразеологічних одиниць. Провівши аналіз цієї таблиці, можна визначити основні можливості та суть кожного інструменту програмного забезпечення.

У додатку Д наведено приклад класифікації фразеологічних одиниць із словника [19] за допомогою ПЗ (Sketch Engine). Для класифікації було обрано 100 ФО. Результати класифікації виявились такими. У результаті роботи ПЗ (Sketch Engine) 27 ФО було віднесено до групи 1, 9 – до групи 2, 6- до групи 3, 19 - до групи 4, 16 - до групи 5 та 23 ФО, які не класифікувались за конкретною ознакою. Час автоматичної класифікації склав 27 секунд.

1.3 Постановка задачі

Завданням даного дослідження є удосконалення методу класифікації фразеологічних одиниць в англомовних текстах для підвищення ефективності автоматичної класифікації фразеологічних одиниць англомовних текстів та зменшення кількості помилкових тлумачень.
Для досягнення мети необхідно вирішити такі задачі:

1) Аналіз існуючих методів автоматичної класифікації фразеологізмів
2) Аналіз існуючого програмного забезпечення автоматичної класифікації фразеологізмів
3) Обгрунтування методу класифікації фразеологізмів англомовних текстів на основі інтеграції алгоритмів машинного навчання

4) Визначення переваг та недоліків розглянутих методів автоматичної класифікації фразеологізмів
5) Розробка структури інформаційної системи автоматичної класифікації фразеологізмів
6) Проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів
7) Розробка алгоритму роботи системи класифікації фразеологічних одиниць
8) Програмна реалізація методу автоматичної класифікації фразеологічних одиниць англомовних текстів

1.4 Висновки

У даному розділі ми дослідили категорії та поняття, які охоплюють предметне середовище автоматичної класифікації фразеологічних одиниць. Була вказана актуальність даного дослідження та встановлено, що людині все складніше обробляти велику кількість інформації та вирішувати задачі пов'язані із класифікацією, а зростання її кількості та одночасне зростання доступної обчислювальної потужності комп'ютерів дозволяють автоматизувати розв’язання задачі класифікації. Окрім цього було описано процес діяльності, який автоматизується, визначено, які етапи потрібно пройти, аби вирішити задачу, наведено коротку характеристику методів, якими пропонується вирішувати поставлену задачу. Проведено аналіз існуючих методів класифікації фразеологічних одиниць, які можна розділити на три основні групи: на основі правил, на основі статистики та на основі машинного навчання. Визначивши методи, було виділено програмне забезпечення, за допомогою якого реалізуються методи автоматичної класифікації фразеологізмів. Виявилось, що головними інструментами можна виділити такі як, лінгвістичне дослідження та підрахунок слів (LIWC), ConText та набір інструментів природної мови (NLTK). Доведено, що

людині стає все важче обробляти великі обсяги даних та вирішувати завдання, пов'язані з класифікацією, збільшення її кількості та зростання доступних обчислювальних потужностей комп'ютера дає можливість працювати в автоматичному режимі. Крім того, описано процес автоматизації, визначено кроки, які необхідно зробити для вирішення проблеми, та наведено короткий опис методів, запропонованих для вирішення проблеми [20]. У цьому розділі ми представили досягнення в галузі автоматичної ідентифікації присутності фразеологізмів у англомовних текстах. Ми вважаємо особливо важливим набір запропонованих гіпотез, тому що вони керуватимуть поточним дослідженням цієї роботи.

2. ПРОЕКТУВАННЯ СТРУКТУРИ ІНФОРМАЦІЙНОЇ СИСТЕМИ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ

2.1 Метод автоматичної класифікації фразеологізмів англомовних текстів на основі інтеграції методу на основі правил з алгоритмами машинного навчання

Класифікація тексту - це розумний розподіл тексту на категорії. А використання машинного навчання для автоматизації цих завдань робить весь процес надшвидким і ефективним. Штучний інтелект і машинне навчання - це, мабуть, найбільш ефективні технології для вирішення зазначеного завдання, які набрали обертів останнім часом. Проте якщо зробити поєднання методу на основі правил та штучного інтелекту, то ми зможемо розширити горизонти можливостей методу на основі машинного навчання та надати його рішенням конструктивної логіки. У цьому розділі розглядаються питання про технологію, додатки, кастомізацію та сегментацію, пов'язані з нашою інтеграцією двох методів автоматизованої класифікації фразеологічних одиниць [21]. Аналіз намірів, емоцій і настроїв текстових даних є одними з найважливіших частин класифікації текстів. Ці випадки використання викликали значний резонанс серед ентузіастів машинного інтелекту. Текстовий класифікатор може працювати з різноманітними текстовими наборами даних. Ви можете навчати класифікатор на тегованих даних або оперувати сирим неструктурованим текстом. Обидві ці категорії мають безліч застосувань. Контрольована класифікація тексту виконується після того, як ви визначили категорії класифікації. Вона працює за принципом навчання і тестування. Ми подаємо марковані дані для роботи алгоритму машинного навчання. Алгоритм навчається на наборі маркованих даних і видає бажаний результат (попередньо визначені категорії) [22]. На етапі тестування алгоритм отримує не спостережувані дані і класифікує їх за категоріями на основі даних, отриманих під час навчання. Описавши основні методи автоматичної класифікації у розділі 1, ми виявили, які є переваги та недоліки кожного методу для того, щоб

розробити метод покращення ефективності автоматичної класифікації фразеологізмів. Основними недоліками машинного навчання є те, що можлива похибка перекладу, проблемність швидко актуалізуватися до нових змін та відсутність прозорості, що ускладнює логічне обгрунтування рішення, до якого прийшов алгоритм. Для того, щоб виправити недоліки та збільшити ефективність машинного методу, так як штучний інтелект та машинне навчання в майбутньому буде одним з найбільш дієвих методів, ми пропонуємо поєднати метод на основі правил, для того, щоб алгоритм спирався на вже доступні варіанти рішення запиту та надавав логіку в свої рішення, використовуючи правила, якими він користувався під час класифікації ФО. Однією з основних переваг машинного методу є масштабованість, що дуже складно дається використовуючи метод правил, тому що підхід на основі правил обмежений можливістю розробляти правила вручну [23]. Проте, моделі машинного навчання можна навчити на великих обсягах даних, що робить їх масштабованими для аналізу тексту з різних джерел і доменів. Штучний інтелект не розрахований на використання базових правил, тому що він вчиться аналізувати матеріал самостійно та будувати свої рішення великою кількістю варіацій одночасно, але залучивши деякі правила, це могло б пришвидшити рішення базових запитів покращити прозорість цього методу. Поєднання методів, заснованих на правилах, і методів машинного навчання для автоматичної класифікації фразеологічних одиниць в англомовних текстах може призвести до більш точної та ефективної системи.

Методи, засновані на правилах, передбачають використання заздалегідь визначених правил або шаблонів для ідентифікації та класифікації фразеологічних одиниць у тексті [24]. Цей метод базується на лінгвістичних знаннях і вимагає ручного створення правил і шаблонів. Хоча він може бути ефективним у визначенні певних типів фразеологічних одиниць, він може бути менш точним у визначенні менш поширених або складних виразів. Приклад правил і шаблонів класифікації фразеологічних одиниць з технічної точки зору такий:

1. Правило: Фразеологізм повинен складатися з двох або більше слів, які функціонують як єдине значення.Приклад: зіграти в ящик (kick the bucket), ні пуху ні пера (break a leg).

2. Правило: Фразеологізм повинен бути сталим і своєрідним за своїм значенням, тобто його значення не можна передбачити зі значень слів-складників. Приклад: проговоритись (spill the beans), влучити прямо в точку або сказати дуже чітко (hit the nail on the head).

3. Зразок: Фразеологізм можна ідентифікувати за його формою, тобто слова в ньому повинні стояти у встановленому порядку. Приклад: кредит довіри (the benefit of the doubt).

4. Зразок: фразеологізм можна ідентифікувати за його колокаційними обмеженнями, тобто він може зустрічатися лише з певними словами чи в певному контексті. Приклад: термінові новини (breaking news).

Ці правила та шаблони можна використовувати в системах обробки природної мови (NLP) для автоматичної ідентифікації та класифікації фразеологізмів у тексті [25]. Поєднуючи ці два методи, можна використовувати сильні сторони обох. Методи, засновані на правилах, можна використовувати для ідентифікації поширених і чітко визначених фразеологічних одиниць, тоді як машинне навчання можна використовувати для ідентифікації більш складних і рідкісних виразів. Наприклад, система може використовувати підхід, заснований на правилах, для ідентифікації поширених словосполучень, таких як «міцна кава» або «швидка машина», і використовувати машинне навчання для визначення менш поширених ідіом, таких як «кинути відро» або «стримуйте коней». У додатку Ґ зображена характеристика методу в порівнянні з іншими та прогнозований ефект від удосконалення.

Загалом поєднання методів на основі правил і машинного навчання може підвищити точність і ефективність автоматичної класифікації фразеологічних одиниць в англомовних текстах. Щоб поєднати сильні сторони обох підходів, один підхід полягає у використанні методів на основі правил для ідентифікації та

виділення конкретних типів фразеологічних одиниць, а потім використання методів машинного навчання для їх класифікації на основі їхніх семантичних і синтаксичних властивостей. Наприклад, можна використовувати метод, заснований на правилах, для ідентифікації іменників, які складаються з іменника та прикметника, а алгоритм машинного навчання можна навчити класифікувати ці іменники на основі їхньої семантичної подібності до інших відомих фразеологічних одиниць. Існує кілька алгоритмів і технічних засобів, які можна використовувати для досягнення цієї комбінації методів на основі правил і машинного навчання. Наприклад, Natural Language Toolkit (NLTK) у Python надає набір інструментів для обробки природної мови на основі правил і машинного навчання. NLTK містить інструменти для зіставлення шаблонів, синтаксичного аналізу та виділення ознак, а також алгоритми для класифікації, кластеризації та пошуку інформації [26]. Іншим популярним інструментом обробки природної мови є spaCy, який також надає комбінацію методів на основі правил і машинного навчання для ідентифікації та класифікації фразеологічних одиниць. SpaCy містить систему зіставлення на основі правил для виявлення конкретних шаблонів у тексті, а також конвеєр машинного навчання для навчання та оцінки користувацьких моделей для класифікації та інших завдань [27].

Таким чином, поєднання методів, заснованих на правилах, і машинного навчання може підвищити точність і ефективність автоматичної класифікації фразеологічних одиниць в англомовних текстах шляхом поєднання сильних сторін обох підходів. Для реалізації цих підходів і досягнення високої точності та ефективності можна використовувати такі технічні засоби, як NLTK і spaCy.

Часто найбільшою перешкодою для використання машинного навчання є відсутність набору даних [28]. Є багато людей, які хочуть використовувати ШІ для класифікації даних, але для цього потрібно створити набір даних, що створює ситуацію, схожу на проблему "курка-яйце".

В останніх дослідженнях пропонують метод навчання з нуля на тексті, коли алгоритм, навчений вивчати зв'язки між реченнями та їхніми категоріями на

великому зашумленому наборі даних, можна узагальнити для нових категорій або навіть нових наборів даних. Ми називаємо цю парадигму "Навчився один раз - тестуй будь-де" [29]. Також пропонується кілька алгоритмів нейронних мереж, які можуть скористатися цією методологією навчання і отримати хороші результати на різних наборах даних. Найкращий метод використовує модель LSTM для завдання вивчення взаємозв'язків [30]. Ідея полягає в тому, що якщо можна змоделювати поняття "приналежності" між реченнями і класами, то знання будуть корисними для невидимих класів або навіть невидимих наборів даних.

Виділення ознак передбачає виявлення характеристик фразеологічних одиниць, за допомогою яких можна відрізнити їх від інших типів виразів. Ці ознаки можуть включати частоту появи, довжину виразу, наявність певних слів або частин мови та інші мовні властивості.

Класифікація передбачає використання алгоритмів машинного навчання для групування схожих фразеологічних одиниць разом на основі ознак, визначених під час процесу виділення ознак. Процес класифікації може бути контрольованим або неконтрольованим. У контрольованому навчанні алгоритм навчається на позначеному наборі даних, тоді як у неконтрольованому навчанні алгоритм визначає шаблони даних без попереднього знання категорій [31]. Існує кілька алгоритмів машинного навчання, які можна використовувати для автоматичної класифікації фразеологізмів, зокрема навчання під контролем, навчання без контролю та напів контрольоване навчання [32]. Навчання під наглядом передбачає навчання алгоритму машинного навчання на маркованому наборі даних фразеологічних одиниць і відповідних їм категорій. Алгоритм вчиться визначати характеристики та властивості різних типів фразеологізмів і використовує ці знання для класифікації нових фразеологізмів. Наприклад, алгоритм керованого навчання можна навчити на наборі даних ідіоматичних виразів і відповідних категорій (наприклад, ідіоми, пов'язані з їжею, ідіоми, пов'язані з погодою тощо), щоб точно класифікувати нові приклади ідіоматичних виразів на основі їхніх семантичних і синтаксичних властивостей [33]. Навчання без нагляду передбачає

навчання алгоритму машинного навчання на немаркованому наборі даних фразеологічних одиниць, що дозволяє алгоритму виявляти закономірності та подібності в даних без будь-яких попередніх знань про категорії. Цей підхід корисний, коли категорії фразеологізмів невідомі, або коли категорій занадто багато, щоб їх можна було позначити вручну. Наприклад, алгоритм неконтрольованого навчання можна використовувати для об'єднання схожих ідіоматичних виразів у групи на основі їхніх семантичних і синтаксичних властивостей [34]. Напівконтрольоване навчання - це комбінація контрольованого і неконтрольованого навчання, де алгоритм навчається на невеликому наборі даних з мітками і великому наборі даних без міток. Алгоритм використовує маркований набір даних для вивчення характеристик і властивостей категорій, а потім застосовує ці знання до немаркованого набору даних для виявлення схожих екземплярів. Цей підхід може бути корисним, коли маркований набір даних невеликий, або коли маркування всього набору даних неможливе [35]. Якщо ми поєднаємо метод машинного навчання з методом на основі правил, ми зможемо використати гібридний підхід для підвищення точності автоматичної класифікації фразеології. Ось кілька алгоритмів, які можна використовувати для цієї мети.

Дерево рішень із правилами: Алгоритми дерева рішень використовують деревоподібну модель рішень та їхніх можливих наслідків. Поєднуючи дерева рішень із правилами, ми можемо створити гібридний підхід, який дозволяє нам включити правила, що стосуються предметної області, які можуть допомогти уточнити результат алгоритму дерева рішень [36]. Випадковий ліс із правилами. Алгоритми випадкового лісу — це метод ансамблевого навчання, який використовує декілька дерев рішень для підвищення точності прогнозів. Поєднуючи випадкові ліси з правилами, ми можемо створити гібридний підхід, який дозволить нам включити доменно-спеціальні правила, які можуть допомогти уточнити результат алгоритму випадкового лісу. Нейронна мережа з правилами: нейронні мережі — це алгоритм машинного навчання, який може навчитися розпізнавати шаблони в даних. Поєднуючи нейронні мережі з правилами, ми

можемо створити гібридний підхід, який дозволить нам включити доменно-спеціальні правила, які можуть допомогти уточнити результат нейронної мережі [37]. Машина опорних векторів із правилами. Машини опорних векторів — це тип алгоритму контрольованого навчання, який можна використовувати для класифікації чи регресії. Поєднуючи машини опорних векторів із правилами, ми можемо створити гібридний підхід, який дозволяє нам включити правила, що стосуються предметної області, які можуть допомогти уточнити вихідні дані машини опорних векторів. Наївний Байєс із правилами: Наївний Байєс — це простий імовірнісний класифікатор, який базується на теоремі Байєса. Поєднуючи наївний байєсівський класифікатор із правилами, ми можемо створити гібридний підхід, який дозволить нам включити предметно-спеціальні правила, які можуть допомогти уточнити результат наївного байєсовського класифікатора [38]. Одним із поширених алгоритмів машинного навчання, що використовується для автоматичної класифікації фразеологічних одиниць, є метод опорних векторів (SVM). SVM - це алгоритм навчання під наглядом, який передбачає визначення оптимальної гіперплощини, що розділяє різні категорії фразеологічних одиниць. Гіперплощина визначається набором коефіцієнтів, які вивчаються в процесі навчання [39]. Алгоритм використовує ці коефіцієнти для класифікації нових фразеологізмів на основі їхнього розташування відносно гіперплощини. Іншим алгоритмом машинного навчання, який використовується для автоматичної класифікації фразеологізмів, є алгоритм k-найближчих сусідів (KNN). KNN - це алгоритм неконтрольованого навчання, який передбачає визначення k найближчих сусідів нового екземпляра на основі їхньої схожості з цим екземпляром. Потім алгоритм відносить новий екземпляр до категорії, яка є найбільш поширеною серед його k-найближчих сусідів. Цей підхід корисний, коли категорії невідомі або коли категорій занадто багато, щоб їх можна було позначити вручну [40]. Отже, інтеграція алгоритмів машинного навчання є ефективним підходом до автоматичної класифікації фразеологізмів в англійських текстах. Алгоритми машинного навчання, такі як контрольоване навчання, неконтрольоване навчання

та напів контрольоване навчання, можна використовувати для точної класифікації різних типів фразеологічних одиниць на основі їхніх семантичних і синтаксичних властивостей.

2.2 Розробка структури інформаційної системи автоматичної класифікації фразеологізмів

Розробка інформаційної системи для автоматичної класифікації фразеологічних одиниць передбачає використання математичних формул та алгоритмів для аналізу та класифікації цих виразів. У даному параграфі ми розглянемо структуру такої системи та наведемо приклади використання математичних формул і алгоритмів.

Структуру інформаційної системи для автоматичної класифікації фразеологічних одиниць можна розділити на три основні компоненти: попередня обробка даних, вилучення ознак і класифікація [41]. Наприклад, уявімо, що у нас є набір фразеологічних одиниць, які були позначені як ідіоми або словосполучення. Ми можемо використати алгоритм k-NN, щоб класифікувати нову фразеологічну одиницю як ідіому або словосполучення на основі її особливостей. Припустимо, нова фразеологічна одиниця - це "kick the bucket". Ми можемо обчислити її характеристики, такі як частота і довжина, і визначити її k найближчих сусідів у наборі даних. Якщо більшість сусідів позначені як ідіоми, ми можемо класифікувати "kick the bucket" як ідіому. Попередня обробка даних передбачає очищення та організацію даних перед тим, як вони потрапляють до алгоритмів вилучення ознак та класифікації. Цей процес включає видалення нерелевантної інформації, наприклад, стоп-слів, і перетворення тексту у формат, який можна легко обробити алгоритмами. Вилучення ознак передбачає визначення характеристик фразеологізмів, за якими їх можна відрізнити від інших типів виразів. Ці ознаки можуть включати частоту вживання, довжину виразу, наявність певних слів або частин мови та інші лінгвістичні властивості. Для вилучення ознак

з даних можна використовувати математичні формули. Наприклад, частоту вживання фразеологізму можна обчислити за такою формулою:

$$f = \frac{n}{N}, \tag{2.1}$$

де n - це кількість входжень фразеологізму;

N - загальна кількість слів у тексті.

Аналогічно можна розрахувати довжину фразеологізму за такою формулою:

$$L = w, \tag{2.2}$$

де L – це довжина;

w – це кількість слів у фразеологізмі

Інші лінгвістичні властивості, такі як наявність певних слів або частин мови, можна визначити за допомогою методів обробки природної мови (NLP) [42]. Класифікація передбачає групування схожих фразеологізмів на основі ознак, виявлених під час процесу вилучення ознак. Математичні алгоритми можна використовувати для класифікації фразеологізмів на основі їхніх особливостей. Одним із найпоширеніших алгоритмів є метод k-найближчих сусідів (k-NN), який передбачає визначення k найближчих сусідів нової фразеологічної одиниці на основі її особливостей і віднесення її до тієї категорії, до якої належить більшість її сусідів [43].

Модель «мішок слів» (bag-of-words, BOW) і модель вбудовування слів – два популярні методи, що використовуються для вилучення ознак. У моделі BOW кожна фразеологічна одиниця представлена у вигляді вектора частот слів. Підраховується кількість разів, коли кожне слово з'являється в одиниці, і результуючий вектор має розмірність, що відповідає кількості унікальних слів у корпусі [44]. Наприклад, фразеологізми «A piece of cake», «Break a leg» і «Hit the sack» будуть представлені наступним чином:

[1, 1, 1, 0, 0, 0, 0]

[0, 1, 0, 1, 0, 0, 0]

[0, 1, 0, 0, 0, 1, 0]

Кожен вимір вектора представляє частоту відповідного слова у фразеологізмі.

Модель вбудовування слів представляє кожне слово у вигляді вектора у високорозмірному просторі. Векторне представлення кожної фразеологічної одиниці створюється шляхом усереднення векторів слів, що входять до складу одиниці. Наприклад, векторне представлення слів фразеологізмів «A piece of cake», «Break a leg» та «Hit the sack» буде таким:

[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]

[0.5, 0.6, 0.7, 0.8, 0.9, 0.1, 0.2]

[0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3]

Кожен вимір вектора представляє середнє значення відповідного вектора слів у фразеологізмі.

Після вилучення ознак наступним кроком є класифікація. Одним із прикладів класифікації фразеологічних одиниць за чіткими ознаками є розрізнення ідіом і словосполучень за їхньою структурою та передбачуваністю. Наприклад, ідіома «kick the bucket» означає «померти», і її неможливо зрозуміти, шукаючи значення «удар» або «відро» в словнику. Ідіоми часто мають метафоричне чи образне значення, яке не пов’язане з їхнім буквальним значенням [45]. З іншого боку, словосполучення. Їх можна певною мірою передбачити на основі значення окремих слів. Наприклад, словосполучення «міцна кава» є передбачуваним, оскільки прикметник «міцний» зазвичай використовується для опису кави, тоді як «міцний чай» також є словосполученням, оскільки «міцний» також зазвичай використовується для опису чаю. Однак словосполучення «важка кава» непередбачуване, оскільки «важка» зазвичай не використовується для опису кави. Розрізнення між ідіомами та словосполученнями на основі їх структури та передбачуваності може допомогти тим, хто вивчає мову, і перекладачам краще

зрозуміти значення та вживання цих типів фразеологічних одиниць. Ми будемо класифікувати фразеологізми за чотирма ознаками:

1) Структура: Фразеологізм «kick the bucket» є ідіомою, що означає, що він має фіксовану структуру і не може бути зрозумілий, шукаючи значення окремих його слів.

2) Семантичний зв'язок: Фразеологізм «міцна кава» є словосполученням, тобто складається зі слів, які часто зустрічаються разом за семантичною спорідненістю.

3) Функція: Фразеологізм «з іншого боку» є дискурсивним маркером, що означає, що він використовується для позначення контрасту або альтернативної точки зору.

4) Походження: Фразеологічна одиниця «faux pas» є запозиченим словом, що означає, що воно було запозичене з французької мови та зазвичай використовується в англійській мові для позначення соціальної помилки чи промаху [46].

Ми використали чотири різні ознаки для класифікації чотирьох різних фразеологічних одиниць: структура (ідіома), семантичний зв'язок (словосполучення), функція (дискурсивний маркер) і походження (запозичене слово). Це демонструє, як різні ознаки можна використовувати для класифікації фразеологічних одиниць і отримати краще розуміння їхніх властивостей і використання в дискурсі. Для класифікації можна використовувати кілька методів, зокрема метод k-найближчих сусідів, дерева рішень, машини опорних векторів (SVM) та нейронні мережі. Вибір методу класифікації залежить від розміру набору даних, складності фразеологічних одиниць і бажаної точності класифікації. У методі k-найближчих сусідів фразеологізм класифікується на основі мітки класу його k найближчих сусідів у просторі ознак. У методі дерева рішень створюється дерево, яке представляє правила прийняття рішень для присвоєння фразеологізмам міток класів. У SVM використовується гіперплощина для розділення фразеологізмів на різні класи на основі їхніх представлень ознак. У нейронних мережах модель глибокого навчання навчається на представленнях ознак фразеологізмів для прогнозування їхньої класової належності [47]. Інформаційна система, яка може автоматично класифікувати фразеологічні одиниці, має

практичне застосування в обробці природної мови, вивченні мови та перекладі [48]. Наприклад, система автоматичної класифікації англійських фразеологізмів може бути корисною у викладанні англійської мови як другої. Система може бути використана для визначення найбільш часто вживаних фразеологізмів та їхніх відповідних значень. Ця інформація може бути використана для створення навчальних матеріалів і вправ, які допоможуть студентам зрозуміти і правильно використовувати фразеологізми. Подібним чином система класифікації фразеологізмів може бути використана в галузі перекладу. Перекладацьке програмне забезпечення може використовувати систему для ідентифікації та точного перекладу фразеологічних одиниць. Розробка інформаційної системи для автоматичної класифікації фразеологічних одиниць не позбавлена труднощів. Однією з найбільших проблем є різноманітність і складність фразеологічних одиниць у різних мовах і культурах. Крім того, точність завдання класифікації сильно залежить від якості навчальних даних і вибору методів вилучення ознак і класифікації. Наприклад, модель BOW є простою та ефективною, але вона не враховує семантичні зв'язки між словами у фразеологізмах. На противагу цьому, модель вбудовування слів враховує семантичні зв'язки між словами, але вона вимагає великої кількості навчальних даних і може бути дорогою в обчислювальному плані. Тому вибір методу вилучення ознак має ґрунтуватися на характеристиках набору даних і доступних обчислювальних ресурсах. Так само вибір методу класифікації залежить від розміру набору даних, складності фразеологічних одиниць і бажаної точності завдання класифікації. Наприклад, метод k-найближчих сусідів є простим і легким у реалізації, але він може не працювати належним чином, коли набір даних великий і кількість класів велика. На противагу цьому, метод нейронних мереж є більш складним і обчислювально дорогим, але він може досягти високої точності навіть з великими і складними наборами даних [49]. Після того, як було проведено дослідження наявних інформаційних систем та визначено їх недоліки, ми навели систему, яка поєднує в собі інструменти методу на основі правил та методу на основі машинного навчання.

Інформаційна система, яка поєднує в собі метод автоматичної класифікації на основі правил і метод машинного навчання, може бути структурована таким чином.

Збір даних: система збирає дані, які включають набір текстів або документів, які необхідно класифікувати. Дані також можуть включати фразеологізми та пов’язані з ними категорії, які можна використовувати як навчальні дані для моделі машинного навчання. Попередня обробка: система попередньо обробляє дані, щоб підготувати їх до класифікації. Сюди входять такі завдання, як токенізація, видалення стоп-слова, формування основи та нормалізація. Метод на основі правил: система застосовує метод на основі правил для ідентифікації та класифікації фразеологізмів у тексті. Це можна зробити шляхом створення набору правил, які відповідають певним шаблонам або послідовностям слів, які відповідають фразеологізмам [50]. Метод на основі машинного навчання: система також використовує метод на основі машинного навчання для класифікації тексту. Це можна зробити шляхом навчання моделі класифікації на навчальних даних, яка включає фразеологізми та пов’язані з ними категорії. Потім модель машинного навчання можна використовувати для автоматичної класифікації фразеологізмів у тексті. Гібридний метод: система поєднує результати методу на основі правил і методу на основі машинного навчання для підвищення точності класифікації. Це можна зробити, застосовуючи модель машинного навчання лише до фразеологізмів, які не були класифіковані методом на основі правил, або використовуючи метод на основі правил для уточнення виходу моделі машинного навчання. Оцінка: система оцінює ефективність класифікації за допомогою таких показників, як точність, запам’ятовування та бал F1. Це можна зробити, порівнявши вихідні дані системи з набором класифікованих вручну текстів.

Вихід: система виводить результати класифікації, які можна використовувати для різних цілей, наприклад для пошуку інформації, аналізу тексту або аналізу настроїв [51]. Для більш доцільної реалізації гібридної автоматичної класифікації фразеологічних одиниць, був проведений аналіз уривку тексту із роману Ф. Скотта Фіцджеральда «Великий Гетсбі», який містить кілька

фразеологізмів: «The one on my right was a colossal affair by any standard — it was a factual imitation of some Hotel de Ville in Normandy, with a tower on one side, spanking new under a thin beard of raw ivy, and a marble swimming pool, and more than forty acres of lawn and garden. It was Gatsby's mansion. Or, rather, as I didn't know Mr. Gatsby, it was a mansion inhabited by a gentleman of that name. My own house was an eyesore, but it was a small eyesore, and it had been overlooked, so I had a view of the water, a partial view of my neighbor's lawn, and the consoling proximity of millionaires — all for eighty dollars a month» [52]. У таблиці 2.1 були визначені та прокласифіковані ФО по конкретним ознакам.

Таблиця 2.1 – Ознаки класифікації ФО

Ознаки	Приклад класифікації за гібридним методом
Ідіоми	Фразеологізм «by any standard»
Семантична класифікація	Фразеологізм «marble swimming pool»
Формульні вирази	Фразеологізм «a view of the water»
Класифікація за походженням	Фразеологізм «Hotel de Ville»
Ад'єктивні	«In the red»
Субстантивні	«Apple of my eye»

Тепер давайте проаналізуємо та класифікуємо фразеологічні одиниці в цьому уривку за чотирма різними ознаками:

1) Структурна класифікація: Фразеологізм «by any standard» є прислівниковим виразом, який має фіксовану структуру і значення, тобто його можна віднести до ідіом.

2) Семантична класифікація: Фразеологізм «marble swimming pool» є словосполученням, тобто складається зі слів, які за семантичною спорідненістю часто зустрічаються разом.

3) Функціональна класифікація: Фразеологізм «a view of the water» є сталим виразом, який передає певне значення, тобто його можна віднести до формульного виразу.

4) Класифікація за походженням: Фразеологізм «Hotel de Ville» є усталеним виразом, який відноситься до типу муніципальних будівель, які зазвичай зустрічаються у Франції, тобто його можна класифікувати як запозичення з іншої мови. Ад'єктивні фразеологізми — це багатослівні вирази, які складаються з прикметника та інших слів і мають переносне чи ідіоматичне значення, яке не обов'язково походить від буквального значення окремих слів. Вони використовуються для опису або вираження якості чи властивості людини, речі чи ситуації. Приклади прикметникових фразеологічних одиниць включають «in the red» (означає бути в боргу), «on cloud nine» (означає бути надзвичайно щасливим) [53]. Субстантивні фразеологічні одиниці — це багатослівні вирази, які складаються з іменника та інших слів і мають переносне чи ідіоматичне значення, яке не обов'язково походить від буквального значення окремих слів. Вони використовуються для позначення людей, речей або ситуацій у творчий або експресивний спосіб. Приклади субстантивних фразеологічних одиниць включають «apple of my eye» (що означає щось або когось дорогого чи коханого), «chicken-hearted» (що означає боягузливий) [54]. Ці класифікації можна використовувати для підвищення точності та ефективності автоматичної класифікації фразеологічних одиниць в англомовних текстах шляхом поєднання методів, заснованих на правилах, і алгоритмів машинного навчання. Наприклад, методи на основі правил можна використовувати для ідентифікації ідіоматичних виразів на основі їх фіксованої структури, тоді як алгоритми машинного навчання можна навчити ідентифікувати словосполучення на основі їхніх семантичних зв'язків. Комбінуючи ці методи, точність класифікації можна підвищити, а процес

можна зробити більш ефективним [55]. Таким чином, інформаційну систему, яка поєднує метод автоматичної класифікації на основі правил і метод машинного навчання на основі машинного навчання, можна структурувати таким чином, щоб використовувати переваги кожного підходу, тим самим підвищуючи точність класифікації.

2.3 Вибір апаратних засобів для автоматичної класифікації фразеологізмів

Автоматична класифікація фразеологічних одиниць в англомовних текстах є складною задачею, яка потребує значних обчислювальних ресурсів, і в галузі обробки природної мови (NLP). Одним з вирішальних факторів успішної реалізації автоматичної класифікації фразеологічних одиниць є вибір відповідного апаратного забезпечення. Апаратне забезпечення, необхідне для автоматичної класифікації фразеологізмів, включає наступні компоненти: Центральний процесор (ЦП), Оперативна пам’ять (RAM), Сховище, Графічний процессор (GPU), Мережева карта (NIC). Центральний процесор є основним компонентом комп'ютерної системи, який виконує всі завдання з обробки даних і прийняття рішень. Для автоматичної класифікації фразеологізмів центральний процесор повинен мати високу обчислювальну потужність, оскільки він повинен швидко обробляти великі обсяги текстових даних. Для автоматичної класифікації фразеологічних одиниць система повинна мати достатній обсяг оперативної пам'яті, щоб обробляти великі обсяги текстових даних і виконувати складні обчислення, яких вимагають алгоритми. Компонент сховища системи відповідає за зберігання текстових даних і алгоритмів, що використовуються для автоматичної класифікації фразеологізмів. Це сховище може бути у вигляді жорстких дисків (HDD) або твердотільних накопичувачів (SSD). У контексті автоматичної класифікації фразеологічних одиниць графічні процесори можна використовувати для прискорення навчання алгоритмів машинного навчання та підвищення загальної продуктивності системи. При виборі апаратного забезпечення для

автоматичної класифікації фразеологічних одиниць необхідно враховувати кілька факторів, зокрема розмір набору даних, складність алгоритмів і бажаний рівень продуктивності. Наприклад, система, яка обробляє великий масив текстових даних і використовує складні алгоритми, такі як глибоке навчання, може потребувати високопродуктивного процесора з декількома ядрами і великим об'ємом оперативної пам'яті. І навпаки, система, яка обробляє менший набір текстових даних і використовує простіші алгоритми, такі як машини опорних векторів (SVM), може потребувати менш потужного апаратного забезпечення. Крім того, вибір графічного процесора може суттєво вплинути на продуктивність системи. Графічні процесори з великою кількістю ядер і пам'яті можуть прискорити навчання алгоритмів машинного навчання і підвищити точність прогнозів системи.

Наприклад, NVIDIA Tesla V100 GPU – це графічний процесор високого класу, призначений для машинного навчання та завдань з інтенсивним використанням даних. Він має 5 120 ядер CUDA і 16 ГБ пам'яті, що робить його ідеальним для навчання складних моделей машинного навчання. З практичної точки зору, використання хмарних апаратних сервісів, таких як Amazon Web Services (AWS) і Microsoft Azure, може надати значні переваги для автоматичної класифікації фразеологічних одиниць. Ці сервіси пропонують доступ до високопродуктивних процесорів, графічних процесорів і компонентів зберігання даних, що дозволяє розробникам масштабувати свої апаратні ресурси відповідно до розміру набору даних і складності алгоритмів. Вибір апаратного забезпечення для автоматичної класифікації фразеологічних одиниць є важливим моментом для ефективного та результативного функціонування системи. У даному параграфі будуть розглянуті теоретичні основи та практичні приклади вибору апаратного забезпечення для автоматичної класифікації фразеологічних одиниць. Обчислювальна потужність обладнання має вирішальне значення для ефективного функціонування методів NLP. Класифікація фразеологічних одиниць передбачає складні математичні обчислення, такі як векторизація, кластеризація та класифікація. Апаратне забезпечення повинно мати достатню обчислювальну

потужність для виконання цих обчислень за розумний проміжок часу. Для цього ідеально підійде потужний процесор, наприклад, Intel i9 або AMD Ryzen 9. Вимоги до оперативної пам'яті також є важливими. Класифікація фразеологічних одиниць передбачає обробку великих обсягів даних, що може призвести до перевантаження пам'яті. Апаратне забезпечення повинно мати достатньо пам'яті для зберігання даних і виконання необхідних обчислень. Для цього рекомендується мінімум 16 ГБ оперативної пам'яті [56]. Вимоги до пам'яті апаратного забезпечення залежать від розміру оброблюваних даних і тривалості їх зберігання. Класифікація фразеологічних одиниць передбачає обробку великих обсягів текстових даних, які можуть займати значний обсяг пам'яті. Апаратне забезпечення повинно мати достатню ємність для зберігання даних і будь-яких проміжних результатів. Для цього рекомендується використовувати твердотільний накопичувач (SSD) ємністю не менше 512 ГБ. Вимоги до підключення апаратного забезпечення залежать від джерел і призначення даних у системі. Автоматична класифікація фразеологічних одиниць передбачає інтеграцію даних з різних джерел, таких як текстові файли, бази даних і веб-джерела. Апаратне забезпечення повинно мати достатньо можливостей для підключення, щоб уможливити інтеграцію та передачу даних. Для цього рекомендується використовувати систему з декількома USB-портами, Ethernet і Wi-Fi. Практичні приклади вибору апаратного забезпечення для автоматичної класифікації фразеологізмів можна побачити в системах, розроблених різними компаніями та організаціями. Наприклад, хмарна платформа Google надає ряд апаратних опцій для NLP-додатків, таких як Compute Engine, що пропонує потужні процесори, великий обсяг пам'яті та SSD-накопичувачі. Іншим прикладом є платформа Amazon Web Services (AWS), яка пропонує низку апаратних опцій для додатків NLP, таких як екземпляри Elastic Compute Cloud (EC2), що надають настроювані обчислювальні ресурси, зокрема потужні процесори, велику пам'ять і можливості зберігання даних. На додаток до хмарних апаратних засобів, існують також локальні апаратні засоби, які можна використовувати для автоматичної класифікації фразеологічних одиниць [57].

Наприклад, для цього можна використовувати високопродуктивну робочу станцію з потужним процесором, великою пам'яттю та SSD-накопичувачами. Для прискорення обчислень НЛП можна також використати спеціально побудовану систему з потужним графічним процесором (GPU). З точки зору програмування, апаратне забезпечення, що використовується для автоматичної класифікації фразеологічних одиниць, повинно відповідати вимогам програмного забезпечення, яке використовується для NLP. Програмне забезпечення, що використовується для NLP, зазвичай включає складні алгоритми і структури даних, які вимагають значного обсягу пам’яті і обчислювальної потужності. Крім того, розмір набору даних, що обробляється, також може впливати на вимоги до апаратного забезпечення. Для ефективної обробки великих наборів даних може знадобитися більше обчислювальної потужності та пам'яті. Дослідивши наявне апаратне забезпечення та технічну базу, ми визначили, що саме буде найбільш ефективним для реалізації, саме, методу поєднуючого метод на основі правил та метод на основі машинного навчання та штучного інтелекту. Для реалізації методу автоматичної класифікації фразеологічних одиниць в англомовних текстах, якщо поєднати метод на основі правил і метод на основі машинного навчання, можуть знадобитися такі апаратно-технічні засоби: такі мови програмування, як Python або Java, можна використовувати для розробки програмного забезпечення для впровадження методу на основі правил і методу на основі машинного навчання. Бібліотеки, такі як NLTK, spaCy і scikit-learn, можна використовувати для реалізації різних методів обробки природної мови та машинного навчання. Інструменти попередньої обробки даних. Такі інструменти, як OpenRefine, Excel або Google Sheets, можна використовувати для попередньої обробки даних шляхом видалення стоп-слів, основ і нормалізації тексту. Метод, заснований на правилах: систему, засновану на правилах, можна розробити за допомогою регулярних виразів або інших методів зіставлення шаблонів для ідентифікації та класифікації фразеологічних одиниць. Такі інструменти, як Тестер регулярних виразів або RegExr, можна використовувати для перевірки та вдосконалення регулярних виразів [58]. Метод

на основі машинного навчання: систему на основі машинного навчання можна розробити за допомогою таких алгоритмів, як дерева рішень, випадкові ліси, опорні векторні машини або нейронні мережі. Показники оцінювання: такі показники оцінювання, як точність, запам’ятовування та оцінка F1, можна використовувати для вимірювання продуктивності системи. Для оцінки моделі машинного навчання можна використовувати такі інструменти, як Weka або scikit-learn. Візуалізація вихідних даних: інструменти візуалізації вихідних даних, такі як Matplotlib, Plotly або Tableau, можна використовувати для візуалізації та інтерпретації результатів класифікації. Центральний процесор: для швидкої обробки даних і навчання моделей машинного навчання рекомендується використовувати багатоядерний ЦП із тактовою частотою не менше 3 ГГц. Прикладами високопродуктивних ЦП є процесори Intel Core i7 або i9, AMD Ryzen 7 або 9 або Xeon. Пам'ять (RAM): принаймні 16 ГБ оперативної пам'яті рекомендовано для обробки великих наборів даних і запуску ресурсомістких алгоритмів машинного навчання. Для дуже великих наборів даних або складних моделей може знадобитися більше пам’яті.

1) Зберігання: твердотільний накопичувач (SSD) ємністю принаймні 256 ГБ рекомендується для більш швидкого доступу та зберігання даних. Для зберігання навчальних даних і файлів моделі може знадобитися додаткове сховище.

2) GPU (додатково): спеціальний графічний процесор (GPU) може пришвидшити навчання моделей машинного навчання, особливо для алгоритмів глибокого навчання.

3) Графічні процесори Nvidia або AMD зазвичай використовуються для завдань машинного навчання.

4) Підключення до мережі: для завантаження бібліотек, наборів даних та інших ресурсів, необхідних для проекту, потрібне стабільне та швидке підключення до Інтернету.

5) Операційна система: вибір операційної системи може залежати від особистих уподобань і сумісності з мовою програмування та використовуваними

бібліотеками. Поширені операційні системи для завдань машинного навчання включають Windows, macOS і Linux.

У таблиці 2.2 були висвітленні складові ПАЗ та їх основні характеристики.

Таблиця 2.2 – складові ПЗ та їх характеристика

Складові програмного алгоритмічного забезпечення	Харектеристики ПАЗ
Центральний процесор	багатоядерний ЦП із тактовою частотою не менше 3 ГГц, Intel Core i7 або i9, AMD Ryzen 7 або 9 або Xeon.
Пам'ять (RAM)	принаймні 16 ГБ
Зберігання	твердотільний накопичувач (SSD) ємністю принаймні 256 ГБ
Складові програмного алгоритмічного забезпечення	Харектеристики ПАЗ
GPU (додатково)	графічні процесори Nvidia або AMD
Мови програмування	Python або Java, Бібліотеки, такі як NLTK, spaCy і scikit-learn
Інструменти попередньої обробки даних	OpenRefine, Excel або Google Sheets
Показники оцінювання	Weka або scikit-learn
Візуалізація вихідних даних	Matplotlib, Plotly або Tableau

2.4 Висновки

У ході написання даного розділу дипломного проекту були описані методи автоматичної класифікації фразеологізмів англомовних текстів на основі інтеграції з алгоритмами машинного навчання та виявивши значущі недоліки, був запропонований авторський метод, який полягає в об'єднанні методу на основі правил та методу на основі машинного навчання, який допоможе замінити недоліки перевагами один одного. Також була розроблена структура інформаційної системи автоматичної класифікації фразеологізмів, яка включає в себе збір даних, попередню обробку, переваги та недоліки методу на основі правил, переваги та недоліки методу на основі машинного навчання, гібридний метод та оцінку. Таким чином, інформаційну систему, яка поєднує метод автоматичної класифікації на основі правил і метод машинного навчання на основі машинного навчання, можна структурувати таким чином, щоб використовувати переваги кожного підходу, тим самим підвищуючи точність класифікації. Після проведеного дослідження, зрозуміло, що автоматична класифікація фразеологічних одиниць у тексті англійською мовою є складним завданням, яке потребує інтеграції алгоритмів ШІ (штучного інтелекту) та методу на основі правил з лінгвістичними даними. Використовуючи алгоритми машинного навчання для виділення ознак і класифікації виразів, можна підвищити точність і ефективність цього процесу. Провівши дослідження, ми визначили основне апаратне забезпечення що використовується для автоматичної класифікації фразеологічних одиниць. Головними вимогами, які потрібні від даного апаратного забезпечення є те, що воно має бути здатним впоратися з обчислювальними вимогами алгоритмів, задіяних у NLP, таких як векторизація, кластеризація та класифікація. Таким чином, апаратні вимоги для реалізації методу автоматичної класифікації фразеологічних одиниць в англомовних текстах, якщо ми поєднуємо метод, заснований на правилах, і метод, заснований на машинному навчанні, можуть включати багатоядерний процесор, щонайменше 16 ГБ. Оперативної пам’яті,

твердотільний накопичувач, виділений графічний процесор (необов’язково), стабільне підключення до Інтернету та операційна система, сумісна з мовою програмування та використовуваними бібліотеками.

3. АЛГОРИТМИ ТА ТЕХНОЛОГІЯ ОБРОБКИ ІНФОРМАЦІЙНИХ ПОТОКІВ У СИСТЕМІ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ

3.1 Алгоритми автоматичної класифікації фразеологізмів

У цьому параграфі ми розглянемо деякі з найпоширеніших алгоритмів автоматичної класифікації фразеологізмів та обговоримо їхні сильні та слабкі сторони. Існує кілька алгоритмів, запропонованих для автоматичної класифікації фразеологізмів. Деякі з найпоширеніших алгоритмів включають підходи, засновані на правилах, статистиці та машинному навчанні. Алгоритм на основі правил може класифікувати вираз як ідіому, якщо він містить не буквальне значення, яке не можна вивести з окремих слів, що входять до складу виразу. Аналогічно, алгоритм на основі правил може класифікувати фразу як словосполучення, якщо вона містить високий ступінь повторюваності слів, що входять до складу виразу. Однією з головних переваг алгоритмів, заснованих на правилах, є їхня прозорість і можливість інтерпретації. Правила, що використовуються в цих алгоритмах, можуть бути легко змінені або розширені для врахування нових типів фразеологічних одиниць. Однак алгоритми, засновані на правилах, часто обмежені складністю правил і варіативністю фразеологічних одиниць. Статистичні алгоритми використовують набір статистичних показників для ідентифікації та класифікації фразеологізмів. Ці показники часто ґрунтуються на частоті та розподілі виразів у корпусі. Наприклад, статистичний алгоритм може ідентифікувати словосполучення, аналізуючи частоту вживання слів у корпусі. Статистичні алгоритми особливо корисні при роботі з великими масивами даних і коли властивості фразеологічних одиниць недостатньо чітко визначені. Однак ці алгоритми часто вимагають значної попередньої обробки та інженерії ознак, а їхня продуктивність може бути обмежена якістю та репрезентативністю корпусу, що використовується для навчання [59]. Алгоритми машинного навчання стають дедалі популярнішими для автоматичної класифікації фразеологізмів. Ці

алгоритми використовують набір ознак, витягнутих з виразів, і маркований набір даних для навчання класифікатора, який може автоматично ідентифікувати і класифікувати нові вирази. Ознаки, що використовуються в цих алгоритмах, можуть базуватися на лінгвістичних або статистичних властивостях виразів. Однією з головних переваг алгоритмів машинного навчання є те, що вони можуть вивчати складні закономірності та взаємозв'язки між ознаками і класифікаційними мітками. Це робить їх особливо корисними при роботі з великими і складними наборами даних. Однак ці алгоритми вимагають значної попередньої обробки та інженерії ознак, а їхня продуктивність може бути обмежена якістю та репрезентативністю набору мічених даних, що використовується для навчання. На рисунку 3.1 зображена UML-діаграма, яка показує високорівневе представлення компонентів і взаємозв'язків підходу на основі машинного навчання для автоматичної класифікації фразеологічних одиниць:

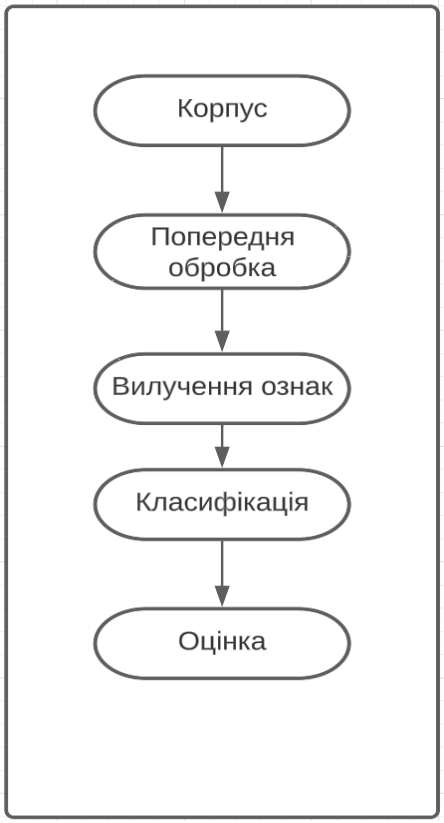


Рисунок 3.1 – UML діаграма автоматичної класифікації на основі машинного навчання

Існують різні алгоритми, які можна використовувати для автоматичної класифікації фразеологічних одиниць, і вибір алгоритму залежить від конкретного завдання та наявних даних. У таблиці 3.1 зображено чотири найпоширеніших підходи:

Таблиця 3.1 – алгоритми автоматичної класифікації ФО

Алгоритм	Суть алгоритму	Метод до якого відноситься
SVM	популярний алгоритм машинного навчання для завдань класифікації, оскільки він може ефективно класифікувати дані, знаходячи найкращу гіперплощину, яка розділяє різні класи у просторі великої розмірності.	Метод на основі машинного навчання
Наївний Байєс	У контексті фразеологічних одиниць алгоритм можна навчити на наборі мічених даних, де кожна фразеологічна одиниця позначена відповідно до її типу (наприклад, ідіома, прислів’я, словосполучення тощо).	Метод на основі машинного навчання
Алгоритм дерева рішень	Алгоритм працює шляхом поділу даних на підмножини на основі значення певних атрибутів, а потім рекурсивного розподілу даних на підмножини, доки не буде прийнято рішення.	Метод на основі правил
Алгоритм вбудовування слів	Алгоритм працює шляхом аналізу спільного використання слів у корпусі та використання цієї інформації для створення матриці, яка представляє зв’язки між словами.	Метод на основі правил

Алгоритм машини на основі опорних векторів (SVM) зображений на рисунку 3.2:

Рисунок 3.2 – UML діаграма алгоритму автоматичної класифікації ФО SVM

На цій діаграмі процес починається з вхідної фрази, яка проходить через виділення ознак, щоб перетворити її на відповідне числове представлення. Етап попередньої обробки включає будь-яке необхідне очищення або нормалізацію даних. Потім алгоритм SVM навчається з використанням позначених даних, щоб дізнатися межі класифікації. Після навчання SVM можна використовувати для класифікації нових, невідомих фразеологічних одиниць. Нарешті, вихідна категорія представляє прогнозовану класифікацію вхідної фрази. SVM – це популярний алгоритм керованого навчання, який можна використовувати для задач класифікації. На основі набору маркованих навчальних прикладів SVM знаходить границю рішення, яка максимізує різницю між класами. Границя рішення представляється гіперплощиною у високорозмірному просторі ознак. Щоб використовувати SVM для класифікації фразеологізмів, потрібно спочатку виділити ознаки з даних. Одним з популярних підходів є представлення кожної фразеологічної одиниці у вигляді мішка слів, де кожне слово представлене як ознака. Потім ми можемо застосувати алгоритм SVM до векторів ознак, щоб вивчити класифікатор, який може передбачити клас нових фразеологічних одиниць. Алгоритм SVM намагається знайти гіперплощину, яка максимізує різницю між двома класами, де різниця визначається як відстань між

гіперплощиною та найближчими навчальними прикладами з кожного класу. Оптимізаційну задачу для SVM можна сформулювати наступним чином:

звести до мінімуму:

$$min\frac{1}{2} * \|w\|^2 + C * sum\_i(yi * (w.xi + b) - 1)^+ \quad , \qquad (3.1)$$

де ||w|| це – евклідова норма вагового вектора w;

C це – гіперпараметр, який контролює компроміс між максимізацією маржі та мінімізацією помилки класифікації;

sum_i це – сума за всіма точками даних

yi – мітка i-ої точки даних (-1 або 1)

xi – i-та точка даних

b – член зсуву

^+ - функція втрат петлі, визначена як max(0, x)

Мета алгоритму SVM полягає в тому, щоб знайти ваговий вектор w і член зміщення b, які мінімізують цільову функцію, за умови, що yi * (w . xi + b) >= 1 для всіх точок даних xi. Ці обмеження гарантують, що всі точки даних будуть правильно класифіковані або лежатимуть за межами поля.

Алгоритм SVM намагається максимізувати різницю між двома класами. Межа – це відстань між гіперплощиною та найближчими точками даних кожного класу. Найближчі точки даних називаються опорними векторами.

Надана формула є цільовою функцією алгоритму SVM. Вона складається з двох частин:

$$\frac{1}{2} * \|w\|^2, \qquad (3.2)$$

де ||w|| представляє евклідову норму вагового вектора w. Мета цього члена – мінімізувати норму вагового вектора, що допомагає уникнути перебору і зробити границю рішення якомога більш загальною.

$$C * sum\_i(yi * (w.xi + b) - 1)^{+}, \quad (3.3)$$

де C це – гіперпараметр, який контролює компроміс між максимізацією маржі та мінімізацією помилки класифікації;

yi – мітка i-ї точки даних (-1 або 1);

xi – i-та точка даних;

b – член зсуву;

^+ позначає функцію втрат на шарнірі, яка визначається як max(0, x) [35]. Функція втрат на петлях карає алгоритм SVM за неправильну класифікацію точок даних, які знаходяться всередині поля або на неправильній стороні гіперплощини. Сума функцій втрат шарнірів для всіх точок даних представляє помилку класифікації алгоритму SVM.

Для мінімізації цільової функції алгоритм SVM використовує алгоритм оптимізації, такий як градієнтний спуск або квадратичне програмування. Алгоритм оптимізації ітеративно оновлює ваговий вектор w і член зсуву b до тих пір, поки цільова функція не сходиться до мінімуму. Результуючий ваговий вектор w представляє параметри границі рішення, а член зміщення b – зміщення границі рішення від початку координат. Наївний Байєс – це імовірнісний алгоритм, який зазвичай використовується для класифікації текстів. Він припускає, що присутність кожної ознаки (слова) в документі не залежить від присутності інших ознак, враховуючи мітку класу. Це припущення відоме як «наївне» припущення і може не виконуватися на практиці, але воно часто добре працює на практиці і має низьку обчислювальну складність. Щоб використовувати наївний Байєс для класифікації фразеологізмів, нам потрібно спочатку виділити ознаки з даних. Одним з популярних підходів є представлення кожної фразеологічної одиниці у вигляді

мішка слів, де кожне слово представлене як ознака. Потім ми можемо застосувати наївний алгоритм Байєса до векторів ознак, щоб отримати класифікатор, який може передбачити клас нових фразеологічних одиниць [60]. Наївний алгоритм Байєса зображений на рисунку 2.3 та обчислює апостеріорну ймовірність кожного класу, враховуючи вектор ознак, використовуючи правило Байєса:

$$P(c|x) = P(x|c) * P(c)/P(x), \tag{3.4}$$

де c – мітка класу, x – вектор ознак;

P(c|x) – апостеріорна ймовірність c для x;

P(x|c) – ймовірність x для c;

P© – попередня ймовірність c;

P(x) – ймовірність доказів.

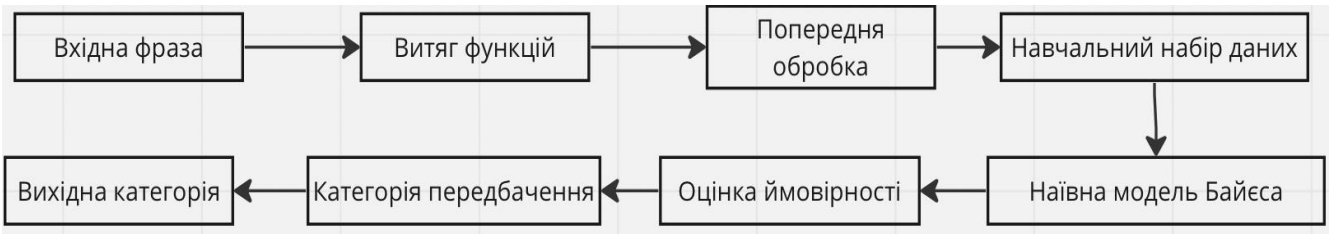


Рисунок 3.3 – UML діаграма алгоритму автоматичної класифікації ФО наївного Баєса

Ймовірність зазвичай моделюється за допомогою мультиноміального розподілу, де кожна ознака розглядається як підрахунок кількості разів, коли вона з'являється в документі. Попередню ймовірність можна оцінити з навчальних даних, а ймовірність доказів можна ігнорувати, оскільки вона однакова для всіх класів. Після навчання наївного байєсівського класифікатора ми можемо використовувати його для прогнозування класу нових фразеологізмів, обчислюючи апостеріорну ймовірність для кожного класу і вибираючи клас з найбільшою ймовірністю. Порівнявши загалом основні алгоритми, які використовуються та рухаючись шляхом «гібридного» методу, який поєднує в собі метод на основі

правил та метод на основі машинного навчання, розроблений також «гібридний» алгоритм, який вийшов після поєднання алгоритму Наївного Байєса та алгоритму розпізнавання іменованих об’єктів. Алгоритм розпізнавання іменованих об’єктів (NER) – це заснований на правилах метод, який використовується для ідентифікації та класифікації іменованих об’єктів у тексті, таких як люди, організації та місця розташування. Він використовує набір правил для ідентифікації та вилучення цих сутностей із тексту.

З іншого боку, Наївний алгоритм Байєса — це метод машинного навчання, який використовує імовірнісні моделі для класифікації тексту за попередньо визначеними категоріями. Він обчислює ймовірність кожної категорії на основі частоти слів у тексті та призначає тексту категорію з найвищою ймовірністю. Поєднання цих двох алгоритмів може передбачати використання вихідних даних алгоритму NER як функцій у алгоритмі Наївного Байєса. Наприклад, іменовані сутності, ідентифіковані алгоритмом NER, можна використовувати як вхідні дані для алгоритму Наївного Байєса як додаткові ознаки для класифікації. Алгоритм NER також можна використовувати для отримання контекстної інформації, яка може бути використана для підвищення точності алгоритму Наївного Байєса. Наприклад, контекст названих сутностей, такий як їх близькість до інших слів у тексті, можна використовувати для надання додаткової інформації алгоритму Наївного Байєса. Гібридний алгоритм, який поєднує алгоритм Naive Bayes і алгоритм Named Entity Recognition (NER) для автоматичної класифікації фразеологічних одиниць, є технологією, яка використовує як методи на основі машинного навчання, так і на основі правил. Поєднуючи ці два алгоритми, гібридний алгоритм може ефективно класифікувати фразеологічні одиниці, спочатку використовуючи алгоритм NER для ідентифікації та позначення іменованих сутностей у вхідному тексті. Вихідні дані алгоритму NER потім використовуються як вхідні дані для алгоритму Наївного Байєса, який обчислює ймовірність кожного класу фразеологічної одиниці з урахуванням ідентифікованих

іменованих сутностей. Тоді класифікація вихідних даних базується на класі з найвищою ймовірністю.

3.2 Технологія обробки інформаційних потоків у системі автоматичної класифікації фразеологізмів

У системі автоматичної класифікації фразеологічних одиниць важливу роль відіграє технологія обробки інформаційних потоків. Ця технологія включає низку алгоритмів і методів, які дозволяють системі обробляти й аналізувати великі обсяги даних, щоб ідентифікувати шаблони та зв'язки, які можна використовувати для класифікації фразеологічних одиниць. Важливим аспектом технології обробки інформаційних потоків у контексті автоматичної класифікації фразеологізмів є оцінювання. Оцінка передбачає вимірювання точності системи класифікації, щоб визначити, наскільки добре вона працює. Є кілька різних показників, які можна використовувати для оцінки ефективності системи класифікації, включаючи точність, запам'ятовування та оцінку F1. Точність вимірює частку справжніх позитивних результатів серед усіх позитивних прогнозів, тоді як пригадування вимірює частку справжніх позитивних результатів серед усіх фактичних позитивних випадків. F1-оцінка – це показник, який поєднує точність і запам'ятовування в один показник.

Підсумовуючи, технологія обробки інформаційних потоків відіграє вирішальну роль у системі автоматичної класифікації фразеологічних одиниць. Ця технологія передбачає використання алгоритмів і методів, таких як обробка природної мови та машинне навчання, для обробки й аналізу даних з метою визначення шаблонів і зв'язків, які можна використовувати для класифікації. Використовуючи ці технології, можна розробити високоточні й ефективні системи автоматичної класифікації фразеологічних одиниць. Технологія обробки інформаційних потоків є неодмінною складовою системи автоматичної класифікації фразеологічних одиниць. Основною метою цієї технології є ефективна

обробка інформації для підвищення точності результатів класифікації. Однією з фундаментальних концепцій обробки інформаційних потоків є використання математичних моделей для представлення даних. Ці моделі дозволяють маніпулювати та перетворювати інформацію, що є важливим у процесі класифікації. У випадку з фразеологізмами найпоширенішою математичною моделлю є модель векторного простору. Ця модель представляє документи та фрази як вектори у багатовимірному просторі. Кожен вимір представляє певний термін, і значення кожного виміру відповідає частоті цього терміна в документі чи фразі. Використовуючи модель векторного простору, ми можемо виконувати різноманітні операції з даними. Наприклад, ми можемо виміряти подібність між двома векторами за допомогою косинусної міри подібності:

$$cos\ \_sim(x, y) = (x.y)/(\|x\| * \|y\|) \quad , \tag{3.5}$$

де x і y це- два вектори;

«.» це- скалярний добуток;

«|| ||» це – величина вектора. Міра косинусної подібності повертає значення від -1 до 1, причому значення, ближчі до 1, вказують на вищу подібність.

Іншою важливою операцією в обробці інформаційних потоків є виділення ознак. Вилучення ознак – це процес вибору та перетворення релевантної інформації з вхідних даних для підвищення точності класифікації. У випадку фразеологічних одиниць релевантні ознаки можуть включати частоту слів, теги частини мови та синтаксичні зв’язки. Одним із методів виділення ознак є модель «мішок слів». Ця модель представляє документ як набір слів, ігноруючи порядок слів і синтаксис. Кожне слово в документі розглядається як функція, а частота кожної функції використовується для створення вектора ознак для документа. Формально нехай D — набір документів, а W — набір усіх унікальних слів у документах. Тоді модель сумки слів представляє кожен документ d як вектор:

$$d = (tf(w1, d), tf(w2, d), \dots, tf(wn, d)), \quad (3.6)$$

де tf(w, d) — частота терміну слова w у документі d;

n — розмір словника W [61].

Окрім виділення ознак, обробка потоку інформації також включає алгоритми класифікації. Алгоритми класифікації використовуються для прогнозування мітки класу даного вхідного даних на основі ознак, отриманих із цього вхідного даних.

Наш гібридний алгоритм може бути реалізований різними мовами програмування, такими як Python, Java або C++. Алгоритм NER можна реалізувати за допомогою бібліотек обробки природної мови, таких як spaCy або NLTK, тоді як алгоритм Naive Bayes можна реалізувати за допомогою бібліотек машинного навчання, таких як scikit-learn або TensorFlow. Вихідні дані алгоритму NER можуть бути попередньо оброблені та використані як вхідні дані для алгоритму Naive Bayes, який потім виводить остаточну класифікацію. По-перше, нам потрібно підготувати дані для навчання та тестування моделі. Ми можемо використовувати такі мови програмування, як Python, Java або R, для попередньої обробки даних і вилучення відповідних функцій із тексту. Наприклад, ми можемо використовувати бібліотеку NLTK у Python, щоб токенізувати текст і витягти з тексту теги частин мови (POS) і іменовані сутності. Після попередньої обробки даних ми можемо використовувати бібліотеку scikit-learn у Python для реалізації алгоритму Naive Bayes для класифікації. Ми можемо використовувати класифікатор MultinomialNB з бібліотеки, який підходить для дискретних даних, таких як частота слів і тегів POS. Вхідними даними для класифікатора були б ознаки, витягнуті з тексту, такі як частота слів і теги POS, а виходом був би прогнозований клас фразеологічної одиниці. Щоб включити алгоритм NER у гібридний алгоритм, ми можемо використовувати такі інструменти, як Stanford Named Entity Recognizer або бібліотека spaCy на Python. Ці інструменти можуть розпізнавати іменовані сутності в тексті та класифікувати їх за попередньо визначеними категоріями, такими як

особа, організація або місцезнаходження. Ми можемо використовувати вихідні дані алгоритму NER як додаткові функції для класифікатора Naive Bayes. Наприклад, ми можемо додати функцію, яка вказує, чи містить фразеологічна одиниця названу сутність певного типу. Щоб навчити та оцінити гібридний алгоритм, ми можемо використовувати фреймворки машинного навчання, такі як TensorFlow, PyTorch або Keras. Ми можемо розділити дані на набори для навчання та тестування та використовувати набір для навчання для навчання простого класифікатора Байєса та алгоритму NER. Потім ми можемо використовувати тестовий набір для оцінки продуктивності гібридного алгоритму, використовуючи такі показники, як точність, запам'ятовування та оцінка F1. Загалом технологія, яка використовується для створення та використання гібридного алгоритму, включає такі мови програмування, як Python, бібліотеки, такі як NLTK, scikit-learn і spaCy, а також інфраструктури машинного навчання, такі як TensorFlow і PyTorch. Гібридний алгоритм поєднує алгоритм Naive Bayes і алгоритм NER, які реалізовані за допомогою цих інструментів, для класифікації фразеологічних одиниць за допомогою комбінації методу на основі правил і методу на основі машинного навчання.

3.3 Проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів

Починаючи дослідження у цьому параграфі, варто почати з основних вказівок щодо компонентів та архітектури інформаційної системи автоматичної класифікації фразеологічних одиниць. Компонент попередньої обробки відповідає за перетворення необробленого тексту в структурований формат, який може використовуватися компонентом вилучення функцій. Він включає такі завдання, як токенізація, сегментація речень і позначення частин мови. Компонент виділення ознак відповідає за ідентифікацію відповідних ознак у попередньо обробленому тексті, які можна використовувати для розрізнення різних фразеологічних одиниць.

Це може включати такі функції, як n-грами, вбудовування слів і синтаксичні шаблони. Компонент класифікації відповідає за використання виділених ознак для класифікації фразеологічних одиниць у заздалегідь визначені категорії. Одним із поширених підходів є використання опорних векторних машин (SVM) із функцією ядра на основі вилучених функцій. Компонент оцінки відповідає за оцінку ефективності системи класифікації. Це може включати такі показники, як точність, запам’ятовування та оцінка F1. Загальна архітектура інформаційної системи може відповідати моделі клієнт-сервер, при цьому клієнтський компонент є інтерфейсом користувача для взаємодії з системою, а серверний компонент містить компоненти попередньої обробки, вилучення ознак, класифікації та оцінки. Ось опис високорівневого дизайну програмного забезпечення: компонент інтерфейсу користувача забезпечить користувачеві засоби для введення тексту для класифікації та відображення результатів процесу класифікації. Модуль попередньої обробки оброблятиме початкові кроки обробки тексту, такі як токенізація, сегментація речень і тегування частини мови. Модуль виділення функцій візьме попередньо оброблений текст і витягне відповідні функції, такі як n-грами, вбудовування слів і синтаксичні шаблони. Модуль класифікації бере витягнуті ознаки та класифікує фразеологічні одиниці за попередньо визначеними категоріями за допомогою SVM із функцією ядра на основі витягнутих ознак. Модуль оцінювання оцінював би продуктивність системи класифікації за допомогою таких показників, як точність, запам’ятовування та оцінка F1. Для зберігання навчальних даних і результатів класифікації можна використовувати базу даних. Зовнішні бібліотеки, такі як scikit-learn, можна використовувати для реалізації алгоритму класифікації на основі SVM та інших методів машинного навчання. Підсумовуючи, проектування інформаційної системи автоматичної класифікації фразеологічних одиниць включає кілька компонентів і вимагає ретельного розгляду архітектури системи та вибору алгоритму. Використовуючи клієнт-серверну модель, модулі попередньої обробки, виділення ознак, класифікації та оцінки, а також зовнішні бібліотеки, можна побудувати надійну та

точну систему класифікації фразеологічних одиниць. Важливі приклади проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів, які являються фундаментальними у цій області є: дизайн інтерфейсу користувача, дизайн бази даних, дизайн попередньої обробки, дизайн класифікації та дизайн оцінювання. Інтерфейс користувача системи повинен бути інтуїтивно зрозумілим і зручним, щоб користувачі могли легко орієнтуватися в системі та використовувати її. Дизайн інтерфейсу користувача повинен включати форми введення текстових даних і виведення результатів класифікації. Форма введення повинна дозволяти користувачеві вводити текстові дані або завантажувати текстові дані у файл. Вихідні дані мають відображати результати класифікації з поясненням того, як система зробила класифікацію. Система повинна мати базу даних для зберігання фразеологічних одиниць, ознак і результатів класифікації. Дизайн бази даних повинен містити таблиці фразеологічних одиниць, ознаки та результати. Кожна таблиця повинна мати необхідні поля для зберігання необхідних даних. Проект попередньої обробки повинен включати очищення даних, нормалізацію даних і виділення ознак. Очищення даних передбачає видалення нерелевантних даних, таких як стоп-слова, і виправлення орфографічних помилок. Нормалізація даних передбачає перетворення даних у стандартну форму. Виділення ознак передбачає ідентифікацію важливих ознак у текстових даних, які використовуватимуться в класифікації. Проект класифікації повинен включати алгоритм, який використовується для класифікації, вибір ознак, які використовуються в класифікації, і навчання системи. Алгоритмом, що використовується для класифікації, може бути машина опорних векторів або нейронна мережа. Вибір ознак може включати частоту слів, словосполучення або n-грами. Навчання системи передбачає використання навчального набору класифікованих фразеологічних одиниць, щоб навчити систему класифікувати нові фразеологічні одиниці. План оцінки повинен включати вибір метрик оцінювання та тестування системи. Показники оцінювання можуть включати точність, запам’ятовування та

оцінку F1. Тестування системи передбачає використання тестового набору фразеологічних одиниць для оцінки ефективності системи. Загалом, кожен елемент дизайну має бути ретельно продуманий і реалізований, щоб система була точною та ефективною у своїй класифікації фразеологічних одиниць. Дизайн краще пояснити більш детально, використовуючи Python як приклад мови програмування.

Спочатку розглянемо модуль попередньої обробки. У цьому модулі нам потрібно виконати такі завдання, як токенізація, виведення коренів і видалення стоп-слова. Для реалізації цих завдань ми можемо використовувати бібліотеку NLTK на Python. На рисунку 3.4 зображено приклад того, як ми можемо токенізувати речення за допомогою бібліотеки NLTK.

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize

sentence = "The quick brown fox jumps over the lazy dog."
tokens = word_tokenize(sentence)
print(tokens)
```

Рисунок 3.4 - Токенізація речення

У наведеному вище коді ми спочатку імпортуємо бібліотеку nltk і завантажуємо punkt tokenizer. Потім ми імпортуємо функцію word_tokenize з модуля nltk.tokenize. Ми визначаємо речення, а потім лексемуємо його за допомогою функції word_tokenize. Нарешті, ми друкуємо жетони. Далі розглянемо модуль вилучення функцій. Тут нам потрібно виділити відповідні ознаки з фразеологізмів. Ми можемо використати підхід сумки слів, щоб представити кожну фразеологічну одиницю як вектор ознак. На рисунку 3.5 зображено приклад того, як ми можемо реалізувати підхід сумки слів у Python.

```
from sklearn.feature_extraction.text import CountVectorizer

corpus = [
    "The quick brown fox jumps over the lazy dog.",
    "She sells seashells by the seashore."
]

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
print(X.toarray())
```

Рисунок 3.5 - Підхід сумми слів

Ми спочатку імпортуємо клас CountVectorizer з модуля sklearn.feature_extraction.text. Визначаємо корпус із двох речень. Потім ми створюємо екземпляр класу CountVectorizer і вписуємо його в корпус за допомогою методу fit_transform. Нарешті, ми друкуємо представлення корпусу у вигляді сумки слів. Переходячи до модуля класифікації, ми можемо використовувати машину опорних векторів (SVM) для класифікації фразеологізмів у різні категорії. На рисунку 3.6 зображено приклад того, як ми можемо реалізувати SVM у Python за допомогою бібліотеки sklearn.

```
from sklearn import svm

X = [[0, 0], [1, 1]]
y = [0, 1]
clf = svm.SVC()
clf.fit(X, y)

print(clf.predict([[2., 2.]]))
```

Рисунок 3.6 - Реалізація методу SVM

Ми визначаємо набір даних X і відповідні мітки y. Ми створюємо екземпляр класу SVC і підбираємо його до набору даних за допомогою методу fit. Нарешті, ми прогнозуємо мітку нової точки даних за допомогою методу predict. Нарешті, давайте розглянемо модуль оцінювання. Модуль оцінки в Python надає різні показники для вимірювання ефективності алгоритму класифікації. Найбільш часто використовувані показники включають точність, запам'ятовування та оцінку F1 [62]. Бібліотека Scikit-learn у Python надає повний набір метрик оцінки для алгоритмів класифікації. Щоб використовувати модуль оцінки в Python, спочатку нам потрібно розділити набір даних на набори для навчання та перевірки. Навчальний набір використовується для навчання алгоритму класифікації, а набір перевірки використовується для оцінки його продуктивності. У додатку Є зображений приклад фрагмента коду, який демонструє, як використовувати модуль оцінки в Python. Вихід цього фрагмента коду відображатиме показники ефективності класифікатора дерева рішень у наборі перевірки. Модуль оптимізації в Python надає методи для точного налаштування алгоритму класифікації та підвищення його точності. Найбільш часто використовувані методи оптимізації включають пошук по сітці, випадковий пошук і байєсовську оптимізацію. Бібліотека Scikit-learn у Python надає повний набір методів оптимізації для алгоритмів класифікації [63]. У додатку Ж зображено приклад фрагмента коду, який демонструє, як використовувати модуль оптимізації в Python. На виході цього фрагмента коду відображатимуться показники ефективності класифікатора дерева рішень із найкращими параметрами, отриманими в результаті оптимізації пошуку в сітці в наборі перевірки. У додатку З можна подивитись приклад реалізації гібридного алгоритму з використанням алгоритму NER і алгоритму Naive Bayes для автоматичної класифікації фразеологічних одиниць. У цій реалізації ми спочатку завантажуємо англомовну модель для розпізнавання іменованих сутностей за допомогою бібліотеки spacy. Потім ми навчаємо класифікатор Naive Bayes на наборі даних позначених фразеологічних одиниць, використовуючи модель MultinomialNB з бібліотеки sklearn. Навчальні дані складаються з фразеологічних

одиниць як векторів ознак та їхніх відповідних міток (ідіом або літералів). Далі ми визначаємо функцію classify_phraseological_unit, яка приймає фразеологічну одиницю як вхідні дані, використовує розпізнавання іменованих сутностей, щоб перевірити наявність будь-яких іменованих сутностей, а якщо ні, витягує ознаки з фразеологічної одиниці та класифікує її за допомогою навченого класифікатора Naive Bayes. Нарешті, ми надаємо приклад використання функції classify_phraseological_unit і роздруковуємо передбачені класифікації для двох різних фразеологічних одиниць. У додатку И зображена діаграма UML, яка показує основні компоненти та їхні зв’язки в гібридному алгоритмі, який поєднує алгоритм Naive Bayes і алгоритм NER для автоматичної класифікації фразеологічних одиниць. Клас «The Phrase Classifier» є основним класом, який поєднує моделі Naive Bayes і NER. Він має дві приватні змінні екземпляра, naiveBayesModel і nerModel, які є екземплярами класу "Model". NaiveBayesModel містить навчений класифікатор Naive Bayes, тоді як nerModel містить навчену модель NER. Клас «Model» — це абстрактний клас, який визначає основні функції моделі машинного навчання. Він має приватну змінну екземпляра "data", яка є екземпляром класу "Data". Клас "Data" представляє дані, які використовуються для навчання моделі, і має приватну змінну екземпляра "path", яка є шляхом до файлу даних. Клас «Data» також має метод «load» для завантаження даних із файлу. Клас «Named Entity Tag» — це простий клас даних, який представляє іменований тег сутності [64]. Він має дві приватні змінні екземпляра, «value» і «type». «Value» — це рядкове значення іменованої сутності, а «type» — це екземпляр переліку «EntityType», який представляє тип іменованої сутності. Перелік "EntityType" - це простий перелік, який містить перелік різних типів іменованих сутностей, які можна класифікувати за моделлю NER [65]. Загалом діаграма UML показує, як різні класи та їхні зв’язки працюють разом для реалізації гібридного алгоритму, який поєднує моделі Наївного Байєса та NER для автоматичної класифікації фразеологічних одиниць.

Таким чином, поєднання методу машинного навчання з методом на основі правил може допомогти підвищити ефективність автоматичної класифікації фразеологічних одиниць.

У додатку Е наведено приклад класифікації фразеологічних одиниць із словника [19] за допомогою авторського ПЗ (Hybrid Soft). Для класифікації було обрано ті ж ФО, що аналізувалися у розділі 1 (100 ФО). Результати класифікації виявились такими. У результаті роботи авторського ПЗ (Hybrid Soft) 27 ФО було віднесено до групи 1, 9 – до групи 2, 6 до групи 3, 19 до групи 4, 16 до групи 5 та 14, які не увійшли до конкретної групи. Більше того, завдяки авторському методу, алгоритм зміг виділити 2 додаткових ознаки та класифікував ще 9 ФО в групу ФО пов'язаних з тілом та 10 ФО пов'язаних з тваринами. Час автоматичної класифікації склав 19 секунд. Таким чином, авторське ПЗ дозволяє не лише зменшити тривалість часу на обробку ФО, а й встановити додаткові закономірності для класифікаційних ознак, за якими класифікуються ФО.

3.4 Висновки

У ході написання даного розділу дипломного проекту були визначені та описані алгоритми автоматичної класифікації фразеологізмів англомовних текстів, які реалізують авторський "гібридний" метод автоматичної класифікації фразеологічних одиниць в англомовних текстах. Основними алгоритмами були визначені такі: SVM, Наївний Байєс, Алгоритм дерева рішень, Алгоритм вбудовування слів. Авторський алгоритм передбачає поєднання алгоритму Наївного Байеса та алгоритму NER (Named Entity Recognition). Також були описані наявні технології обробки інформаційних потоків у системі автоматичної класифікації фразеологізмів та авторська технологія. Технологія, яка використовується для створення та використання гібридного алгоритму, включає такі мови програмування, як Python, бібліотеки, такі як NLTK, scikit-learn і spaCy, а також інфраструктури машинного навчання, такі як TensorFlow і PyTorch Окрім

цього, були розглянуті основні приклади проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів. Серед основних етапів цього дослідження можна виділити такі: токенізація речення за допомогою бібліотеки NLTK, реалізація підходу сумки слів у Python, реалізація SVM у Python за допомогою бібліотеки sklearn, демонстрація того, як використовувати модуль оцінки в Python. Підсумовуючи параграф 3.3 ми провели загальне дослідження використовуючи мову Python для того щоб зобразити роботу гібридного алгоритму.

4. ПРОГРАМНО-ТЕХНІЧНА СИСТЕМА РЕАЛІЗАЦІЇ МЕТОДУ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ

4.1 Опис середовища розробки програмно-технічної системи реалізації методу автоматичної класифікації фразеологізмів

"Hybrid soft" - це потенційна програма, яку ми розробляємо для того, щоб реалізувати поєднання методів автоматичної класифікації фразеологічних одиниць в англомовних текстах. Далі будемо називати її "HS". Загалом - це бібліотека Python для обробки природної мови, яка може бути розроблена і запущена на широкому спектрі комп'ютерних систем. Програмно-технічне середовище для розробки HS включає наступне:

1) Операційна система: HS сумісна з різними операційними системами, включаючи Windows, macOS та дистрибутиви Linux, такі як Ubuntu, Fedora та CentOS.

2) Python: HS реалізовано на Python, тому потрібно встановити Python. HS підтримує версії Python 2.x та Python 3.x. Однак рекомендується використовувати Python 3.x, оскільки Python 2.x більше не підтримується.

3) Пакети Python: HS залежить від декількох пакетів Python, які можна встановити за допомогою pip, інсталятора пакетів Python. Необхідні пакунки включають numpy, scipy, matplotlib та інші. HS надає зручну команду встановлення для встановлення всіх необхідних пакунків, яку зображено на рисунку 4.1.

```
pip install nltk[all]
```

Рисунок 4.1 - Команда встановлення

Вимоги до апаратного забезпечення для розробки HS є відносно скромними і залежать від розміру набору даних та складності виконуваних операцій. Однак для оптимальної продуктивності рекомендується мати комп'ютер з наступними

характеристиками: Процесор: Багатоядерний процесор (наприклад, Intel Core i5 або вище) для більш швидкої обробки. Оперативна память: Щонайменше 4 ГБ оперативної пам'яті для обробки великих наборів даних і виконання операцій, що вимагають багато пам'яті. Сховище: Достатній обсяг пам'яті для зберігання бібліотеки HS, наборів даних і будь-яких додаткових ресурсів, необхідних для ваших завдань NLP. Підключення до мережі: Деякі функції HS, такі як завантаження додаткових наборів даних, можуть вимагати підключення до Інтернету. Інтегроване середовище розробки (IDE): HS можна розробляти за допомогою будь-якого Python-сумісного IDE або текстового редактора. Серед популярних варіантів: PyCharm, Visual Studio Code, Atom, Sublime Text та Jupyter Notebook. Виберіть IDE, з якою вам комфортно працювати і яка підтримує розробку на Python. Варто зазначити, що HS - це дуже гнучка бібліотека, яка легко налаштовується, а її функціональність можна розширити за допомогою додаткових інструментів та ресурсів. Залежно від конкретних вимог, може знадобитися встановити та налаштувати інші інструменти, такі як специфічні моделі NLP, корпуси або сторонні бібліотеки, щоб розширити можливості HS. Загалом, HS забезпечує зручне та доступне середовище для обробки природної мови, яке можна легко налаштувати на широкому спектрі комп'ютерних систем. Інструментарій програмного забезпечення HS в основному розроблений за допомогою мови програмування Python і використовує різні інструменти та технології в екосистемі Python. Ось деякі ключові інструменти та технології, що використовуються при написанні програми HS:

5. Python: HS написана мовою програмування високого рівня Python, відомою своєю простотою, читабельністю та широкою підтримкою наукових обчислень і аналізу даних. Python надає багатий набір бібліотек і фреймворків, які використовуються в HS.

6. Бібліотеки HS: HS - це бібліотека для обробки природної мови в Python. Вона надає широкий спектр модулів і пакетів, включаючи токенізатори, стеммери,

тегери частин мови, синтаксичні аналізатори та класифікатори, для виконання різних завдань NLP. Ці бібліотеки є основними компонентами HS.

7. NumPy та SciPy: NumPy (Numerical Python) та SciPy (Scientific Python) - це фундаментальні бібліотеки для чисельних та наукових обчислень на Python [66]. Вони надають ефективні структури даних, чисельні алгоритми та математичні функції, які використовуються в HS для маніпулювання даними, матричних операцій та статистичних обчислень.

8. Scikit-learn: Scikit-learn - це популярна бібліотека машинного навчання на мові Python, яка надає повний набір інструментів для різних алгоритмів керованого та некерованого навчання [67]. HS може використовувати класифікатори Scikit-learn, такі як наївний Байєс, дерева рішень та класифікатори з максимальною ентропією, для завдань навчання та класифікації.

9. Matplotlib та seaborn: Matplotlib та seaborn - це бібліотеки для побудови графіків у Python, які дозволяють візуалізувати дані та результати. Ці бібліотеки можна використовувати в HS для візуалізації показників ефективності, кривих навчання або розподілу фразеологічних одиниць під час аналізу.

10. Pandas: Pandas - це потужна бібліотека для маніпулювання та аналізу даних [68]. Вона надає структури даних, такі як DataFrames, які широко використовуються в HS для обробки структурованих даних, виконання попередньої обробки даних і полегшення завдань вилучення ознак.

11. Jupyter Notebook: Jupyter Notebook - це інтерактивне веб-середовище, яке дозволяє створювати та обмінюватися документами, що містять живий код, візуалізації та пояснювальний текст [69]. Його часто використовують разом з HS для розробки та представлення коду, результатів та аналізів у більш інтерактивний та читабельний спосіб.

У контексті застосування HS не існує конкретного головного органу або головного сервера, який приймає запити в традиційному розумінні клієнт-сервер. HS - це насамперед бібліотека та інструментарій для задач обробки природної мови, і її використання зазвичай передбачає написання коду на Python для виконання

конкретних операцій NLP. Однак HS можна інтегрувати у веб-додатки або сервіси, де задіяний сервер. У таких випадках основна частина програми буде залежати від конкретної архітектури та фреймворку, що використовується. Ось загальний огляд архітектури веб-додатків на основі HS:

1. Веб-сервер: Веб-сервер, такий як Apache або Nginx, обробляє вхідні HTTP-запити від клієнтів і перенаправляє їх на відповідний сервер додатків або фреймворк.

2. Сервер/фреймворк додатків: Сервер додатків або фреймворк, такий як Django, Flask або FastAPI, отримує HTTP-запити від веб-сервера і перенаправляє їх до відповідних кінцевих точок або представлень [70].

3. Кінцева точка/подання: Кінцева точка або подання на сервері додатків отримує запити, пов'язані з класифікацією фразеологічних одиниць. Він може містити певну URL-адресу, наприклад, "/classify", яка запускає виконання коду для класифікації фразеологізмів.

4. Інтеграція з HS: У межах кінцевої точки або подання пишеться код на мові Python з використанням бібліотеки HS для виконання необхідних завдань з обробки тексту, вилучення ознак і класифікації. Цей код використовує функціональність HS для аналізу вхідного тексту і відповідної класифікації фразеологічних одиниць.

5. 5) Реакція: Після класифікації фразеологізмів сервер додатків генерує відповідну відповідь, яка може бути у формі JSON, HTML або будь-якому іншому форматі залежно від вимог [71]. Відповідь може містити результати класифікації або будь-яку іншу релевантну інформацію. Важливо зазначити, що сама бібліотека HS немає вбудованого сервера або можливостей обробки запитів. Замість цього HS зазвичай використовується в рамках більшого додатку або системи, де логіка на стороні сервера реалізована за допомогою веб-фреймворків та інструментів. Конкретні деталі реалізації основної частини та сервера залежать від обраного фреймворку та загальної архітектури програми. Розглянемо структуру бібліотеки

HS та її компоненти: бібліотека HS організована у каталоги та файли, які слугують певним цілям.

Деякі важливі каталоги та файли у структурі проекту HS включають:

1. HS: Це основний каталог, який містить вихідний код HS та основні модулі.

2. HS/corpus: У цьому каталозі зберігаються різні корпуси (великі збірки текстів), які постачаються разом з HS. Ці корпуси слугують наборами даних для навчання і тестування моделей NLP.

3. 3)HS/data: У цьому каталозі зберігаються додаткові файли даних, що використовуються бібліотекою HS, такі як мовні моделі, лексичні ресурси та попередньо навчені моделі.

4. HS/tokenize: Цей каталог містить модулі, пов'язані з токенізацією, тобто процесом розбиття тексту на окремі токени або слова.

5. HS/tag: У цьому каталозі містяться модулі, пов'язані з тегуванням частин мови, яке передбачає присвоєння граматичних тегів словам у тексті.

6. HS/parse: Цей каталог містить модулі для синтаксичного аналізу, який передбачає аналіз граматичної структури речень.

7. HS/sentiment: Цей каталог містить модулі для аналізу сентиментів - процесу визначення емоційного тону або настрою, вираженого в тексті.

8. HS/classify: Цей каталог містить модулі для завдань класифікації, таких як створення та навчання класифікаторів.

Це лише кілька прикладів, і HS надає набагато більше каталогів і файлів для різних завдань і функцій НЛП.

Алгоритми та технічні аспекти включають в себе широкий спектр алгоритмів і методів з області обробки природної мови. Він включає алгоритми машинного навчання для таких завдань, як класифікація, кластеризація та тегування послідовностей. HS також використовує статистичні моделі, підходи на основі правил і лінгвістичні ресурси для забезпечення точної обробки мови. Філософія проектування HS наголошує на модульності, розширюваності та простоті

використання. Він надає уніфікований API, який абстрагується від основних алгоритмів і технічних деталей, дозволяючи розробникам зосередитися на створенні NLP-додатків без необхідності реалізовувати алгоритми з нуля. Крім того, HS надає вичерпну документацію, навчальні посібники та приклади коду, що робить його доступним як для початківців, так і для досвідчених практиків НЛП.

Загалом, HS - це потужна та гнучка бібліотека НЛП, яка пропонує повний набір інструментів, алгоритмів та ресурсів для різних завдань НЛП. Модульна структура, широкі функціональні можливості та підтримка програмування на Python роблять її популярним вибором серед дослідників, розробників та викладачів у галузі обробки природної мови.

4.2 Програмна реалізація методу автоматичної класифікації фразеологічних одиниць англомовних текстів

Для того щоб доцільно розглянути програмну реалізацію нашої програми, треба оглянути основні модулі та структуру, яка приймає відповідні запити від користувача, для цього можна розглянути UML діаграму програмного забезпечення, яка зображена на рисунку 4.2.

На цій діаграмі UML інструментарій гібридного програмного забезпечення складається з кількох модулів, які надають різні функціональні можливості для завдань обробки природної мови. Ось розбивка компонентів:

1. Hybrid Soft: основний компонент представляє бібліотеку, яка служить основою для різноманітних операцій обробки природної мови.
2. Модуль токенізації: цей модуль керує процесом токенізації вхідного тексту в окремі слова або токени. Він може використовувати такі алгоритми, як регулярні вирази або методи на основі правил для виконання токенізації.
3. Модуль тегування POS: цей модуль виконує тегування частини мови (POS), призначаючи граматичні теги кожному слову в реченні. Він використовує статистичні моделі або підходи на основі правил для ідентифікації тегів POS.

4. Модуль визначення основи: Модуль визначення основи реалізує такі алгоритми, як Porter Stemmer або Snowball Stemmer, щоб скоротити слова до їхньої основи або кореневої форми [72]. Це допомагає скоротити варіації слів до їх загальної форми.

5. Модуль поділу: модуль поділу відповідає за ідентифікацію та групування слів у значущі частини або фрази. Він може використовувати такі методи, як регулярні вирази або правила граматики для виконання фрагментації.

6. Корпусний модуль: Корпусний модуль надає доступ до колекції текстових корпусів для навчання та оцінювання. Він містить різні набори даних і ресурси для різних мов і доменів.

7. Модуль класифікації: модуль класифікації пропонує функції для завдань класифікації тексту. Він включає такі алгоритми, як Naive Bayes, Decision Trees або SVM для навчання та застосування моделей класифікації до текстових даних. Діаграма демонструє високорівневі компоненти програмного забезпечення HS та їхні зв’язки. Важливо відзначити, що фактична реалізація може включати додаткові модулі, класи та взаємодії на основі конкретних функцій і вимог до завдань обробки природної мови. Модулі та пакети Python відіграють важливу роль у забезпеченні функціональності HS. Ці модулі та пакети використовуються для покращення різних аспектів задач обробки природної мови (NLP).

Регулярні вирази (re): модуль re у Python забезпечує підтримку роботи з регулярними виразами [73]. HS широко використовує регулярні вирази для таких завдань, як токенізація, зіставлення шаблонів і попередня обробка тексту. Регулярні вирази забезпечують гнучкі та потужні можливості маніпулювання текстом та зіставлення. Структури даних (колекції): модуль колекцій у Python пропонує спеціалізовані структури даних, які є корисними для завдань NLP. Наприклад, клас Counter з модуля collections часто використовується в HS для підрахунку входжень слів або лексем у тексті. Статистичні моделі (статистика): Модуль статистики надає функції для статистичних обчислень. HS використовує статистичні моделі та алгоритми для різних завдань НЛП, таких як аналіз настроїв,

моделювання мови та пошук інформації. Статистичні функції в модулі статистики дозволяють обчислювати такі показники, як середнє, медіана, мода, стандартне відхилення тощо. Файловий ввід/вивід (io): модуль io надає інструменти для обробки операцій вводу-виводу [74].

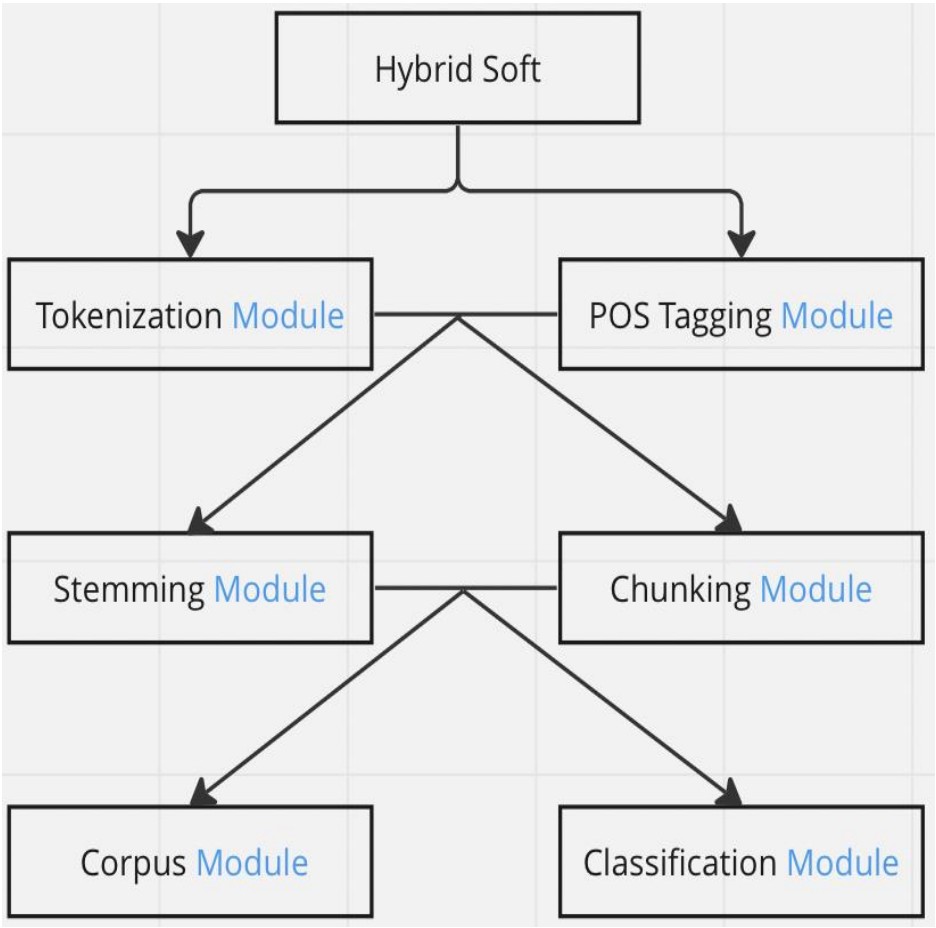


Рисунок 4.2 - UML діаграма ПЗ HS

Ось деякі важливі модулі та пакети Python, які зазвичай використовуються разом з HS:

HS використовує операції файлового вводу/виводу для читання і запису даних, завантаження мовних ресурсів і доступу до корпусів. Модуль io дозволяє читати текстові файли, маніпулювати шляхами до файлів та виконувати інші операції, пов'язані з файлами. scikit-learn: Універсальна бібліотека машинного навчання, яка включає широкий спектр алгоритмів для класифікації, регресії, кластеризації та зменшення розмірності [75]. HS може використовувати алгоритми

scikit-learn для таких завдань, як класифікація тексту або аналіз настроїв. Структура HS-сервісу, який приймає запити від користувача, зазвичай включає два основних компоненти: серверний додаток і користувацький інтерфейс. Розглянемо кожен компонент більш детально: Серверний додаток відповідає за обробку вхідних запитів від користувачів, обробку цих запитів за допомогою функціональних можливостей HS і повернення результатів. Він може бути побудований за допомогою веб-фреймворків, таких як Flask або Django, які надають необхідну інфраструктуру для створення кінцевих точок HTTP і обробки клієнт-серверного зв'язку. Серверний додаток складається з наступних компонентів – це маршрути/кінцеві точки: Визначає шляхи URL-адрес, до яких користувачі можуть звертатися для виконання запитів [76]. Кожен маршрут пов'язаний з певною функціональністю або завданням HS, наприклад, токенізацією, тегуванням частин мови або аналізом настрою.

Обробник запитів: Отримує вхідні запити, витягує необхідні дані або параметри і викликає відповідну функціональність або алгоритм HS для обробки запиту користувача. Інтеграція HS: Використовує бібліотеку HS та її різні модулі для виконання запитуваних завдань NLP. Це може включати завантаження мовних ресурсів, застосування попередньо навчених моделей або запуск певних алгоритмів. Генерація відповідей: Після завершення обробки HS сервер генерує відповідь, що містить результати виконання завдання NLP. Ця відповідь може бути в різних форматах, таких як JSON, XML або звичайний текст. Інтерфейс користувача надає користувачам можливість взаємодіяти зі службою HS і робити запити. Він може бути реалізований як веб-інтерфейс з використанням HTML, CSS і JavaScript або як окремий додаток з графічним інтерфейсом користувача (GUI) з використанням фреймворків, таких як PyQt або Tkinter.

Інтерфейс користувача зазвичай включає наступні елементи:

1) Поля введення: Дозволяють користувачам вводити текст або дані, над якими вони хочуть виконати завдання НЛП [77]. Це можуть бути текстові поля, опції завантаження файлів або меню вибору для вибору певних функцій HS.

2) Кнопка "Надіслати": Запускає відправку запиту користувача на сервер для обробки.

3) Дисплей результатів: Показує результати завдань NLP, виконаних сервером, такі як токенізований текст, мітки частин мови, результати аналізу настрою або будь-яку іншу відповідну інформацію. Текстове представлення структури зображено на рисунку 4.3.

```
NLTK Service Structure:
- Server Application
  - Routes/Endpoints
  - Request Handler
  - NLTK Integration
  - Response Generation

- User Interface
  - Input Fields
  - Submit Button
  - Output Display
```

Рисунок 4.3 - Структура ПЗ

Перед початком роботи з цією програмою ми завантажемо бази даних.

Це необов'язково, але якщо ви відчуваєте, що вам потрібні ці набори даних перед початком роботи над задачею, завантажте їх.

import HS

HS.download()

Інтерфейс програмного забезпечення Hybrid Soft за своєю структурою та візуальною частиною максимально наближений до інтерфейсу Natural Language Toolkit (NLTK) (рис. 4.4, 4.5) [78].

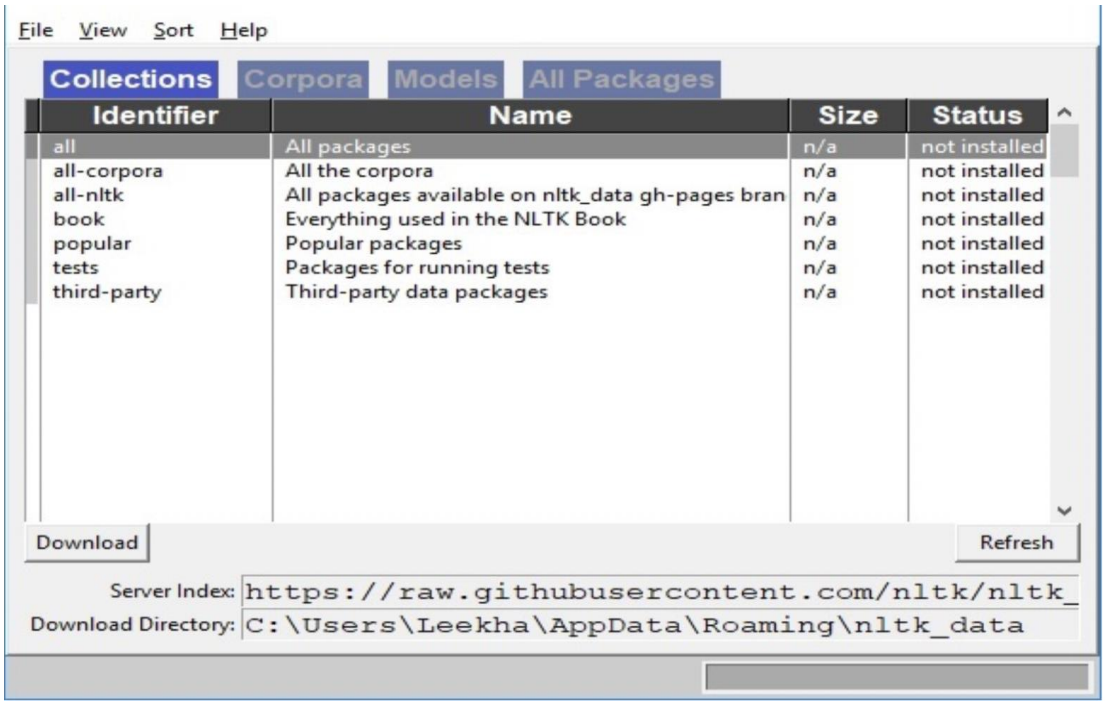


Рисунок 4.4 - Інтерфейс ПЗ NLTK

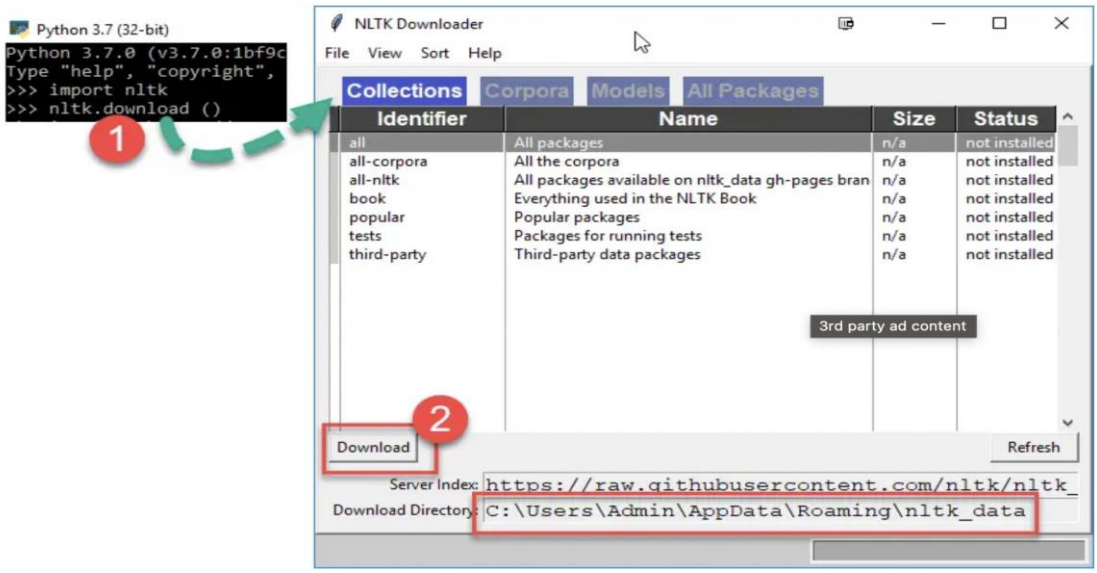


Рисунок 4.5 - Інтерфейс ПЗ NLTK

Важливою складовою нашого софта HS є токенізація. Це так зване розбиття тексту на менші одиниці, які називаються токенами. Якщо ми маємо фразеологічну одиницю, яка включає в себе більше ніж 2 лексичні одиниці, ідея полягає в тому, щоб відокремити кожне слово і проаналізувати кожне слово окремо, для

визначення попередніх ознак. Кожне слово являється окремим токеном. На рисунку 4.6 та 4.7 зображена токенізація Python коду нашої програми HS [79].

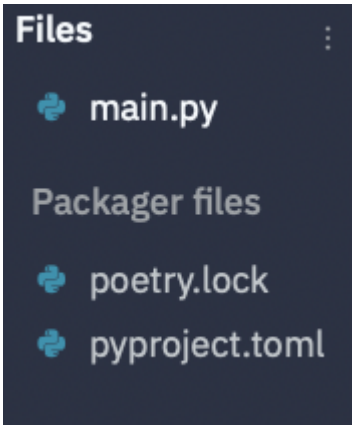


Рисунок 4.6 - Токенізація коду у програмі HS

```
1  import nltk
2  nltk.download('punkt')
3  from nltk.tokenize import sent_tokenize,
   word_tokenize
4
5  text = "Natural language processing is an
   exciting area. Huge budget have been allocated
   for this."
6
7  print(sent_tokenize(text))
8  print(word_tokenize(text))
```

Рисунок 4.7 - Токенізація коду у програмі HS

Щоб класифікувати базу з 20 фразеологічних одиниць за допомогою HS, ми виконали ці покрокові дії:

1) Встановлення HS: Інсталяція бібліотеки HS у системі. HS можна встановити за допомогою pip, інсталятора пакунків Python. Далі ми відкрили інтерфейс командного рядка і виконали наступну команду: "pip install HS".

2) Імпорт HS та попередня обробку даних: Потрібно імпортувати необхідні модулі та пакети HS до нашого коду Python. Крім того, попередньо обробити набір даних фразеологічних одиниць, щоб забезпечити узгоджене форматування та видалити будь-яку нерелевантну інформацію. Це може включати видалення розділових знаків, перетворення на малі літери або застосування стемінгу.

3) Вилучення ознак: Визначення ознаки, які будуть використані для класифікації. У випадку фразеологізмів ознаки можуть ґрунтуватися на частоті вживання, частиномовних тегах або інших лінгвістичних характеристиках. Ми використали функції та інструменти HS, щоб виокремити ці ознаки з нашого набору даних.

4) Розділ набору даних: Розділяємо попередньо оброблений набір даних на навчальний і тестовий. Навчальний набір буде використаний для навчання моделі класифікації, тоді як тестовий набір буде використаний для оцінки продуктивності моделі.

5) Вибір алгоритму класифікації: Вибір алгоритму класифікації з доступних варіантів HS. Деякі з найпоширеніших алгоритмів класифікації тексту включають наївний Байєс, максимальну ентропію та опорно-векторні машини (SVM). Ми обрали алгоритм наївного Байєса та алгоритм Named Entity Recognition (NER) [80].

Навчання класифікатора: Використовуємо навчальну вибірку для навчання обраного алгоритму класифікації. Надаємо ознаки, витягнуті з навчальних даних, разом з відповідними мітками (категоріями).

6) Оцінка моделі: Проведення тесту навченого класифікатора на тестовому наборі. Далі ми виміряємо точність або інші відповідні показники ефективності, щоб оцінити, наскільки добре модель здатна класифікувати фразеологічні одиниці.

7) Використовуємо модель для прогнозування: Після того, як модель навчена й оцінена, ми можемо використовувати її для прогнозування категорій нових фразеологізмів, які ще не зустрічалися. Надаємо витягнуті ознаки нових даних класифікатору і отримуємо передбачувані категорії.

Приклад фрагмента коду, який демонструє реалізацію вищезазначених кроків за допомогою наївного байєсівського класифікатора HS зображений на рисунку 4.8.

```
import hs
from hs.corpus import stopwords
from hs.tokenize import word_tokenize

# Preprocess the dataset
phraseological_units = [
    ("All thumbs", "Category1"),
    ("Busy as a bee", "Category2"),
    # ... Add more phraseological units with their corresponding categories
]

# Remove stopwords and perform tokenization
stop_words = set(stopwords.words('english'))
tokenized_units = []
for unit, category in phraseological_units:
    words = word_tokenize(unit)
    filtered_words = [word for word in words if word.lower() not in stop_words]
    tokenized_units.append((filtered_words, category))

# Extract features
all_words = hs.FreqDist(word.lower() for unit, _ in tokenized_units for word in unit)
word_features = list(all_words.keys())[:50]  # Use top 50 frequent words as features

def extract_features(unit):
    unit_words = set(unit)
    features = {}
    for word in word_features:
        features[word] = (word in unit_words)
    return features

# Split dataset into training and test sets
split_ratio = 0.8
split_index = int(len(tokenized_units) * split_ratio)
training_set = [(extract_features(unit), category) for unit, category in tokenized_units[:split_index]]
test_set = [(extract_features(unit), category) for unit, category in tokenized_units[split_index:]]

# Train the Naive Bayes classifier
classifier = hs.NaiveBayesClassifier.train(training_set)

# Evaluate the classifier
accuracy = hs
```

Рисунок 4.8 - Реалізація автоматичної класифікації ФО

4.3 Висновки

У даному розділі описано середовище розробки прогамно-технічної системи реалізації методу автоматичної класифікації фразеологічних одиниць в англомовних текстах та визначено основні програмні вимоги для реалізації софту. Багатоядерний процесор (наприклад, Intel Core i5 або вище) для більш швидкої обробки. Щонайменше 4 ГБ оперативної пам'яті для обробки великих наборів

даних і виконання операцій, що вимагають багато пам'яті. Достатній обсяг пам'яті для зберігання бібліотеки HS, наборів даних і будь-яких додаткових ресурсів, необхідних для ваших завдань NLP. Деякі функції HS, такі як завантаження додаткових наборів даних, можуть вимагати підключення до Інтернету. Далі ми зазначили, що HS сумісна з різними операційними системами, включаючи Windows, macOS та дистрибутиви Linux, такі як Ubuntu, Fedora та CentOS. Основною мовою програмування, за допомогою якої реалізований софт "Hybrid Soft" є Python 3.x. У другому параграфі була проведена програмна реалізація та розглянута більш детальна структура бібліотеки HS, а саме, алгоритми та технічні аспекти, модулі та пакети Python. Інтерфейс авторського програмного забезпечення максимально наближений до інтерфейсу програмного забезпечення Natural Language Toolkit (NLTK). Фінально була проведена класифікація фразеологічних одиниць та наведений покроковий алгоритм дій, як користувачу реалізувати класифікацію власноруч.

ВИСНОВОК

У роботі за результатами виконаних теоретичних та практичних досліджень розроблено гібридний метод автоматичної класифікації фразеологічних одиниць англомовних текстів, який включає в себе поєднання алгоритмів методу на основі правил та методу на основі машинного навчання. Поєднавши алгоритми, було розроблено програмне забезпечення Hybrid Soft, яке допомогло скоротити час автоматичної класифікації фразеологічних одиниць та зменшити кількість невизначених ФО із загальної бази.

У першому розділі визначено актуальність дослідження та проаналізовано предметне середовище автоматичної класифікації фразеологічних одиниць, визначено задачі, які потребують вирішення для досягнення поставленої мети.

У другому розділі опрацьовано авторський метод автоматичної класифікації ФО на основі інтеграції з алгоритмами машинного навчання та методу на основі правил. Також розроблена структура інформаційної системи, до якої входять: збір даних; попередня обробка; основні компоненти, що реалізують переваги методу на основі правил і методу на основі машинного навчання; гібридний метод та оцінку.

У третьому розділі описано основні алгоритми автоматичної класифікації ФО, які реалізують гібридний метод. Основними алгоритмами були визначені такі: SVM, Наївний Байєс, Алгоритм дерева рішень, Алгоритм вбудовування слів. Авторський метод передбачає поєднання алгоритму Наївного Байеса та алгоритму NER (Named Entity Recognition). Також була визначена технологія, яка використовується для створення та використання гібридного алгоритму, включає такі мови програмування, як Python, бібліотеки, такі як NLTK, scikit-learn і spaCy, а також інфраструктури машинного навчання, такі як TensorFlow і PyTorch.

У четвертому розділі описано опрацьоване програмне забезпечення для автоматичної класифікації ФО. Визначено необхідні вимоги для реалізації ПЗ та мову програмування.

У роботі удосконалено метод автоматичної класифікації фразеологічних одиниць англомовних текстів на основі інтеграції алгоритмів методу на основі правил і машинного навчання та удосконалено програмно-технічну систему реалізації методу автоматичної класифікації фразеологізмів.

Практична цінність отриманих результатів: у результаті виконання наукового дослідження опрацьовано програмне забезпечення автоматичної класифікації фразеологічних одиниць на основі реалізації алгоритму гібридного методу, що поєднує метод на основі правил і метод на основі машинного навчання. Це дозволило збільшити ефективність класифікації ФО англомовних текстів, зменшило час класифікації та збільшило кількість ознак для класифікації. Матеріали дипломної роботи апробовані на конференції "Автоматизація та комп’ютерно-інтегровані технології у виробництві та освіті: стан, досягнення, перспективи розвитку: матеріали Всеукраїнської науково-практичної Internet-конференції. – Черкаси, 2023. - 196 с.".

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Martinez-Blasco I. Verbos soporte y fijacio´n lexica in: *Las construcciones verbo- nominales libres y fijas*. 2008. Volume 7 (9). P.p. 47–59.

2. Huerta P.M. Estudio contrastivo lingu¨´ıstico y sema´ntico de las construcciones ver- bales fijas diato´picas mexicanas/espan˜olas in. *Las construcciones verbo-nominales libres y fijas*. 2010. Volume 5 (6). P.p. 179–198.

3. Lamiroy B. Les expressions fig´ees. *a´ la recherche d'une d´efinition*. 2008. P.p. 85–98.

4. Mejri S. Le figement lexical. *Descriptions linguistiques et structuration s´emantique.* 2015. Volume 3 . P. 45.

5. Kunin A. V. A phraseology course of the contemporary English language. *A manual for institutes and foreign languages faculties*. 2013. P.p. 78-79

6. Fazlyeva Z. K. Types of interlanguage phraseological correspondences (based on English and Turkish languages). *Review of European Studies*. 2015. Volume 7(9). P.p. 1-9.

7. Rodríguez-Piñero A. Variación y sinonimia en las locuciones. *Revista de Lingüística y Lenguas Aplicadas*. 2012. P.p. 225-238.

8. Dobrovol'skij D., Baranov A. Semanticheskie otnoshenia vo frazeologii. *Semantic relations in Phraseology*. 2014. P.p. 43.

9. Pawar A., Mago V. Calculating the similarity between words and sentences using a lexical database and corpus statistics. 2010. P.p. 1-14.

10. Darchuk N. P. Kompiuterna linhvistyka (avtomatychne opratsiuvannya textu). *Computational linguistics (automatic text processing).* 2013. P.p. 23-29.

11. Thomas J. Discovering English with Sketch Engine. *A Corpus-Based Approach to Language Exploration.* 2016. Volume 12. P. 228.

12. Tausczik Y., Pennebaker J. The psychological meaning of words. *LIWC and computerized text analysis methods*. 2011. P. p. 24–54.

13. Natural Language Toolkit. url: https://www.nltk.org/ (дата звернення: 19.03.2023)

14. Steven B. Natural Language Processing with Python. *O'Reilly Media.* 2019. P.504.

15. Stevenson S., Fazly A., North R. Statistical measures of the semi-productivity of light verb constructions. *Proceedings of the Workshop on Multiword Expres*. 2014. P.p. 1-8.

16. Cook P., Fazly A., Stevenson S. The VNC-tokens dataset. *Proceedings of the MWE workshop ACL*. 2018. P.p. 19–22.

17. Fazly A., Stevenson S. Automatically constructing a lexicon of verb phrase idiomatic combinations. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. 2016. P.p. 337–344.

18. Sinclair J. How to build a corpus. *A guide to good practice*. 2012. P.p. 78-83.

19. Ayto J. The Oxford Dictionary of Idioms. *Oxford University Press*. 2020. P.p. 46-84.

20. NER-моделі для MITIE. lang.org.ua. url: https://lang.org.ua/uk/models (дата звернення: 17.02.2023)

21. Springer Ling. Evaluating Polarity for Verbal Phraseological Units. url: https://link.springer.com/chapter/10.1007/978-3-319-13647-9_19 (дата звернення: 18.03.2023)

22. Goldberg Y. Neural Network Methods for Natural Language Processing. Morgan &amp. 2017. P.p. 19-20.

23. Indurkhya N. Nandbook of Natural Language Processing. *Chapman and Hall.* 2010. P.p. 112– 113.

24. Sager J. Language Engineering and Translation. *Consequenses of Automation.* 2014. P.23.

25. Twitto-Shmuel N., Ordan N.,Wintner S. Statistical machine translation with automatic identification of translationese. *Proceedings of WMT*. 2015. P.p. 67-68.

26. Raymond J. Mining Knowledge from Text Using Information Extraction. 2015. P.p. 3-4.

27. Mark Lutz. Learning Python. 2014. P. 246.

28. Tomas M., Iutskever I., Chen K., Corrado G., Dean J. 2011. P.32

29) Strzalkowski T., Broadwell1 G., Taylor S., Feldman L., Yamrom B., Shaikh S. Robust Extraction of Metaphors from Novel Data. *Proceedings of the First Workshop on Metaphor in NLP*. 2013. P.p. 24-25.

30) Levchenko O. Romanyshyn N. Modern approaches to automated identification of metaphor. *Philological series.* 2019. P.p. 288–298.

31) Wilson M. Mrc psycholinguistic database. *Machine-usable dictionary, version 2.00*. 2012. P.p. 6–10.

32) Kyrylyuk Y. Algorithm of automatic identification of zoomorphic metaphor. *Master's thesis*. 2010. Volume 2(3). P.p. 37-56.

33) Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications. *Ain Shams Engineering Journ*. 2014. P.p. 1093–1113.

34) Ghiassi M., Olschimke M., Moon B., Arnaudo P. Automated text classification using a dynamic artificial neural net-work model. *Expert Systems with Applications*. 2012. P.p. 967–1076.

35) Ellis N. Phraseology the periphery and the heart of language. *Phraseology in foreign language learning and teaching.* 2018. P.p. 1-13.

36) Barry V. Decision Trees for Business Intelligence and Data Mining. *Using SAS Enterprise Miner*. 2016. P.p. 1-6.

37)Analytics Vidhya. NLTK: A Beginners Hands-on Guide to Natural Language Processing. url: https://www.analyticsvidhya.com/blog/2021/07/nltk-a-beginners-hands-on-guide-to-natural-language-processing/?# (дата звернення: 03.04.2023)

38) Study Tonight. Introduction to NLP using NLTK Library in Python. url: https://www.studytonight.com/post/natural-language-processing-with-python-nltk-library (дата звернення: 03.04.2023)

39) Hunston S. Corpora and language teaching: issues of language. *Corpora in applied linguistics.* 2017. P.p. 137-169.

40) Martinez R., Schmitt N. A phrasal expressions list. *Applied Linguistics*. 2012. P.p. 299-320.

41) Paquot M. Exemplification in learner writing. *Phraseology in foreign language learning and teaching.* 2012. P.p. 101-119.

42) Skorokhodko Ye. F. Linhvistychni osnovy avtomatyzatsii informatsiynoho poshuku. *Linguistic bases of information search automation*. 2012. P.p. 1-11.

43) Popović S. Onomasiological dictionary in bilingual phraseology. *Foreign Language Teaching and Lexicography.* 2020. P. p. 141-149.

44) Davis R., Barrett L. Lexical semantic factors in the acceptability of english support-verb-nominalization constructions. *ACM Trans*. 2012. P.p. 5-15.

45) Vincze V., Zsibrita J. Learning to detect english and hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*. 2013. P.p. 6-15.

46) Riehemann Z., Wasow T., Copestake A., Clark E., Zwicky M. A constructional approach to idioms and word formation. *Tech. rep*. 2011. P.p. 1-11.

47) Sebastiani F. Machine Learning in Automated Text Categorization. *Computing Surveys*. Volume 34 (1) 2014. P. p. 45 – 48.

48) Lewis D.D., An evaluation of phrasal and clustered representations on a text categorization task. *In Proceedings of SIGIR-92, 15th ACM International Conference on Researchand Development in Information Retrieval.* 2012. P.p. 37-50.

49) Aggarwal C. Data Classification. Algorithms and Applications. *CRC Press.* 2014. P.p. 245–273.

50) Zhang X., Zhao J., LeCun Y. Character-level Convolutional Networks for Text Classification. *Proc. of the Neural Information Processing Systems Conf.* 2016. P.p. 67-68.

51) Ju R. An Efficient Method for Document Categorization Based on Word2vec and Latent Semantic Analysis. Ubiquitous Computing and Communications.

Dependable, Auto-nomic and Secure Computing. *Pervasive Intelligence and Computing*. 2015. P.p. 2276–2283.

52) Moraes R., Valiati J.F., Gavião Neto W.P. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*. 2013. P. p. 621–633.

53) Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. *The MIT Press*. 2014. P. p. 2-7.

54) Pedersen T., Patwardhan S. Michelizzi J. WordNet: Similarity—Measuring the Relatedness of Concepts. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. 2016. P. p. 1-7.

55) Turney P., Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*. 2013. P. p. 141-188.

56) Riloff E., Jones, R. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *In Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence.* 2014. P. p.474-479.

57) Li S., Zhang Z., Zhu Q. Phraseological Unit Extraction from Unstructured Texts via Linguistic Patterns and Semantic Constraints. *Cognitive Computation*. 2020. P. p. 152-167.

58) Tufiş D., Ion R., Ştefănescu D. Automatic Extraction of Phraseological Units from Corpora. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation*. 2013. P. p. 3212-3218.

59) Bojarski P., iasecki M. Automatic Identification and Extraction of Phraseological Units with the Sketch Engine. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.* 2013. P. p. 1284-1294.

60) Leech G., Hundt M., Mair C., Smith N. Change in Contemporary English: A Grammatical Study. *Cambridge University Press*. 2014. P. p. 34-38.

61) Benson M., Benson E., Ilson R. The BBI Combinatory Dictionary of English: A Guide to Word Combinations. *John Benjamins Publishing*. 2016. P. p. 66-89.

62) Gläser R., Ladewig S. H. Handbook of Phraseological Units Across Languages. *Walter de Gruyter*. 2019. P. p. 23-27.

63) Gibbs W. The Poetics of Mind: Figurative Thought, Language, and Understanding. *Cambridge University Press*. 2012. P. p. 1-4.

64) Nation P. Learning Vocabulary in Another Language. *Cambridge University Press*. 2015. P. p. 877-878.

65) Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. *Cambridge University Press*. 2017. P. p. 34-37.

66) Calzolari N. Building a Lexical Knowledge Base for Natural Language Processing: Theory and Implementation of the CIRM Project. *Springer*. 2013. P. p.66-67.

67) Cook G. The Discourse of Advertising. *Routledge*. 2013. P. p. 93-94.

68) Fellbaum C. WordNet: An Electronic Lexical Database. *The MIT Press.* 2018. P. p.45-47.

69) Fillmore J. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences.* 2013. P. p. 20-32.

70) Hanks P. Lexical Analysis. Norms and Exploitations. *MIT Press*. 2013. P. p. 4-7.

71) Tsur O., Rappoport A. Using Lexical Chains for Text Summarization. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2014. P. p. 13-18.

72) Wiebe J., Riloff E. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2015. P. p. 58-63.

73) Baroni M., Dinu G., Kruszewski G. Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* 2014. P. p. 23-26.

74) Lappin S., Leass, H. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*. 2014. P. p. 535-561.

75) Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress.* 2018. P. p. 32-43.

76) Erk K. A Simple, Similarity-Based Model for Selectional Preference. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.* 2013. P. 342.

77) McCarthy D., Keller B., Carroll J. Detecting Stepping-Stones: The Concept of Salience in Wordnet. *Proceedings of the 10th EURALEX International Congress*. 2019. P. 12.

78) Kilgarriff A., Grefenstette G. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*. 2011. P. p. 333-347.

79) Ramisch C., Gros C., Villavicencio A. From Words to Multiword Expressions: *An Integrated Account of Compositionality in Lexical Semantics. Language and Linguistics Compass*. 2017. P. p. 260-277.

80) Baker M., Ellsworth M., Erk K. SemEval. Evaluating Word Sense Induction and Discrimination Systems. *Proceedings of the 4th International Workshop on Semantic Evaluations.* 2016. P.

ДОДАТОК А

КОПІЯ НАУКОВОЇ ПУБЛІКАЦІЇ

Зінюк Євгеній Ростиславович
Хмельницький національний університет, м. Хмельницький
Боровик Олег Васильович, д.т.н., професор
Адміністрація Державної прикордонної служби України, м. Київ

ЩОДО АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ АНГЛОМОВНИХ ТЕКСТІВ

У зв'язку з бурхливим розвитком комп'ютерних технологій в останній період дедалі більше дослідників приділяють свою увагу проблемам автоматичної обробки текстів. Одним із актуальних завдань, яке стосується автоматичної обробки текстів, є автоматична класифікація фразеологічних одиниць англомовних текстів [1-2].

Це пояснюється наступними причинами.

Величезний обсяг текстових даних, що доступні сьогодні, унеможливлює «вручну» аналізувати та класифікувати всі фразеологічні одиниці, які присутні в англомовних текстах.

Фразеологічні одиниці часто відіграють вирішальну роль у розумінні та створенні мови.

Автоматична класифікація фразеологічних одиниць має численні застосування в різних сферах, включаючи навчання мови, переклад, обробку природної мови та комп'ютерну лінгвістику.

Фразеологічні одиниці можуть становити серйозну проблему для систем машинного перекладу, яким часто важко точно перекладати ідіоматичні вирази та усталені фрази.

На даний час існують різні підходи до класифікації фразеологічних одиниць англомовних текстів. Серед найбільш поширених можна виокремити такі.

Корпусний підхід. Підхід передбачає аналіз великих колекцій текстів (корпусів) для виявлення моделей і тенденцій у використанні фразеологізмів англійською мовою.

Когнітивний лінгвістичний підхід. Підхід підкреслює роль когнітивних процесів у формуванні та вживанні фразеологічних одиниць.

Граматичний підхід. Підхід акцентує увагу на граматичній структурі фразеологічних одиниць англійської мови, зокрема на їхніх синтаксичних і семантичних властивостях.

Стилістичний підхід. Підхід зосереджується на використанні фразеологічних одиниць у різних стилях англомовного письма, таких як офіційний чи неофіційний, літературний чи нелітературний.

Міжлінгвістичний підхід. Підхід порівнює фразеологічні одиниці англійської мови з фразеологічними одиницями інших мов, щоб визначити схожість і відмінності в їх структурі та вживанні.

Складність задачі класифікації фразеологічних одиниць обумовлює застосування для її вирішення різних програмних засобів. Серед найбільш поширених засобів можна виокремити такі.

Sketch Engine: Sketch Engine - програмний інструмент, який спеціалізується на корпусній лінгвістиці та дозволяє користувачам аналізувати та класифікувати фразеологізми в англомовних текстах.

Tregex - інструмент для синтаксичного аналізу та зіставлення шаблонів у тексті природною мовою.

TreeTagger - програмний інструмент для тегування частин мови та лемматизації текстів природною мовою.

Хоча використання програмних засобів для автоматичної класифікації фразеологічних одиниць в англомовних текстах має багато переваг, є також деякі недоліки, які слід враховувати. До числа таких можна віднести обмежену точність, труднощі з контекстом, обмежене охоплення, необхідність налаштування, вартість.

Одним із можливих шляхів підвищення точності та ефективності програмних засобів автоматичної класифікації фразеологічних одиниць є їх інтеграція з алгоритмами машинного навчання. Це може включати навчання алгоритмів на великих наборах даних анотованого тексту з перевіркою людиною для підтвердження точності класифікацій. Постійно вдосконалюючи точність алгоритмів за допомогою машинного навчання, програмні інструменти можуть стати більш ефективними в ідентифікації та класифікації ширшого діапазону фразеологічних одиниць, у тому числі тих, які є складними або неоднозначними.

Список використаних джерел

1. Peter M. Lee. Bayesian Statistics: An Introduction, 4th Edition / Peter M. Lee. – 486 с.

2. Sfar, I. (2008). Polylexicalite et continuite pr´edicative: le cas des locutions verbales fig´ees. Las construcciones verbo-nominales libres y fijas. Aproximaci´on contrastiva y traductol´ogica, pp. 213–221.

ДОДАТОК Б

ПРЕЗЕНТАЦІЯ ДИПЛОМНОЇ РОБОТИ

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Кафедра комп'ютерної інженерії та системного програмування

«Метод автоматичної класифікації фразеологічних одиниць англомовних текстів»

Виконав ст. групи КІ2м-21-1:
Зінюк Є.Р.
Науковий керівник:
д.т.н., проф. Боровик О.В.

Актуальність роботи

У зв'язку з бурхливим розвитком комп'ютерних технологій в останній період дедалі більше дослідників приділяють свою увагу проблемам автоматичної обробки текстів. Одним із актуальних завдань, яке стосується автоматичної обробки текстів, є автоматична класифікація фразеологічних одиниць англомовних текстів.

Актуальність роботи

Причини:

- Величезний обсяг текстових даних, що доступні сьогодні, унеможливлює «вручну» аналізувати та класифікувати всі фразеологічні одиниці, які присутні в англомовних текстах.

- Фразеологічні одиниці часто відіграють вирішальну роль у розумінні та створенні мови.
- Автоматична класифікація фразеологічних одиниць має численні застосування в різних сферах, включаючи навчання мови, переклад, обробку природної мови та комп'ютерну лінгвістику.

- Фразеологічні одиниці можуть становити серйозну проблему для систем машинного перекладу, яким часто важко точно перекладати ідіоматичні вирази та усталені фрази.

МЕТА РОБОТИ

Метою дипломної роботи є підвищення ефективності автоматичної класифікації фразеологічних одиниць англомовних текстів та зменшення кількості помилкових тлумачень.

Часткові завдання дослідження

- Аналіз існуючих методів автоматичної класифікації фразеологізмів
- Аналіз існуючого програмного забезпечення автоматичної класифікації фразеологізмів;
- Обгрунтування методу автоматичної класифікації фразеологізмів англомовних текстів на основі інтеграції алгоритмів машинного навчання;

- Розробка структури інформаційної системи автоматичної класифікації фразеологізмів;
- Проектування програмного забезпечення інформаційної системи автоматичної класифікації фразеологізмів;

- Розробка алгоритму роботи системи класифікації фразеологічних одиниць;
- Програмна реалізація методу автоматичної класифікації фразеологічних одиниць англомовних текстів.

Об’єкт та предмет дослідження

Об'єкт

Об’єктом дослідження є класифікація фразеологічних одиниць в англомовних текстах.

Предмет

Предметом дослідження є науково-методичний апарат автоматичної класифікації фразеологічних одиниць в англомовних текстах.

Практична цінність

У результаті виконання наукового дослідження опрацьовано програмне забезпечення автоматичної класифікації фразеологічних одиниць на основі реалізації алгоритму гібридного методу, що поєднує метод на основі правил і метод на основі машинного навчання.

Це дозволило збільшити ефективність класифікації ФО англомовних текстів, зменшило час класифікації та збільшило кількість ознак для класифікації.

Наукова новизна

Наукова новизна отриманих результатів:

1. Удосконалено метод автоматичної класифікації фразеологічних одиниць англомовних текстів на основі інтеграції алгоритмів методу на основі правил і машинного навчання.

2. Удосконалено програмно-технічну систему реалізації методу автоматичної класифікації фразеологізмів.

Публікації за матеріалами магістерської роботи

Матеріали дипломної роботи апробовані на конференції "Автоматизація та комп’ютерно-інтегровані технології у виробництві та освіті: стан, досягнення, перспективи розвитку: матеріали Всеукраїнської науково-практичної Internet конференції. – Черкаси, 2023. - 196 с.";.

Фразеологічні одиниці

є невід'ємною частиною мови і можуть бути класифіковані за їхньою структурою та значенням.

Розрізняють ідіоми, словосполучення, фразові дієслова та прислів'я.

Автоматична класифікація фразеологізмів

в англійських текстах є важливим завданням в обробці природної мови та може використовувати методи машинного навчання для ефективної ідентифікації та категоризації фразеологічних виразів.

Підхід	Опис	Математичні методи	Переваги	Недоліки
На основі правил	Спирається на заздалегідь визначені правила для ідентифікації та класифікації фразеологізмів	Формальні граматики, регулярні вирази	Гнучкість, зрозумілість, висока точність	Обмежена масштабованість, Схильність до помилок, Накладні витрати на технічне обслуговування, Відсутність адаптивності

Статистичні	Використовує статистичні показники для ідентифікації та класифікації фразеологізмів на основі частоти та розподілу виразів	Імовірнісні моделі, кластеризація, видобуток асоціативних правил	Масштабованість, управляючий даними, Автоматичне виділення ознак, швидкість обробки даних, Можливість узагальнення	Відсутність розуміння контексту, Зміщення даних, Доменно-залежність

Машинне навчання	Використовує набір ознак, витягнутих з виразів, і маркований набір даних для навчання класифікатора, який може автоматично ідентифікувати і класифікувати нові вирази	Дерева рішень, машини опорних векторів, глибокі нейронні мережі	Висока точність, Контекстуальне розуміння, Надійність, Масштабованість.	Потрібні великі обсяги навчальних даних, Переобладнання, Можливість інтерпретації, Відсутність прозорості

Програмне забезпечення автоматичної класифікації

GATE	General Architecture for Text Engineering (GATE) — це набір програм із відкритим вихідним кодом, який надає інструменти для обробки природної мови	Метод на основі правил
DKPro	DKPro — це набір інструментів і бібліотек для обробки природної мови, що містить засновану на правилах систему ідентифікації та класифікації фразеологічних одиниць	
LingPipe	Пакет програмного забезпечення на основі Java, який надає інструменти для обробки природної мови, включаючи систему на основі правил для фразеологічної класифікації.	

R	R є популярною мовою програмування для статистичних обчислень і має кілька бібліотек, які можна використовувати для статистичного аналізу, включаючи аналіз тексту для фразеологічної класифікації.	Метод на основі статистики
Weka	Weka надає кілька інструментів попередньої обробки для текстових даних, включаючи токенізацію, основну частину та видалення стоп-слова.	
RapidMiner	Можна створювати n-грами для фіксації контексту фразеологічних одиниць, а ознаки можна створювати на основі частоти появи слів у тексті.	

Scikit-learn	Популярна бібліотека Python, яка використовується для завдань машинного навчання, включаючи класифікацію.	Метод на основі машинного навчання
TensorFlow	Його можна використовувати для різних завдань машинного навчання, включаючи класифікацію.	
Sketch Engine	Є універсальним інструментом для аналізу фразеологічних одиниць на основі корпусу, пропонуючи функціональні можливості для створення корпусу, узгодження, вилучення, статистичного аналізу, анотування, запитів і фільтрації.	
Keras	Це високорівневий API для нейронних мереж, написаний на Python і здатний працювати поверх TensorFlow.	

Авторський метод	Суть методу	Характеристика методу в порівнянні з іншими		
		Метод на основі правил	Статистичний метод	Метод на основі машинного навчання
Гібридний метод	Гібридний метод автоматичної класифікації фразеологічних одиниць поєднує в собі підхід, заснований на правилах, для виявлення фіксованих структур і підхід машинного навчання для виявлення семантичних зв'язків, що призводить до підвищення точності та ефективності.	Найважливішою характеристикою гібридного методу порівняно з простим методом на основі правил є те, що він дозволяє виявляти більш складні та варіативні фразеологічні одиниці за допомогою алгоритмів машинного навчання.	Найважливішою характеристикою гібридного методу порівняно з простим методом, заснованим на статистичних даних, є те, що він враховує структурно-семантичні особливості фразеологічних одиниць, що призводить до більш точної та повної класифікації.	Найважливішою характеристикою є те, що він включає в себе досвід і знання людини за допомогою підходів на основі правил, які можуть підвищити точність і можливість інтерпретації класифікації.

Авторський метод	Прогнозований ефект від удосконалення
Гібридний метод	Гібридний метод може підвищити точність машинного перекладу, автоматично ідентифікуючи та перекладаючи фразеологізми на основі їхніх структурних і семантичних особливостей, а не просто покладаючись на статистичні закономірності. у задачі аналізу настрою поєднання підходів, заснованих на правилах, і машинного навчання може допомогти ідентифікувати вирази, що несуть сентимент, такі як ідіоми, словосполучення і фразові дієслова. Ці вирази можуть мати переносне значення, яке неможливо вивести з окремих слів, що ускладнює їх ідентифікацію та інтерпретацію за допомогою суто машинного навчання. Однак, залучаючи людський досвід і знання за допомогою підходів, заснованих на правилах, гібридний метод може підвищити точність і інтерпретованість аналізу настроїв.

Структура інформаційної системи

$$f = \frac{n}{N}$$

де n - це кількість входжень фразеологізму;

N - загальна кількість слів у тексті.

Ми можемо використати алгоритм k-NN, щоб класифікувати нову фразеологічну одиницю як ідіому або словосполучення на основі її особливостей. Припустимо, нова фразеологічна одиниця - це "kick the bucket". Ми можемо обчислити її характеристики, такі як частота і довжина, і визначити її k найближчих сусідів у наборі даних.

ПРОГРАМНІ ВИМОГИ ДЛЯ РЕАЛІЗАЦІЇ ГІБРИДНОГО МЕТОДУ

Складові програмного алгоритмічного забезпечення	Харектеристики ПАЗ
Центральний процесор	багатоядерний ЦП із тактовою частотою не менше 3 ГГц, Intel Core i7 або i9, AMD Ryzen 7 або 9 або Xeon.
Пам'ять (RAM)	принаймні 16 ГБ

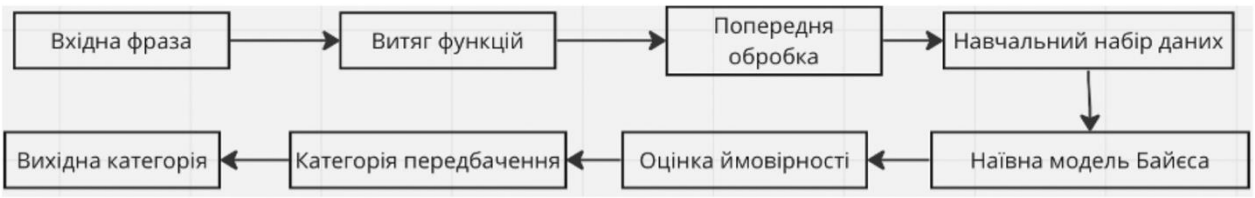
Зберігання	твердотільний накопичувач (SSD) ємністю принаймні 256 ГБ
GPU (додатково)	графічні процесори Nvidia або AMD
Мови програмування	Python або Java, Бібліотеки, такі як NLTK, spaCy і scikit-learn
Інструменти попередньої обробки даних	OpenRefine, Excel або Google Sheets
Показники оцінювання	Weka або scikit-learn
Візуалізація вихідних даних	Matplotlib, Plotly або Tableau

Алгоритми автоматичної класифікації ФО

Алгоритм	Суть алгоритму	Метод до якого відноситься
SVM	популярний алгоритм машинного навчання для завдань класифікації, оскільки він може ефективно класифікувати дані, знаходячи найкращу гіперплощину, яка розділяє різні класи у просторі великої розмірності.	Метод на основі машинного навчання
Наївний Байєс	У контексті фразеологічних одиниць алгоритм можна навчити на наборі мічених даних, де кожна фразеологічна одиниця позначена відповідно до її типу (наприклад, ідіома, прислів'я, словосполучення тощо).	Метод на основі машинного навчання
Алгоритм дерева рішень	Алгоритм працює шляхом поділу даних на підмножини на основі значення певних атрибутів, а потім рекурсивного розподілу даних на підмножини, доки не буде прийнято рішення.	Метод на основі правил
Алгоритм вбудовування слів	Алгоритм працює шляхом аналізу спільного використання слів у корпусі та використання цієї інформації для створення матриці, яка представляє зв'язки між словами.	Метод на основі правил

UML діаграма алгоритму автоматичної класифікації ФО SVM

UML діаграма алгоритму автоматичної класифікації ФО наївного Баєса



Опис середовища розробки програмно-технічної системи реалізації методу автоматичної класифікації фразеологізмів

Операційна система: HS сумісна з різними операційними системами, включаючи Windows, macOS та дистрибутиви Linux, такі як Ubuntu, Fedora та CentOS.

Python: HS реалізовано на Python, тому потрібно встановити Python. HS підтримує версії Python 2.x та Python 3.x. Однак рекомендується використовувати Python 3.x, оскільки Python 2.x більше не підтримується.

Пакети Python: HS залежить від декількох пакетів Python, які можна встановити за допомогою pip, інсталятора пакетів Python. Необхідні пакунки включають numpy, scipy, matplotlib та інші. HS надає зручну команду встановлення для встановлення всіх необхідних пакунків, яку зображено на рисунку 4.1

Опис середовища розробки програмно-технічної системи реалізації методу автоматичної класифікації фразеологізмів

Процесор: Багатоядерний процесор (наприклад, Intel Core i5 або вище) для більш швидкої обробки.

Оперативна память: Щонайменше 4 ГБ оперативної пам'яті для обробки великих наборів даних і виконання операцій, що вимагають багато пам'яті.

Сховище: Достатній обсяг пам'яті для зберігання бібліотеки HS, наборів даних і будь-яких додаткових ресурсів, необхідних для ваших завдань NLP.

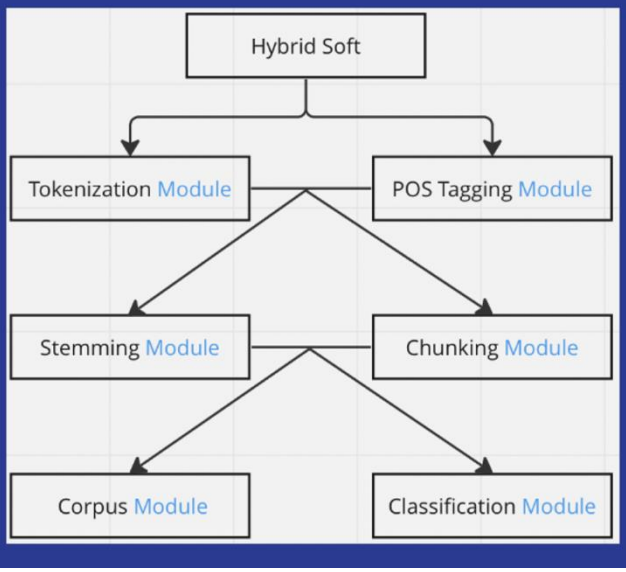
Підключення до мережі: Деякі функції HS, такі як завантаження додаткових наборів даних, можуть вимагати підключення до Інтернету.

Інтегроване середовище розробки (IDE): HS можна розробляти за допомогою будь-якого Python-сумісного IDE або текстового редактора.

Технології, які використовуються для реалізації

1) Python
2) Бібліотеки HS
3) NumPy та SciPy
4) Scikit-learn
5) Matplotlib та seaborn
6) Pandas
7) Jupyter Notebook

UML діаграма ПЗ HS

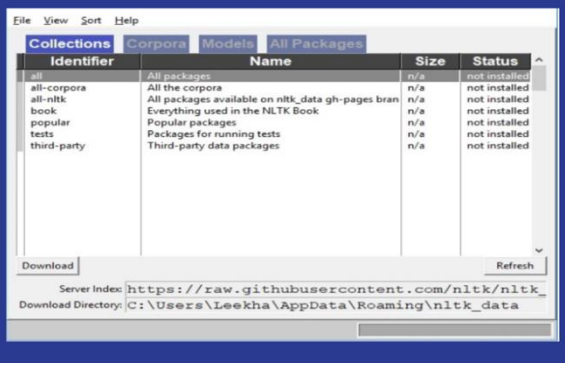


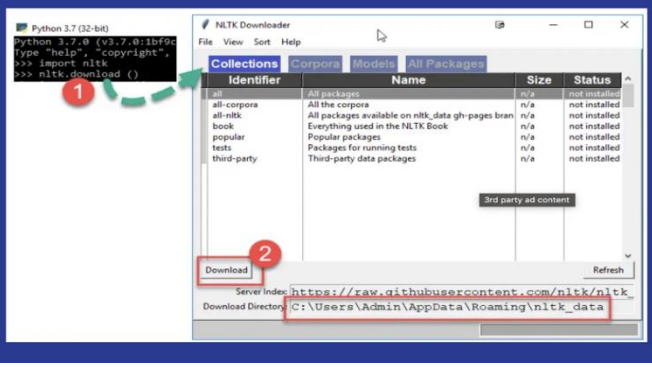
Інтерфейс користувача зазвичай включає наступні елементи:

1) Поля введення
2) Кнопка "Надіслати"
3) Дисплей результатів

Інтерфейс користувача зазвичай включає наступні елементи:

Інтерфейс програмного забезпечення Hybrid Soft за своєю структурою та візуальною частиною максимально наближений до інтерфейсу Natural Language Toolkit (NLTK)





Автоматична класифікація ФО на основі правил

Ознаки	Приклад класифікації
Ідіоми (27)	A bird in the hand is worth two in the bush, A dime a dozen, A piece of cake, All ears, Barking up the wrong tree, Bite the bullet, Break a leg, Cut the mustard, Devil's advocate, Don't count your chickens before they hatch, Drop a dime, Face the music, Fit as a fiddle, Go the extra mile, Hit the hay, Kick the bucket, Let the cat out of the bag, Piece of the action, Pull someone's leg, Rule of thumb, Silver lining, Take a rain check, Take the bull by the horns, Under the weather, When pigs fly, Wild goose chase, You can't judge a book by its cover
Формульні вирази (9)	How do you do?, It's raining cats and dogs, The whole nine yards, Let's call it a day, The early bird catches the worm, Break the ice, A stitch in time saves nine, It takes two to tango, Beat around the bush
Класифікація за походженням (6)	Ad hoc, Catharsis, Avant-garde, Coup d'état, Feng shui, Yin and yang
Ад'єктивні (19)	blue-blooded, high-tech, fair-haired, two-faced, easygoing, hardworking, big-hearted, well-off, narrow-minded, long-lasting, one-of-a-kind, old-fashioned, fast-paced, top-notch, high-spirited, short-tempered, well-known, easy-to-use, open-minded
Субстантивні (16)	the pot calling the kettle black, a safe bet, a skeleton in the closet, a square peg in a round hole, the straw that broke the camel's back, a swan song, the tail wagling the dog, the tip of the iceberg, a tough cookie, the milk of human kindness, a money pit, a loose cannon, a nest egg, the new kid on the block, a pain in the neck, a penny for your thoughts
Не увійшли у виділені ознаки класифікації (23)	All thumbs, Busy as a bee, Catch-22, Diamond in the rough, Egg on your face, Fly off the handle, Get cold feet, Hold your horses, In the same boat, Jump the gun, Keep your chin up, Leave no stone unturned, Money talks, On the ball, Pull yourself together, Saved by the bell, The whole nine yards, Up in arms, Vanishing point, Walking on eggshells, X marks the spot, You can lead a horse to water, but you can't make it drink, Zero hour

Автоматична класифікація ФО на основі гібридного методу

Ознаки	Фразеологічні одиниці
Ідіоми (27)	A bird in the hand is worth two in the bush, A dime a dozen, A piece of cake, All ears, Barking up the wrong tree, Bite the bullet, Break a leg, Cut the mustard, Devil's advocate, Don't count your chickens before they hatch, Drop a dime, Face the music, Fit as a fiddle, Go the extra mile, Hit the hay, Kick the bucket, Let the cat out of the bag, Piece of the action, Pull someone's leg, Rule of thumb, Silver lining, Take a rain check, Take the bull by the horns, Under the weather, When pigs fly, Wild goose chase, You can't judge a book by its cover
Формульні вирази (9)	How do you do?, It's raining cats and dogs, The whole nine yards, Let's call it a day, The early bird catches the worm, Break the ice, A stitch in time saves nine, It takes two to tango, Beat around the bush
Класифікація за походженням (6)	Ad hoc, Catharsis, Avant-garde, Coup d'état, Feng shui, Yin and yang
Ад'єктивні (19)	blue-blooded, high-tech, fair-haired, two-faced, easygoing, hardworking, big-hearted, well-off, narrow-minded, long-lasting, one-of-a-kind, old-fashioned, fast-paced, top-notch, high-spirited, short-tempered, well-known, easy-to-use, open-minded
Субстантивні (16)	the pot calling the kettle black, a safe bet, a skeleton in the closet, a square peg in a round hole, the straw that broke the camel's back, a swan song, the tail wagling the dog, the tip of the iceberg, a tough cookie, the milk of human kindness, a money pit, a loose cannon, a nest egg, the new kid on the block, a pain in the neck, a penny for your thoughts
Фразеологічні одиниці пов'язані з тваринами (10)	Busy as a bee, Hold your horses, You can lead a horse to water, but you can't make it drink, A bird in the hand is worth two in the bush, Let the cat out of the bag, When pigs fly, Wild goose chase, It's raining cats and dogs, The early bird catches the worm, the tail wagling the dog, a swan song
Фразеологічні одиниці пов'язані з тілом (9)	All thumbs, Egg on your face, Fly off the handle, Get cold feet, Keep your chin up, All ears, Break a leg, Pull someone's leg, Up in arms
Не увійшли у виділені ознаки класифікації (14)	Catch-22, Diamond in the rough, In the same boat, Jump the gun, Leave no stone unturned, Money talks, On the ball, Pull yourself together, Saved by the bell, The whole nine yards, Vanishing point, Walking on eggshells, X marks the spot, Zero hour

ВИСНОВКИ

У роботі за результатами виконаних теоретичних та практичних досліджень розроблено гібридний метод автоматичної класифікації фразеологічних одиниць англомовних текстів, який включає в себе поєднання алгоритмів методу на основі правил та методу на основі машинного навчання. Поєднавши алгоритми, було розроблено програмне забезпечення Hybrid Soft, яке допомогло скоротити час автоматичної класифікації фразеологічних одиниць та зменшити кількість невизначених ФО із загальної бази.

ДОДАТОК В

ОПИС ПІДХОДІВ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФО

Підхід	Опис	Математичні методи	Переваги	Недоліки
На основі правил	Спирається на заздалегідь визначені правила для ідентифікації та класифікації фразеологізмів	Формальні граматики, регулярні вирази	Гнучкість, зрозумілість, висока точність	Обмежена масштабованість, Схильність до помилок, Накладні витрати на технічне обслуговування, Відсутність адаптивності
Статистичні	Використовує статистичні показники для ідентифікації та класифікації фразеологізмів на основі частоти та розподілу виразів	Імовірнісні моделі, кластеризація, видобуток асоціативних правил	Масштабованість, управляючий даними, Автоматичне виділення ознак, швидкість обробки даних, Можливість узагальнення	Відсутність розуміння контексту, Зміщення даних, Доменно-залежність

Машинне навчання	Використовує набір ознак, витягнутих з виразів, і маркований набір даних для навчання класифікатора, який може автоматично ідентифікувати і класифікувати нові вирази	Дерева рішень, машини опорних векторів, глибокі нейронні мережі	Висока точність, Контекстуальне розуміння, Надійність, Масштабованість.	Потрібні великі обсяги навчальних даних, Переобладнання, Можливість інтерпретації, Відсутність прозорості

ДОДАТОК Г

МОЖЛИВОСТІ ПЗ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФО

ПЗ автоматичної класифікації методів	Можливості ПЗ	Метод класифікації, який відповідає ПЗ
GATE	General Architecture for Text Engineering (GATE) — це набір програм із відкритим вихідним кодом, який надає інструменти для обробки природної мови	Метод на основі правил
DKPro	DKPro — це набір інструментів і бібліотек для обробки природної мови, що містить засновану на правилах систему ідентифікації та класифікації фразеологічних одиниць	
LingPipe	Пакет програмного забезпечення на основі Java, який надає інструменти для обробки природної мови, включаючи систему на основі правил для фразеологічної класифікації.	
R	R є популярною мовою програмування для статистичних обчислень і має кілька бібліотек, які можна використовувати для статистичного аналізу, включаючи аналіз тексту для фразеологічної класифікації.	Метод на
Weka	Weka надає кілька інструментів попередньої обробки	

	для текстових даних, включаючи токенізацію, основну частину та видалення стоп-слова. Weka надає кілька інструментів оцінювання, які можна використовувати для оцінки ефективності моделей класифікації, включаючи перехресну перевірку та аналіз матриці плутанини.	основі статистики
RapidMiner	Можна створювати n-грами для фіксації контексту фразеологічних одиниць, а ознаки можна створювати на основі частоти появи слів у тексті. Інструменти для оптимізації моделі, можуть бути використані для оптимізації продуктивності моделей класифікації та визначення найбільш інформативних ознак для фразеологічної класифікації.	
Scikit-learn	Популярна бібліотека Python, яка використовується для завдань машинного навчання, включаючи класифікацію. Він надає різні алгоритми для класифікації, такі як логістична регресія, дерева рішень і опорні векторні машини.	Метод на основі машинного навчання
TensorFlow	Його можна використовувати для різних завдань машинного навчання, включаючи класифікацію. Інструмент надає кілька алгоритмів, таких як глибокі нейронні мережі та згорткові нейронні мережі.	
Sketch Engine	Є універсальним інструментом для аналізу фразеологічних одиниць на основі корпусу, пропонуючи функціональні можливості для створення корпусу, узгодження, вилучення,	

	статистичного аналізу, анотування, запитів і фільтрації.	
Keras	Це високорівневий API для нейронних мереж, написаний на Python і здатний працювати поверх TensorFlow. Він забезпечує простий і легкий у використанні інтерфейс для побудови нейронних мереж	

ДОДАТОК Ґ

ХАРАКТЕРИСТИКА МЕТОДУ В ПОРІВНЯННІ З ІНШИМИ ТА ПРОГНОЗОВАНИЙ ЕФЕКТ ВІД УДОСКОНАЛЕННЯ

Авторський метод	Суть методу	Характеристика методу в порівнянні з іншими			Прогнозований ефект від удосконалення
		Метод на основі правил	Статистичний метод	Метод на основі машинного навчання	
Гібридний метод	Гібридний метод автоматичної класифікації фразеологічних одиниць поєднує в собі підхід, заснований на правилах, для виявлен	Найважливішою характеристикою гібридного методу порівняно з простим методом на основі правил	Найважливішою характеристикою гібридного методу порівняно з простим методом, заснованим на статистичних даних, є те, що він враховує структурно-	Найважливішою характеристикою є те, що він включає в себе досвід і знання людини за допомогою підходів на основі правил, які можуть	Гібридний метод може підвищити точність машинного перекладу, автоматично ідентифікуючи та перекладаючи фразеологізми на основі їхніх структурних і семантичних особливостей, а не просто покладаючись на статистичні закономірності. у задачі аналізу настрою поєднання підходів, заснованих на правилах, і машинного навчання може

	ня фіксованих структур і підхід машинного навчання для виявлення семантичних зв’язків, що призводить до підвищення точності та ефективності.	є те, що він дозволяє виявляти більш складні та варіативні фразеологічні одиниці за допомогою алгоритмів машинного навчання.	семантичні особливості фразеологічних одиниць, що призводить до більш точної та повної класифікації.	підвищити точність і можливість інтерпретації класифікації.	допомогти ідентифікувати вирази, що несуть сентимент, такі як ідіоми, словосполучення і фразові дієслова. Ці вирази можуть мати переносне значення, яке неможливо вивести з окремих слів, що ускладнює їх ідентифікацію та інтерпретацію за допомогою суто машинного навчання. Однак, залучаючи людський досвід і знання за допомогою підходів, заснованих на правилах, гібридний метод може підвищити точність і інтерпретованість аналізу настроїв.

ДОДАТОК Д

АВТОМАТИЧНА КЛАСИФІКАЦІЯ ФО НА ОСНОВІ ЗАСТОСУВАННЯ SKETCH ENGINE

Ознаки	**Приклад класифікації**
Ідіоми (27)	A bird in the hand is worth two in the bush, A dime a dozen, A piece of cake, All ears, Barking up the wrong tree, Bite the bullet, Break a leg, Cut the mustard, Devil's advocate, Don't count your chickens before they hatch, Drop a dime, Face the music, Fit as a fiddle, Go the extra mile, Hit the hay, Kick the bucket, Let the cat out of the bag, Piece of the action, Pull someone's leg, Rule of thumb, Silver lining, Take a rain check, Take the bull by the horns, Under the weather, When pigs fly, Wild goose chase, You can't judge a book by its cover
Формульні вирази (9)	How do you do?, It's raining cats and dogs, The whole nine yards, Let's call it a day, The early bird catches the worm, Break the ice, A stitch in time saves nine, It takes two to tango, Beat around the bush
Класифікація за походженням (6)	Ad hoc, Catharsis, Avant-garde, Coup d'état, Feng shui, Yin and yang
Ад’єктивні (19)	blue-blooded, high-tech, fair-haired, two-faced, easygoing, hardworking, big-hearted, well-off, narrow-minded, long-lasting, one-of-a-kind, old-fashioned, fast-paced, top-notch, high-spirited, short-tempered, well-known, easy-to-use, open-minded
Субстантивні (16)	the pot calling the kettle black, a safe bet, a skeleton in the closet, a square peg in a round hole, the straw that broke the camel's back, a swan song, the tail wagling the dog, the tip of the iceberg, a tough cookie, the milk of human kindness, a money pit, a loose cannon, a nest egg, the new kid on the block, a pain in the neck, a penny for your thoughts
Не увійшли у виділені ознаки класифікації (23)	All thumbs, Busy as a bee, Catch-22, Diamond in the rough, Egg on your face, Fly off the handle, Get cold feet, Hold your horses, In the same boat, Jump the gun, Keep your chin up, Leave no stone unturned, Money talks, On the ball, Pull yourself together, Saved by the bell, The whole nine yards, Up in arms, Vanishing

	point, Walking on eggshells, X marks the spot, You can lead a horse to water, but you can't make it drink, Zero hour
Фразеологічна одиниця	**Пояснення**
the pot calling the kettle black	хтось звинувачує іншу людину в чомусь, у чому сам винен.
A bird in the hand is worth two in the bush	краще триматися за те, що маєш, ніж ризикувати втратити це, намагаючись отримати щось краще
a safe bet	те, що, ймовірно, станеться або вдасться
How do you do?	Офіційне привітання, яке використовується для демонстрації ввічливості.
a skeleton in the closet	таємниця або незручний факт, який хтось хоче зберегти прихованим.
Ad hoc	латинська фраза означає «для цієї мети» і використовується для опису чогось, що створюється або робиться спеціально для конкретної ситуації чи проблеми, а не є частиною загального плану.
A dime a dozen	дуже поширене або легкодоступне
the spice of life	різноманітність і хвилювання в житті
It's raining cats and dogs	спосіб описати сильний дощ.
a square peg in a round hole	людина, яка не підходить для певної ситуації.
All thumbs	незграбні руки
the straw that broke the camel's back	остання, здавалося б, незначна подія, що спричиняє велику проблему або невдачу.
a swan song	остаточне виконання або досягнення перед відставкою або смертю
A piece of cake	те, що дуже легко зробити
the tail wagling the dog	ситуація, в якій незначний або незначний фактор має непропорційний вплив на більшу систему або процес.
Busy as a bee	дуже працьовитий або працьовитий
the tip of the iceberg	невелика або видима частина набагато більшої проблеми або питання.
Catch-22	ситуація, в якій людина потрапляє в пастку суперечливих або парадоксальних правил або обставин
All ears	повністю слухають і звертають увагу

a tough cookie	сильна або витривала людина
the milk of human kindness	добра і співчутлива поведінка
Diamond in the rough	хтось або щось з великим потенціалом або талантом, але потребує доопрацювання або розвитку
a money pit	проект або підприємство, яке постійно потребує більше грошей, ніж створює.
a loose cannon	людина непередбачувана і некерована, часто завдає збитків або неприємностей.
Barking up the wrong tree	дотримуватися помилкового або неправильного курсу дій
a nest egg	сума грошей, відкладена на майбутнє
Egg on your face	бути збентеженим або приниженим власними діями або помилками
Yin and yang	Ця китайська концепція стосується взаємодоповнюючих і протилежних сил Всесвіту, часто представлених у вигляді двох половин кола або символу.
Fly off the handle	стати раптово і нестримно розлюченим
Bite the bullet	погодитеся з чимось важким або неприємним
Get cold feet	нервувати або боятися рішення чи способу дії
the new kid on the block	нова особа чи річ у певній сфері чи місці.
It takes two to tango	фраза, яка використовується для припущення, що обидві сторони несуть відповідальність за ситуацію чи проблему.
old-fashioned	застарілий або не відповідає сучасній моді чи стилю
Hold your horses	почекати або проявити терпіння, перш ніж приступити до чогось
a pain in the neck	хтось або щось дратує або дратує
Break a leg	удача (часто використовується в контексті сценічного мистецтва)
a penny for your thoughts	прохання, щоб хтось поділився своїми думками або почуттями.
In the same boat	в тій же складній або складній ситуації, що і хтось інший
A stitch in time saves nine	фраза, яка використовується для припущення, що вжиття профілактичних заходів зараз може заощадити більше часу та зусиль пізніше.
Jump the gun	діяти занадто швидко або передчасно

Cut the mustard	до вдалого виконання або виправдання очікувань
Keep your chin up	залишатися позитивним і оптимістичним перед обличчям труднощів або негараздів
fast-paced	відбувається швидко і з великою енергією
Leave no stone unturned	докласти всіх зусиль для досягнення мети або знайти рішення
Devil's advocate	той, хто займає позицію, з якою він не обов'язково погоджується, заради аргументації чи дослідження процесу прийняття рішення.
Money talks	влада або вплив грошей на прийняття рішень або переговори
Don't count your chickens before they hatch	не припускайте, що щось трапиться, перш ніж це станеться
Drop a dime	подзвонити
On the ball	пильний, уважний, ефективний
well-off	мати фінансову забезпеченість або багатство
Pull yourself together	відновити самовладання або контроль після важкого або емоційного досвіду
top-notch	найвищої якості або стандарту
Face the music	змиритися з неприємними наслідками своїх дій
Saved by the bell	ледве уникнути важкої або неприємної ситуації в останню хвилину
Fit as a fiddle	бути в хорошому фізичному стані
The whole nine yards	викластися на повну або зробити щось на повну
Feng shui	ця китайська фраза відноситься до системи дизайну та облаштування, заснованої на принципах гармонії та балансу, яка часто використовується в архітектурі та дизайні інтер'єру.
Up in arms	бути дуже злим або засмученим через щось
long-lasting	тривалий або триваючий протягом тривалого часу
Vanishing point	точка на відстані, де об'єкти ніби зливаються в одну лінію або зовсім зникають
Go the extra mile	докласти додаткових зусиль або вийти за межі очікуваного

Walking on eggshells	бути дуже обережним і обережним у своїх діях або словах, щоб не засмутити когось
one-of-a-kind	унікальний, винятковий і ні на що не схожий
Hit the hay	лягти спати
Kick the bucket	померти
X marks the spot	фраза, яка використовується для позначення місця розташування чогось важливого чи цінного
Beat around the bush	фраза, яка використовується для опису людини, яка уникає переходу на суть або ухиляється.
big-hearted	щедрий, добрий і турботливий
Let the cat out of the bag	передчасно розкрити таємницю або здивувати
You can lead a horse to water, but you can't make it drink	ви можете дати комусь можливість або пораду, але ви не можете змусити їх прийняти це
Piece of the action	частка або частина чого-небудь
Break the ice	фраза, яка використовується для опису процесу початку розмови з кимось, хто незнайомий або стриманий.
Coup d'état	ФО французького походження, ця фраза відноситься до раптового і часто насильницького повалення уряду або правлячої влади.
Pull someone's leg	дражнити або жартувати з ким-небудь добродушно
easygoing	розслаблений і толерантний, не легко засмучуватися або піддаватися стресу
Rule of thumb	загальна настанова чи правило, засноване на досвіді чи практиці
Zero hour	точний момент, коли очікується, що станеться щось важливе або значуще.
The early bird catches the worm	Фраза, яка використовується для припущення, що найбільше виграють ті, хто діє на початку ситуації.
Silver lining	позитивний аспект у важкій або неприємній ситуації
hardworking	вкладати багато зусиль і часу в завдання
Take a rain check	відхиліть запрошення, але запропонуйте прийняти його пізніше
Avant-garde	це французьке словосполучення означає «передовий захисник» і використовується для опису інноваційних або експериментальних ідей або практик, особливо в мистецтві.

Take the bull by the horns	зіткнутися з проблемою лоб в лоб
Under the weather	погане самопочуття
Let's call it a day	фраза, яка використовується для підказки того, що настав час припинити працювати чи щось робити.
When pigs fly	такого ніколи не буде
high-tech	передбачають передові технології
Wild goose chase	безнадійна або марна гонитва
narrow-minded	небажання розглядати нові ідеї чи думки
Catharsis	це грецьке слово означає очищення або звільнення від емоцій, часто за допомогою певної форми мистецтва чи вистави.
The whole nine yards	фраза, яка використовується для вказівки повного обсягу чогось.
You can't judge a book by its cover	ви не можете судити про щось лише за зовнішнім виглядом
easy-to-use	простий або зрозумілий в експлуатації або поводженні
blue-blooded	мають дворянське або аристократичне походження
well-known	відомий або знайомий багатьом людям
open-minded	сприйнятливий до нових ідей і готовий враховувати різні думки
fair-haired	прихильність або перевага; також відноситься до світлого кольору волосся
two-faced	нечесний або брехливий; говорити одне, а робити інше
high-spirited	жваві, енергійні та повні ентузіазму
short-tempered	легко піддається провокаціям або роздратуванням, має запальний і інтенсивний характер

Ідіоми

Для виявлення ідіоматичних фразеологізмів за допомогою Sketch Engine користувач може скористатися інструментом "Word Sketch", який надає детальний аналіз певного слова чи словосполучення та їхніх відповідників.

Щоб скористатися інструментом "Word Sketch" для виявлення ідіоматичних фразеологізмів, спочатку потрібно ввести ключове слово, яке, ймовірно, зустрічається в ідіоматичних фразах.

Після того, як введене ключове слово, можна переглянути ескіз слова, щоб побачити всі слова і фрази, які часто зустрічаються в тому ж контексті, що і ключове слово. Потім він може проаналізувати фрази, щоб виявити ті з них, які мають не буквальне значення, відмінне від буквального значення окремих слів.

Наприклад, словесний ескіз для ключового слова "cake" може включати такі фрази, як "A piece of cake". Ця фраза має ідіоматичне значення, яке відрізняється від буквального значення слів.

На додаток до інструменту "Схеми слів", Sketch Engine також включає модуль "Фразеологія", який можна використовувати для ідентифікації та вилучення різних типів фразеологізмів, включаючи ідіоматичні фразеологізми. Модуль містить кілька підкатегорій для ідіоматичних фразеологізмів, таких як "сталі вирази" та "словосполучення". Та якщо вибрати усі підкатегорії, то модуль буде виділяти усі ідіоматичні фразеологічні одиниці.

A bird in the hand is worth two in the bush - краще триматися за те, що маєш, ніж ризикувати втратити це, намагаючись отримати щось краще
A dime a dozen - дуже поширене або легкодоступне
A piece of cake - те, що дуже легко зробити
All ears - повністю слухають і звертають увагу
Barking up the wrong tree - дотримуватися помилкового або неправильного курсу дій
Bite the bullet - погодитеся з чимось важким або неприємним
Break a leg - удача (часто використовується в контексті сценічного мистецтва)
Cut the mustard - до вдалого виконання або виправдання очікувань
Devil's advocate - той, хто займає позицію, з якою він не обов'язково погоджується, заради аргументації чи дослідження процесу прийняття рішення.
Don't count your chickens before they hatch - не припускайте, що щось трапиться, перш ніж це станеться
Drop a dime - подзвонити
Face the music - змиритися з неприємними наслідками своїх дій
Fit as a fiddle - до хорошого фізичного здоров'я
Go the extra mile - докласти додаткових зусиль або вийти за межі очікуваного
Hit the hay — лягти спати
Kick the bucket - померти
Let the cat out of the bag - передчасно розкрити таємницю або здивувати
Piece of the action - частка або частина чого-небудь
Pull someone's leg - дражнити або жартувати з ким-небудь добродушно
Rule of thumb - загальна настанова чи правило, засноване на досвіді чи практиці
Silver lining - позитивний аспект у важкій або неприємній ситуації
Take a rain check – відхиліть запрошення, але запропонуйте прийняти його пізніше
Take the bull by the horns- зіткнутися з проблемою лоб в лоб
Under the weather - погане самопочуття
When pigs fly - такого ніколи не буде
Wild goose chase - безнадійна або марна гонитва
You can't judge a book by its cover - ви не можете судити про щось лише за зовнішнім виглядом

Формульні вирази

Щоб використовувати Sketch Engine для ідентифікації шаблонних виразів, користувачеві потрібно буде вибрати модуль «Фразеологія» та вибрати мову та корпус, який він хоче проаналізувати. Потім ми використали інструмент «Формульні вирази» для пошуку певних типів формульних виразів.

Інструмент надасть список шаблонних виразів, які відповідають критеріям пошуку, а також інформацію про їх частоту, поширення та інші мовні особливості. Користувачі також можуть аналізувати узгодженість кожного шаблонного виразу, щоб зрозуміти, як він використовується в контексті, і визначити будь-які варіації чи підтипи.

На додаток до інструменту «Формульні вирази», Sketch Engine також містить інші інструменти, які можна використовувати для аналізу формульних виразів, наприклад інструмент «Ескіз слів», який забезпечує детальний аналіз окремого слова чи фрази та його співрозмовників, а також інструмент «Тезаурус», який дозволяє користувачам досліджувати споріднені та подібні слова та фрази.

How do you do? - Офіційне привітання, яке використовується для демонстрації ввічливості.
It's raining cats and dogs - спосіб описати сильний дощ.
The whole nine yards - фраза, яка використовується для вказівки повного обсягу чогось.
Let's call it a day - фраза, яка використовується для підказки того, що настав час припинити працювати чи щось робити.
The early bird catches the worm Фраза, яка використовується для припущення, що найбільше виграють ті, хто діє на початку ситуації.
Break the ice - фраза, яка використовується для опису процесу початку розмови з кимось, хто незнайомий або стриманий.
Beat around the bush - фраза, яка використовується для опису людини, яка уникає переходу на суть або ухиляється.
A stitch in time saves nine - фраза, яка використовується для припущення, що вжиття профілактичних заходів зараз може заощадити більше часу та зусиль пізніше.
It takes two to tango - фраза, яка використовується для припущення, що обидві сторони несуть відповідальність за ситуацію чи проблему.

ФО за походженням

Для ідентифікації фразеологічних одиниць за походженням, програмі необхідний доступ до повної бази даних фразеологічних одиниць та їх походження. Цього можна досягти за допомогою поєднання методів обробки природної мови, алгоритмів машинного навчання та лінгвістичної експертизи.
Щоб використати Sketch Engine для ідентифікації фразеологічних одиниць за походженням, спочатку потрібно вибрати модуль «Фразеологія» та вибрати мову та корпус, які він хоче проаналізувати. Потім потрібно використати функцію пошуку, щоб знайти фразеологічні одиниці, класифіковані за походженням, наприклад «латинські фрази» або «грецькі вирази».
Крім того, можна використовувати інструмент узгодження, щоб витягти та класифікувати фразеологічні одиниці з текстового корпусу на основі їх походження. Це передбачає виконання пошуку всіх екземплярів фразеологічних одиниць у корпусі та подальший аналіз результатів для визначення їх походження.

Ad hoc – ця латинська фраза означає «для цієї мети» і використовується для опису чогось, що створюється або робиться спеціально для конкретної ситуації чи проблеми, а не є частиною загального плану.

Catharsis – це грецьке слово означає очищення або звільнення від емоцій, часто за допомогою певної форми мистецтва чи вистави.

Avant-garde – це французьке словосполучення означає «передовий захисник» і використовується для опису інноваційних або експериментальних ідей або практик, особливо в мистецтві.
Coup d'état - також французького походження, ця фраза відноситься до раптового і часто насильницького повалення уряду або правлячої влади.

Feng shui - ця китайська фраза відноситься до системи дизайну та облаштування, заснованої на принципах гармонії та балансу, яка часто використовується в архітектурі та дизайні інтер'єру.

Yin and yang Ця китайська концепція стосується взаємодоповнюючих і протилежних сил Всесвіту, часто представлених у вигляді двох половин кола або символу.

Ад’єктивні

Щоб використати інструмент «Word Sketch» для визначення фразеологічних одиниць прикметників, потрібно спочатку ввести прикметник у пошуковий рядок інструменту. Потім переглянути ескіз слова, щоб побачити всі слова та фрази, які часто зустрічаються в тому самому контексті, що й прикметник.
параметри та фільтри які ми використали:

Виберіть корпус: Виберіть корпус, у якому ви хочете шукати прикметникові фразеологізми.
Налаштуйте параметри пошуку: виберіть «Фраза» як параметр пошуку, який дозволить вам шукати багатослівні фрази.
Введіть пошуковий термін: введіть прикметник, який, на вашу думку, є частиною прикметникового фразеологізму, наприклад «червоний» або «синій».
Виберіть вкладку «Колокації». Тут ви побачите слова, які зазвичай зустрічаються в тому самому контексті, що й пошуковий термін.
Фільтрувати за частиною мови: у стовпці «Частина мови» виберіть фільтр «Прикметник», щоб відображати лише прикметники, які зазвичай зустрічаються разом із пошуковим терміном.
Фільтрувати за семантичним типом: у стовпці «Семантичний тип» виберіть фільтр «Фразеологічні», щоб відобразити лише фразеологізми, які зазвичай зустрічаються разом із пошуковим терміном.
Ознайомтеся з результатами: перегляньте список фразеологізмів ад’єктивної форми, які відображені в інструменті «Word Sketch». Ці одиниці будуть фразами, які містять пошуковий термін і мають значення прикметника, яке відрізняється від буквального значення окремих слів.

blue-blooded - мають дворянське або аристократичне походження
high-tech - передбачають передові технології

fair-haired - прихильність або перевага; також відноситься до світлого кольору волосся
two-faced - нечесний або брехливий; говорити одне, а робити інше
easygoing - розслаблений і толерантний, не легко засмучуватися або піддаватися стресу
hardworking - вкладати багато зусиль і часу в завдання
big-hearted - щедрий, добрий і турботливий
well-off - мати фінансову забезпеченість або багатство
narrow-minded - небажання розглядати нові ідеї чи думки
long-lasting - тривалий або триваючий протягом тривалого часу
one-of-a-kind - унікальний, винятковий і ні на що не схожий
old-fashioned - застарілий або не відповідає сучасній моді чи стилю
fast-paced - відбувається швидко і з великою енергією
top-notch - найвищої якості або стандарту
high-spirited - жваві, енергійні та повні ентузіазму
short-tempered - легко піддається провокаціям або роздратуванням, має запальний і інтенсивний характер
well-known - відомий або знайомий багатьом людям
easy-to-use - простий або зрозумілий в експлуатації або поводженні
open-minded - сприйнятливий до нових ідей і готовий враховувати різні думки

Субстантивні

Щоб визначити лише субстантивні фразеологічні одиниці зі змішаного списку фразеологічних одиниць у Sketch Engine, ми скористалися інструментом «Word Sketch» і встановити певні параметри та фільтри. Ось приклад того, які параметри та фільтри можна використовувати:

Виберіть корпус: Виберіть корпус, у якому ви хочете шукати субстантивні фразеологічні одиниці.
Налаштуйте параметри пошуку: виберіть «Фраза» як параметр пошуку, який дозволить вам шукати багатослівні фрази.
Введіть пошуковий термін: введіть іменник, який, на вашу думку, є частиною субстантивної фразеологічної одиниці, наприклад «серце» або «голова».
Використовуйте інструмент «Word Sketch».
Виберіть вкладку «Колокації». Тут ви побачите слова, які зазвичай зустрічаються в тому самому контексті, що й пошуковий термін.
Фільтрувати за частиною мови: у стовпці «Частина мови» виберіть фільтр «Іменник», щоб відображати лише іменники, які зазвичай зустрічаються разом із пошуковим терміном.
Фільтрувати за семантичним типом: у стовпці «Семантичний тип» виберіть фільтр «Фразеологічні», щоб відобразити лише фразеологізми, які зазвичай зустрічаються разом із пошуковим терміном.
Ознайомтеся з результатами: перегляньте список субстантивних фразеологізмів, які відображені в інструменті «Word Sketch». Ці одиниці будуть фразами, які містять пошуковий термін і мають змістовне значення, яке відрізняється від буквального значення окремих слів.

the pot calling the kettle black- хтось звинувачує іншу людину в чомусь, у чому сам винен.
a safe bet - те, що, ймовірно, станеться або вдасться
a skeleton in the closet - таємниця або незручний факт, який хтось хоче зберегти прихованим.
a square peg in a round hole - людина, яка не підходить або не підходить для певної ситуації.
the straw that broke the camel's back - остання, здавалося б, незначна подія, що спричиняє велику проблему або невдачу.
a swan song - остаточне виконання або досягнення перед відставкою або смертю.
the tail wagling the dog - ситуація, в якій незначний або незначний фактор має непропорційний вплив на більшу систему або процес.
the tip of the iceberg - невелика або видима частина набагато більшої проблеми або питання.
a tough cookie - сильна або витривала людина.
the milk of human kindness - добра і співчутлива поведінка.
a money pit - проект або підприємство, яке постійно потребує більше грошей, ніж створює.
a loose cannon - людина непередбачувана і некерована, часто завдає збитків або неприємностей.
a nest egg - сума грошей, відкладена на майбутнє.
the new kid on the block - нова особа чи річ у певній сфері чи місці.
a pain in the neck - хтось або щось дратує або дратує.
a penny for your thoughts - прохання, щоб хтось поділився своїми думками або почуттями.

Не увійшли у виділені ознаки класифікації

All thumbs - незграбні або незграбні руки
Busy as a bee - дуже працьовитий або працьовитий
Catch-22 - ситуація, в якій людина потрапляє в пастку суперечливих або парадоксальних правил або обставин
Diamond in the rough - хтось або щось з великим потенціалом або талантом, але потребує доопрацювання або розвитку
Egg on your face - бути збентеженим або приниженим власними діями або помилками
Fly off the handle - стати раптово і нестримно розлюченим
Get cold feet - нервувати або боятися рішення чи способу дії
Hold your horses - почекати або проявити терпіння, перш ніж приступити до чогось
In the same boat - в тій же складній або складній ситуації, що і хтось інший
Jump the gun - діяти занадто швидко або передчасно
Keep your chin up - залишатися позитивним і оптимістичним перед обличчям труднощів або негараздів
Leave no stone unturned - докласти всіх зусиль для досягнення мети або знайти рішення
Money talks - влада або вплив грошей на прийняття рішень або переговори
On the ball - пильний, уважний, ефективний
Pull yourself together - відновити самовладання або контроль після важкого або емоційного досвіду
Saved by the bell - ледве уникнути важкої або неприємної ситуації в останню хвилину
The whole nine yards - викластися на повну або зробити щось на повну
Up in arms - бути дуже злим або засмученим через щось
Vanishing point - точка на відстані, де об'єкти ніби зливаються в одну лінію або зовсім зникають
Walking on eggshells - бути дуже обережним і обережним у своїх діях або словах, щоб не засмутити когось

X marks the spot - фраза, яка використовується для позначення місця розташування чогось важливого чи цінного
You can lead a horse to water, but you can't make it drink - ви можете дати комусь можливість або пораду, але ви не можете змусити їх прийняти це
Zero hour - точний момент, коли очікується, що станеться щось важливе або значуще.

ДОДАТОК Е

АВТОМАТИЧНА КЛАСИФІКАЦІЯ ФО НА ОСНОВІ ЗАСТОСУВАННЯ ГІБРИДНОГО ПЗ HYBRID SOFT

Ознаки	**Фразеологічні одиниці**
Ідіоми (27)	A bird in the hand is worth two in the bush, A dime a dozen, A piece of cake, All ears, Barking up the wrong tree, Bite the bullet, Break a leg, Cut the mustard, Devil's advocate, Don't count your chickens before they hatch, Drop a dime, Face the music, Fit as a fiddle, Go the extra mile, Hit the hay, Kick the bucket, Let the cat out of the bag, Piece of the action, Pull someone's leg, Rule of thumb, Silver lining, Take a rain check, Take the bull by the horns, Under the weather, When pigs fly, Wild goose chase, You can't judge a book by its cover
Формульні вирази (9)	How do you do?, It's raining cats and dogs, The whole nine yards, Let's call it a day, The early bird catches the worm, Break the ice, A stitch in time saves nine, It takes two to tango, Beat around the bush

Класифікація за походженням (6)	Ad hoc, Catharsis, Avant-garde, Coup d'état, Feng shui, Yin and yang
Ад’єктивні (19)	blue-blooded, high-tech, fair-haired, two-faced, easygoing, hardworking, big-hearted, well-off, narrow-minded, long-lasting, one-of-a-kind, old-fashioned, fast-paced, top-notch, high-spirited, short-tempered, well-known, easy-to-use, open-minded
Субстантивні (16)	the pot calling the kettle black, a safe bet, a skeleton in the closet, a square peg in a round hole, the straw that broke the camel's back, a swan song, the tail wagling the dog, the tip of the iceberg, a tough cookie, the milk of human kindness, a money pit, a loose cannon, a nest egg, the new kid on the block, a pain in the neck, a penny for your thoughts
Фразеологічні одиниці пов'язані з тваринами (10)	Busy as a bee, Hold your horses, You can lead a horse to water, but you can't make it drink, A bird in the hand is worth two in the bush, Let the cat out of the bag, When pigs fly, Wild goose chase, It's raining cats and dogs, The early bird catches the worm, the tail wagling the dog, a swan song

Фразеологічні одиниці пов'язані з тілом (9)	All thumbs, Egg on your face, Fly off the handle, Get cold feet, Keep your chin up, All ears, Break a leg, Pull someone's leg, Up in arms
Не увійшли у виділені ознаки класифікації (14)	Catch-22, Diamond in the rough, In the same boat, Jump the gun, Leave no stone unturned, Money talks, On the ball, Pull yourself together, Saved by the bell, The whole nine yards, Vanishing point, Walking on eggshells, X marks the spot, Zero hour

Фразеологічна одиниця	**Пояснення**
the pot calling the kettle black	хтось звинувачує іншу людину в чомусь, у чому сам винен.
A bird in the hand is worth two in the bush	краще триматися за те, що маєш, ніж ризикувати втратити це, намагаючись отримати щось краще
a safe bet	те, що, ймовірно, станеться або вдасться
How do you do?	Офіційне привітання, яке використовується для демонстрації ввічливості.
a skeleton in the closet	таємниця або незручний факт, який хтось хоче зберегти прихованим.
Ad hoc	латинська фраза означає «для цієї мети» і використовується для опису чогось, що створюється або робиться спеціально для конкретної ситуації чи проблеми, а не є частиною загального плану.

A dime a dozen	дуже поширене або легкодоступне
the spice of life	різноманітність і хвилювання в житті
It's raining cats and dogs	спосіб описати сильний дощ.
a square peg in a round hole	людина, яка не підходить для певної ситуації.
All thumbs	незграбні руки
the straw that broke the camel's back	остання, здавалося б, незначна подія, що спричиняє велику проблему або невдачу.
a swan song	остаточне виконання або досягнення перед відставкою або смертю
A piece of cake	те, що дуже легко зробити
the tail wagling the dog	ситуація, в якій незначний або незначний фактор має непропорційний вплив на більшу систему або процес.
Busy as a bee	дуже працьовитий або працьовитий
the tip of the iceberg	невелика або видима частина набагато більшої проблеми або питання.
Catch-22	ситуація, в якій людина потрапляє в пастку суперечливих або парадоксальних правил або обставин
All ears	повністю слухають і звертають увагу
a tough cookie	сильна або витривала людина
the milk of human kindness	добра і співчутлива поведінка

Diamond in the rough	хтось або щось з великим потенціалом або талантом, але потребує доопрацювання або розвитку
a money pit	проект або підприємство, яке постійно потребує більше грошей, ніж створює.
a loose cannon	людина непередбачувана і некерована, часто завдає збитків або неприємностей.
Barking up the wrong tree	дотримуватися помилкового або неправильного курсу дій
a nest egg	сума грошей, відкладена на майбутнє
Egg on your face	бути збентеженим або приниженим власними діями або помилками
Yin and yang	Ця китайська концепція стосується взаємодоповнюючих і протилежних сил Всесвіту, часто представлених у вигляді двох половин кола або символу.
Fly off the handle	стати раптово і нестримно розлюченим
Bite the bullet	погодитеся з чимось важким або неприємним
Get cold feet	нервувати або боятися рішення чи способу дії
the new kid on the block	нова особа чи річ у певній сфері чи місці.
It takes two to tango	фраза, яка використовується для припущення, що обидві сторони несуть відповідальність за ситуацію чи проблему.

old-fashioned	застарілий або не відповідає сучасній моді чи стилю
Hold your horses	почекати або проявити терпіння, перш ніж приступити до чогось
a pain in the neck	хтось або щось дратує або дратує
Break a leg	удача (часто використовується в контексті сценічного мистецтва)
a penny for your thoughts	прохання, щоб хтось поділився своїми думками або почуттями.
In the same boat	в тій же складній або складній ситуації, що і хтось інший
A stitch in time saves nine	фраза, яка використовується для припущення, що вжиття профілактичних заходів зараз може заощадити більше часу та зусиль пізніше.
Jump the gun	діяти занадто швидко або передчасно
Cut the mustard	до вдалого виконання або виправдання очікувань
Keep your chin up	залишатися позитивним і оптимістичним перед обличчям труднощів або негараздів
fast-paced	відбувається швидко і з великою енергією
Leave no stone unturned	докласти всіх зусиль для досягнення мети або знайти рішення

Devil's advocate	той, хто займає позицію, з якою він не обов'язково погоджується, заради аргументації чи дослідження процесу прийняття рішення.
Money talks	влада або вплив грошей на прийняття рішень або переговори
Don't count your chickens before they hatch	не припускайте, що щось трапиться, перш ніж це станеться
Drop a dime	подзвонити
On the ball	пильний, уважний, ефективний
well-off	мати фінансову забезпеченість або багатство
Pull yourself together	відновити самовладання або контроль після важкого або емоційного досвіду
top-notch	найвищої якості або стандарту
Face the music	змиритися з неприємними наслідками своїх дій
Saved by the bell	ледве уникнути важкої або неприємної ситуації в останню хвилину
Fit as a fiddle	бути в хорошому фізичному стані
The whole nine yards	викластися на повну або зробити щось на повну
Feng shui	ця китайська фраза відноситься до системи дизайну та облаштування, заснованої на принципах гармонії та

	балансу, яка часто використовується в архітектурі та дизайні інтер'єру.
Up in arms	бути дуже злим або засмученим через щось
long-lasting	тривалий або триваючий протягом тривалого часу
Vanishing point	точка на відстані, де об'єкти ніби зливаються в одну лінію або зовсім зникають
Go the extra mile	докласти додаткових зусиль або вийти за межі очікуваного
Walking on eggshells	бути дуже обережним і обережним у своїх діях або словах, щоб не засмутити когось
one-of-a-kind	унікальний, винятковий і ні на що не схожий
Hit the hay	лягти спати
Kick the bucket	померти
X marks the spot	фраза, яка використовується для позначення місця розташування чогось важливого чи цінного
Beat around the bush	фраза, яка використовується для опису людини, яка уникає переходу на суть або ухиляється.
big-hearted	щедрий, добрий і турботливий

Let the cat out of the bag	передчасно розкрити таємницю або здивувати
You can lead a horse to water, but you can't make it drink	ви можете дати комусь можливість або пораду, але ви не можете змусити їх прийняти це
Piece of the action	частка або частина чого-небудь
Break the ice	фраза, яка використовується для опису процесу початку розмови з кимось, хто незнайомий або стриманий.
Coup d'état	ФО французького походження, ця фраза відноситься до раптового і часто насильницького повалення уряду або правлячої влади.
Pull someone's leg	дражнити або жартувати з ким-небудь добродушно
easygoing	розслаблений і толерантний, не легко засмучуватися або піддаватися стресу
Rule of thumb	загальна настанова чи правило, засноване на досвіді чи практиці
Zero hour	точний момент, коли очікується, що станеться щось важливе або значуще.
The early bird catches the worm	Фраза, яка використовується для припущення, що найбільше виграють ті, хто діє на початку ситуації.
Silver lining	позитивний аспект у важкій або неприємній ситуації

hardworking	вкладати багато зусиль і часу в завдання
Take a rain check	відхиліть запрошення, але запропонуйте прийняти його пізніше
Avant-garde	це французьке словосполучення означає «передовий захисник» і використовується для опису інноваційних або експериментальних ідей або практик, особливо в мистецтві.
Take the bull by the horns	зіткнутися з проблемою лоб в лоб
Under the weather	погане самопочуття
Let's call it a day	фраза, яка використовується для підказки того, що настав час припинити працювати чи щось робити.
When pigs fly	такого ніколи не буде
high-tech	передбачають передові технології
Wild goose chase	безнадійна або марна гонитва
narrow-minded	небажання розглядати нові ідеї чи думки
Catharsis	це грецьке слово означає очищення або звільнення від емоцій, часто за допомогою певної форми мистецтва чи вистави.
The whole nine yards	фраза, яка використовується для вказівки повного обсягу чогось.
You can't judge a book by its cover	ви не можете судити про щось лише за зовнішнім виглядом

easy-to-use	простий або зрозумілий в експлуатації або поводженні
blue-blooded	мають дворянське або аристократичне походження
well-known	відомий або знайомий багатьом людям
open-minded	сприйнятливий до нових ідей і готовий враховувати різні думки
fair-haired	прихильність або перевага; також відноситься до світлого кольору волосся
two-faced	нечесний або брехливий; говорити одне, а робити інше
high-spirited	жваві, енергійні та повні ентузіазму
short-tempered	легко піддається провокаціям або роздратуванням, має запальний і інтенсивний характер

ДОДАТОК Є

ПРИКЛАД ФРАГМЕНТУ КОДУ, ЯКИЙ ДЕМОНСТРУЄ, ЯК ВИКОРИСТОВУВАТИ МОДУЛЬ ОЦІНКИ В PYTHON

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
import pandas as pd

# Load dataset
df = pd.read_csv('dataset.csv')

# Split dataset into training and validation sets
X = df.drop('class', axis=1)
y = df['class']
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

# Train decision tree classifier
clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)

# Evaluate performance on validation set
y_pred = clf.predict(X_val)
accuracy = accuracy_score(y_val, y_pred)
precision = precision_score(y_val, y_pred, average='macro')
recall = recall_score(y_val, y_pred, average='macro')
f1 = f1_score(y_val, y_pred, average='macro')

print('Accuracy:', accuracy)
print('Precision:', precision)
print('Recall:', recall)
print('F1 score:', f1)
```

ДОДАТОК Ж

ПРИКЛАД ФРАГМЕНТУ КОДУ, ЯКИЙ ДЕМОНСТРУЄ, ЯК ВИКОРИСТОВУВАТИ МОДУЛЬ ОПТИМІЗАЦІЇ В PYTHON

```
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
import pandas as pd

# Load dataset
df = pd.read_csv('dataset.csv')

# Split dataset into training and validation sets
X = df.drop('class', axis=1)
y = df['class']
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

# Define decision tree classifier and parameter grid
clf = DecisionTreeClassifier()
param_grid = {'max_depth': [1, 5, 10, 15, 20]}

# Perform grid search optimization
grid_search = GridSearchCV(clf, param_grid=param_grid, cv=5)
grid_search.fit(X_train, y_train)

# Evaluate performance on validation set
y_pred = grid_search.predict(X_val)
accuracy = accuracy_score(y_val, y_pred)

print('Accuracy:', accuracy)
print('Best parameters:', grid_search.best_params_)
```

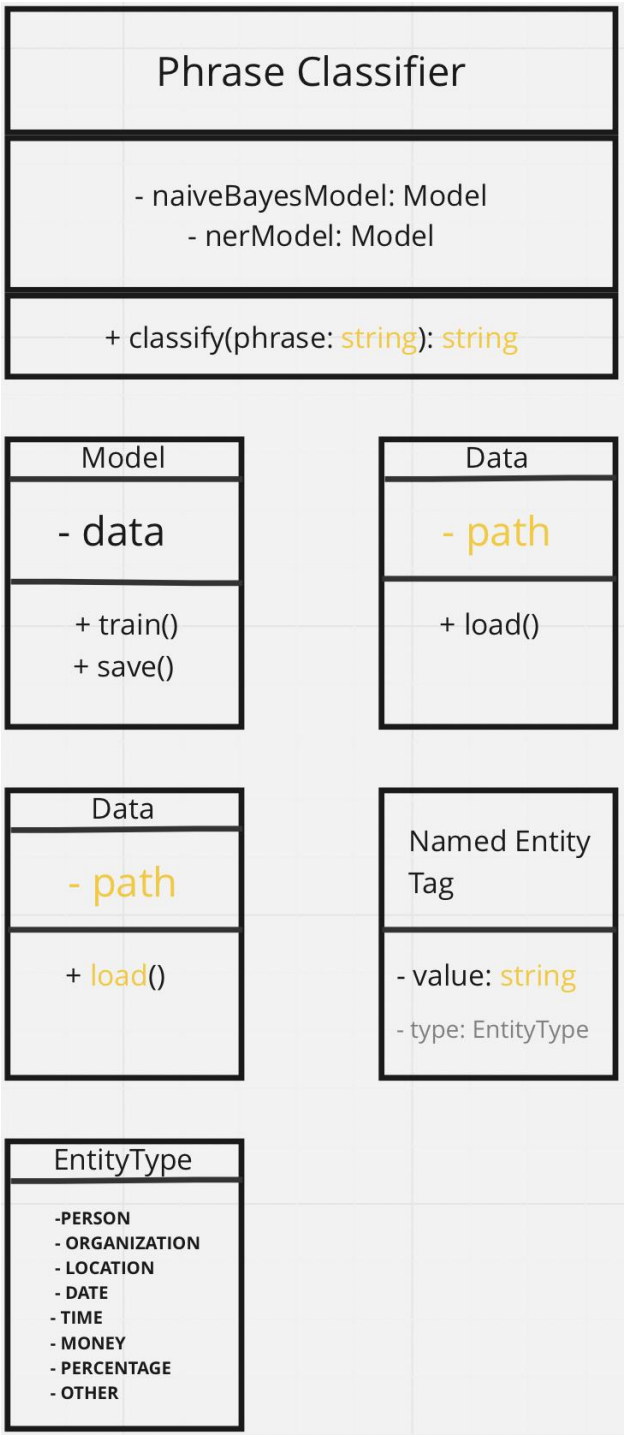
ДОДАТОК З

ПРИКЛАД РЕАЛІЗАЦІЇ ГІБРИДНОГО АЛГОРИТМУ З ВИКОРИСТАННЯМ АЛГОРИТМУ NER І АЛГОРИТМУ NAIVE BAYES

```
# Import the necessary libraries and models
import spacy
from sklearn.naive_bayes import MultinomialNB

# Load the English language model for Named Entity Recognition
nlp = spacy.load("en_core_web_sm")

# Train a Naive Bayes classifier on a dataset of labeled phraseological units
classifier = MultinomialNB()
X_train = ["kick the bucket", "break a leg", "barking up the wrong tree"]
y_train = ["idiom", "idiom", "literal"]
X_train_transformed = [] # Convert phraseological units to numerical features
# Fit the classifier on the transformed features and labels
classifier.fit(X_train_transformed, y_train)

# Define a function to classify phraseological units using Named Entity Recognition and the trained classifier
def classify_phraseological_unit(phraseological_unit):
    doc = nlp(phraseological_unit)
    if doc.ents: # Check if any named entities are detected in the phraseological unit
        return "literal" # If so, classify as literal
    else:
        features = [] # Extract features from phraseological unit
        features_transformed = [] # Convert features to numerical form
        return classifier.predict(features_transformed) # Classify using the trained classifier

# Example usage
print(classify_phraseological_unit("I kicked the bucket yesterday"))
# Output: literal
print(classify_phraseological_unit("She's barking up the wrong tree"))
# Output: idiom
```

ДОДАТОК И

UML ДІАГРАМА, ЯКА ПОКАЗУЄ ОСНОВНІ КОМПОНЕНТИ ТА ЇХ ЗВ'ЯЗКИ В ГІБРИДНОМУ АЛГОРИТМІ



UNICHECK

Ім'я користувача:
Кафедра КІ

ID перевірки:
1015123934

Дата перевірки:
17.05.2023 08:26:46 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
17.05.2023 08:27:25 EEST

ID користувача:
100005591

Назва документа: **Зінюк_Метод автоматичної класифікації фразеологічних одиниць англомовних текстів**

Кількість сторінок: **114** Кількість слів: **21365** Кількість символів: **164981** Розмір файлу: **4.26 MB** ID файлу: **1014806006**

6.1%
Схожість

Найбільша схожість: **0.56%** з Інтернет-джерелом (http://ddpu-filolvisnyk.com.ua/uploads/arkhiv-nomerov/2021/NV_2021.

5.77% Джерела з Інтернету	865	Сторінка 116
0.8% Джерела з Бібліотеки	68	Сторінка 124

0.55% Цитат

Цитати	10	Сторінка 125

Не знайдено жодних посилань

0%
Вилучень

Немає вилучених джерел

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: 113476 Назва: МКР Метод автоматичної класифікації фразеологічних одиниць англомовних текстів Додано в БД: 2023-05-17 Автора: Зінюк Є.Р. Керівники: Боровик О.В. Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	153620	1129	4419 (3%)	40 (4%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ

РЕЦЕНЗІЯ НА ДИПЛОМНУ РОБОТУ

Магістр Онишко Оксана Григорівна

Тема Метод та програмні засоби препроцесингу вхідного текстового контенту

Спеціальність 121 – Інженерія програмного забезпечення

Обсяг дипломної роботи:

Кількість листів креслень 0; кількість сторінок записки 76

1.Короткий зміст ДР та прийнятих рішень У дипломній роботі удосконалено метод автоматичного визначення рольових структур висловлень, заснованих на комунікативній граматиці української мови. Удосконалено метод комп'ютерного семантико-синтаксичного аналізу текстів, в який інтегровані методи побудови синтаксичних дерев залежностей і визначення рольових структур висловлень. Застосування розробленого методу надає можливість покращити точність та повноту синтаксичного та семантичного аналізу.
На основі розроблених методів було створено програмний засіб препроцесингу вхідного текстового контенту.

2. Висновок про відповідність ДР поставленому завданню Дипломна робота освітнього ступеня «магістр» у достатній мірі відповідає поставленому завданню як у теоретичній , так і в практичній її частині

3. Характеристика виконання кожного розділу роботи, ступінь використання останніх досягнень науки і техніки і передових методів роботи:
Розділ 1- Проведено огляд відомих методів синтаксичного та семантичного аналізу та визначено їх недоліки. Розділ 2 – Розроблено метод визначення рольових структур тверджень у текстах українською мовою та метод семантико-синтаксичного аналізу текстів . Розділ 3 – Проведенно дослідження розроблених методів семантико-синтаксичного аналізу тексту та визначення рольових структур тверджень. Розділ 4 – На основі розроблених методів створено програмний засіб препроцесингу вхідного текстового контенту

4.Позитивні сторони роботи Застосування розробленого програмного засобу препроцесингу вхідного текстового контенту надає можливість підвищення якості вирішення задачі визначення рольових структур тверджень. Зокрема, надає можливість підвищити повноту на 1,7%, що призводить до збільшення F_1–міри на 1,2%. При цьому точність не зменшується

5. Негативні сторони роботи Недостатньо детально в роботі прописана програмна реалізація

6. Оцінка графічного оформлення та пояснювальної записки роботи
Оформлення роботи відповідає всім необхідним нормам та вимогогам. Викладення матеріалу є логічним. Послідовним та стилістично грамотним

7. Відгук про роботу в цілому
Кваліфікаційна робота магістранта Онишко Оксани Григорівни рекомендується до захисту.

8. Інші зауваження

9. Оцінка дипломної роботи Робота заслуговує на оцінку «добре»

РЕЦЕНЗЕНТ (прізвище, ім'я, по-батькові, посада, місце роботи)
доц. кафедри комп'ютерної інженерії та кібербезпеки
Бобровський К.Ю.

"6" 12 2021 р. (підпис)

Завідувачу кафедри КІІС
д-р.техн.наук, проф. Говорущенко Т. О.

Зінюк Євгеній Ростиславович

ПІБ здобувача вищої освіти

ФІТ, 2 курсу, групи КІ2м-21-1

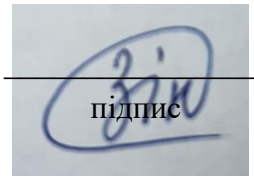
ЗАЯВА

З правилами чинного Положення «Про дотримання академічної доброчесності в Хмельницькому національному університеті» від 26.09.2020 (зі змінами від 26.11.2020), згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування заходів дисциплінарної та академічної відповідальності, ознайомлений (а). Про використання програмно-технічних засобів для перевірки кваліфікаційних робіт здобувачів вищої освіти на плагіатоповіщений (а) та надаю свою згоду на обробку та збереження університетом моєї роботи в інституційному репозитарії університету.

Також надаю університету право на передачу моєї роботи для обробки та збереження в базах даних програмно-технічних засобів (Unicheck та Anti-Plagiarism) та використання роботи для виявлення плагіату в інших роботах, які перевіряються програмно-технічними засобами та користувачами, що мають доступ до цих програмно-технічних засобів, виключно в обмежених цілях для виявлення плагіату в текстах робіт.

Робота для перевірки університетом надається в друкованому та електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

18.05.2023
дата


підпис

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ

КАФЕДРИ КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОМАЦІЙНИХ СИСТЕМ

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

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

Назва: Метод автоматичної класифікації фразеологічних одиниць англомовних текстів

Автор: Зінюк Євгеній Ростиславович

Спеціальність: 123 – Компютерна інженерія

Освітня програма: освітньо-наукова

Науковий керівник: Боровик О.В. , д.т.н, професор

Після аналізу звіту подібності зроблено такий висновок:

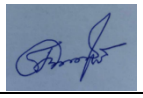
№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий 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.	

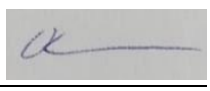
Підтвердження:

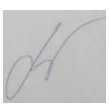
Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) усі запозичення фрагментарні, або мають належним чином оформленні посилання;
2) окремі виявлені збіги є загальновживаними фразами або виразами, про що свідчить посилання системи на збіг з джерелами на один фрагмент речення;
3) всі зафіксовані системою ознаки модифікації тексту відносяться до комбінування латинських символів зі україномовними скороченнями індексів в формулах, що не є модифікацією тексту.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості Unicheck, складає 6.1% і адресується до 933 першоджерела; та системою Anti-Plagiarism складає 2%, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи  О.В. Боровик

Гарант ОП  О. С. Савенко

Завідувач кафедри КІІС  Т. О. Говорущенко