Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації
Кафедра системного аналізу та інформаційних технологій

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему:

“Інформаційна NLP-технологія класифікації англомовної інформації про стан водних ресурсів”

Виконав: студент 2 курсу, групи 2ІСТ-21м
спеціальності 126 – «Інформаційні системи та технології»
_____ Радецький О.В.

Керівник: д.т.н., проф. каф. САІТ
_____ Мокін В. Б.
«01» 12 2022 р.

Опонент: к.т.н., доц. каф. АІІТ
_____ Сторчак В. Г.
«15» 12 2022 р.

Допущено до захисту

Завідувач кафедри САІТ
_____ д.т.н., проф. Мокін В. Б.
«05» 12 2022 р.

Вінниця ВНТУ – 2022 рік

Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації
Кафедра системного аналізу та інформаційних технологій
Рівень вищої освіти – ІІ-й (магістерський)
Галузь знань – 12 Інформаційні технології
Спеціальність – 126 Інформаційні системи та технології
Освітньо-професійна програма – Інформаційні технології аналізу даних та зображень

ЗАТВЕРДЖУЮ
Завідувач кафедри САІТ
_____ д.т.н., проф. Мокін В. Б.
«16» ____09____ 2022 р.

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
Радецькому Олександру Володимировичу

1. Тема роботи: «Інформаційна NLP-технологія класифікації англомовної інформації про стан водних ресурсів»
керівник роботи: Мокін В. Б., д-р техн. наук, професор. каф. САІТ,
затверджені наказом закладу вищої освіти від «14» ____09____2022 року №203
2. Строк подання студентом роботи «01» ____12____2022 року
3. Вихідні дані до роботи: Kaggle Dataset «NLP : Reports & News Classification» https://www.kaggle.com/datasets/vbmokin/nlp-reports-news-classification
4. Зміст розрахунково-пояснювальної записки:
   - аналіз проблеми автоматичної класифікації україномовної екологічної інформації;
   - вибір оптимальних налаштувань інформаційної технології для розв'язання поставленої задачі;
   - проєктування та створення інтелектуальної інформаційної технології класифікації україномовної текстової інформації про масиви вод;
   - економічна частина.

5. Перелік графічного матеріалу:
   - алгоритм розробки моделі класифікаторів;
   - алгоритм інформаційної технології;
   - сформований датасет для тестування моделей;
   - результат роботи відібраних моделей;
   - класифікація за допомогою моделі лінійної регресії;
   - класифікація за допомогою моделі випадкових лісів (Random Forest).

6. Консультанти розділів МКР

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв
4	Бурєннікова Н.В., д.е.н., проф. каф. ЕПВМ	01.11.2022	30.11.22

7. Дата видачі завдання «16» 09 2022 року

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів МКР	Строк виконання етапів роботи	Примітка
1	Характеристика проблеми класифікації англомовної інформації про стан водних ресурсів	09.2022	
2	Вибір оптимальних налаштувань інформаційної технології для розв`язання поставленої задачі	10.2022	
3	Створення інтелектуальної інформаційна технології класифікації англомовної інформації про масиви вод басейну річки Південний Буг	11.2022	
4	Економічна частина	11.2022	
5	Оформлення матеріалів до захисту МКР	11.2022	

Студент _____ Радецький О.В.

Керівник роботи _____ Мокін В.Б.

АНОТАЦІЯ

УДК 004.8:338+332.628

Радецький О.В. Інформаційна NLP-технологія класифікації англомовної інформації про стан водних ресурсів. Магістерська кваліфікаційна робота зі спеціальності 126 – інформаційні системи та технології, освітньо-професійна програма – інформаційні технології аналізу даних та зображень. Вінниця: ВНТУ, 2022. 84 с.

На укр. мові. Бібліогр.: 21 назв; рис.: 62; табл.: 5

В магістерській кваліфікаційній роботі проведено аналіз предметної області класифікації текстової англомовної інформації. Здійснено огляд аналогічних рішень, запропоновано алгоритм створення ІТ класифікації текстової англомовної інформації, на основі якого проведено розроблення ІТ. Виконано пошук необхідних текстових даних, проведено попереднє очищення й сформовано датасет. Проведено розвідувальний аналіз даних, обрано регресійні моделі машинного навчання, здійснено їх тренування та визначено оптимальну модель серед них, точність передбачення якої – 0.95, що є більшим за 0.80, як у найкращого аналога.

Ілюстративна частина складається з 6 плакатів.

У розділі економічної частини розглянуто питання про доцільність розроблення та впровадження інформаційної NLP-технології класифікації англомовної інформації про стан водних ресурсів.

Ключові слова: інформаційна технологія, розвідувальний аналіз даних, класифікація тексту, водні ресурси, ознаки, моделі машинного навчання.

ABSTRACT

UDC 004.8:338+332.628

Radetsky O.V. Informational NLP technology for the classification of English-language information about the state of water resources. Master's qualification thesis on specialty 126 - information systems and technologies, educational and professional program - information technologies of data and image analysis. Vinnytsia: VNTU, 2022. 84 p.

In Ukrainian speech Bibliography: 21 titles; Fig.: 62; tab.: 5

In the master's qualification work, an analysis of the subject area of the classification of textual English-language information was carried out. A review of similar solutions was carried out, an algorithm for creating IT classification of English-language textual information was proposed, on the basis of which IT was developed. The search for the necessary text data was performed, preliminary cleaning was carried out, and the dataset was formed. Exploratory data analysis was carried out, machine learning regression models were selected, their training was carried out, and the optimal model among them was determined, the prediction accuracy of which is 0.95, which is higher than 0.80, as the best analogue.

The illustrative part consists of 6 posters.

In the section of the economic part, the issue of the feasibility of developing and implementing NLP information technology for the classification of English-language information about the state of water resources is considered.

Keywords: information technology, data intelligence analysis, text classification, water resources, features, machine learning models.

ЗМІСТ

ВСТУП

Актуальність теми. Через те, що люди з кожним роком все більше і більше забруднюють навколишнє середовище актуальність пошуку рішень покращення його стану не викликає питань. Під згубний вплив потрапляє все живе і всі ресурси. Детальніше зупинимось на воді, оскільки вона є одним із основних, для того щоб життя на нашій планеті не завершилось.

Для того, щоб утримувати контроль над станом водних ресурсів та покращувати їх якість, потрібно знати в якому стані вони знаходяться. Щоб це реалізувати, необхідно провести аналіз доступної інформації, яка описує стан води або ж має відношення до нього.

На даний момент є багато інформації, яка потенційно може бути корисною та містити необхідні дані. Але в цей же час виникає інше питання, як саме проаналізувати та класифікувати всю цю інформацію, адже якщо робити це старими методами, тобто вручну, це може зайняти дуже багато часу, і такий аналіз може містити помилки через людський фактор. Подібні технології вже існують, але для інших задач. Зокрема, не має адаптацій до класифікації інформації англомовних джерел про масиви вод басейну річки Південний Буг. Отже, розробка є актуальною.

Мета і задачі дослідження. Метою дослідження є підвищення точності класифікації текстової англомовної інформації про масиви вод басейну річки Південний Буг шляхом створення відповідної інтелектуальної інформаційної технології.

Розробка технології передбачає виконання наступних задач:

- огляд і вибір оптимальних інформаційних технологій;
- створення датасету та вибір моделей;
- вибір оптимальної моделі, створення інформаційної технології та її застосування для класифікації даних.

Об’єктом дослідження магістерської кваліфікаційної роботи є класифікація текстової англомовної інформації про масиви вод басейну р. Південний Буг.

Предметом дослідження магістерської кваліфікаційної роботи є інтелектуальна інформаційна технологія класифікації текстової англомовної інформації про масиви вод басейну р. Південний Буг.

Методи дослідження. У дослідженнях використано методи розвідувального аналізу даних та методи машинного навчання для побудови моделей та методи оброблення природномовних текстів на Python.

Новизна одержаних результатів. Наукова новизна полягає у тому, що дістала подальший розвиток інформаційна технологія класифікації текстової англомовної інформації про масиви вод басейну річки Південний Буг з використанням методів машинного навчання та NLP-технологій, яка дозволяє підвищити точність такої класифікації у порівнянні з аналогами.

На основі теоретичних результатів розроблено і реалізовано програмне забезпечення інтелектуальної інформаційної технології класифікації текстової англомовної інформації про масиви вод басейну р. Південний Буг. Результати впроваджено у діяльність басейнової ради Південного Бугу та у навчальний процес кафедри системного аналізу та інформаційних технологій ВНТУ.

Апробація результатів магістерської кваліфікаційної роботи. Результати роботи доповідались на Всеукраїнській науково-практичній інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи» (Вінниця, 2021-2022 рр.).

Публікація результатів магістерської кваліфікаційної роботи.

Опубліковано тези на Всеукраїнській науково-практичній інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи» (Вінниця, 2021-2022 рр.) [1].

1 ХАРАКТЕРИСТИКА ПРОБЛЕМИ КЛАСИФІКАЦІЇ АНГЛОМОВНОЇ ІНФОРМАЦІЇ ПРО СТАН ВОДНИХ РЕСУРСІВ

1.1 Суть проблеми

На даний момент пошук необхідної інформації уже не є проблемою, адже інформаційні технології як ніколи на висоті, але кількість інформації не означає, що вся вона достовірна і актуальна, тому для того, щоб знайти щось дійсно корисне, потрібно опрацювати великі об’єми даних. Це завдання можна реалізувати двома способами, перший – це зробити це вручну, що не є оптимальним варіантом, адже ручна праця це – дуже довгий процес, до того ж вона містить високий рівень помилок через людський фактор; другий – автоматизувати цей процес, що зекономить багато часу та інших ресурсів, створивши інтелектуальну інформаційну систему.

Передусім, слід автоматизувати процес видобування інформації з веб-ресурсів та різних текстових документів. Для цього існують спеціальні Python-бібліотеки та технології.

Щоб система могла функціонувати, дані з якими вона працюватиме, потрібно заздалегідь опрацювати та звести до потрібного формату. Для вирішення цього завдання необхідно створити парсер, який буде виконувати вищевказані перетворення. Також необхідно залучити спеціально навчених експертів, для того щоб зробити правильну розмітку.

Наступним кроком є формування датасету із відповідним набором відформатованих даних для навчання моделі, яку потім можна буде застосувати в подальшому для розв’язання поставленої задачі.

Потім за допомогою технології NLP класифікуємо інформацію, але перед цим модель потрібно адаптувати спеціально до потреб даного завдання.

1.2 Огляд технології NLP як методу вирішення проблеми класифікації

Обробка природної мови (з англ.: Natural Language Processing (NLP)) дозволяє машинам розбивати та інтерпретувати людську мову. Вона є основою інструментів, які використовуються кожного дня – починаючи з програмного забезпечення для перекладу, чат-ботів, фільтрів спаму та пошукових систем закінчуючи програмним забезпеченням для виправлення граматики, голосових помічників та інструментів моніторингу соціальних мереж.

NLP використовується для розуміння структури та значення людської мови шляхом аналізу різних аспектів, таких як синтаксис, семантика, прагматика та морфологія. Потім технологія перетворює ці лінгвістичні знання в алгоритми машинного навчання, що можуть вирішувати конкретні проблеми та виконувати бажані завдання.

Візьмемо, наприклад, Gmail, в якому електронні листи автоматично класифікуються як рекламні , соціальні , основні або спам завдяки завданню NLP, яке називається вилучення ключових слів . «Зчитуючи» слова в темах і пов’язуючи їх із заздалегідь визначеними тегами, машини автоматично дізнаються, яку категорію призначати електронним листам [2].

Технологія NLP має багато переваг, найважливіші з них:

– Виконання широкомасштабного аналізу: обробка природньої мови дає можливість системам розуміти й аналізувати практично будь-які обсяги неструктурованих текстових даних, таких як коментарі, телефонні розмови, чи новини;

– Автоматизація процесів в режимі реального часу: інструменти обробки природньої вони є практично автономними, тобто вони можуть дати можливість машинам сортувати та маршрутизувати інформацію практично без участі людини – швидко, ефективно, точно та цілодобово;

– Адаптація інструментів NLP до конкретної задачі: алгоритми обробки природньої мови можна застосувати до необхідних потреб та критеріїв.

З використанням векторизації тексту , інструменти NLP перетворюють текст до такого виду, який машина може зрозуміти, далі алгоритми машинного навчання подають навчальні дані та очікувані результати (теги), щоб навчити машини створювати асоціації між конкретним входом і його відповідним виходом. Після цього машини використовують методи статистичного аналізу, щоб створити власний «багаж знань» і визначити, які характеристики найкраще представляють тексти, перш ніж робити прогнози для невидимих даних (нових текстів) (рис. 1.1) [2].

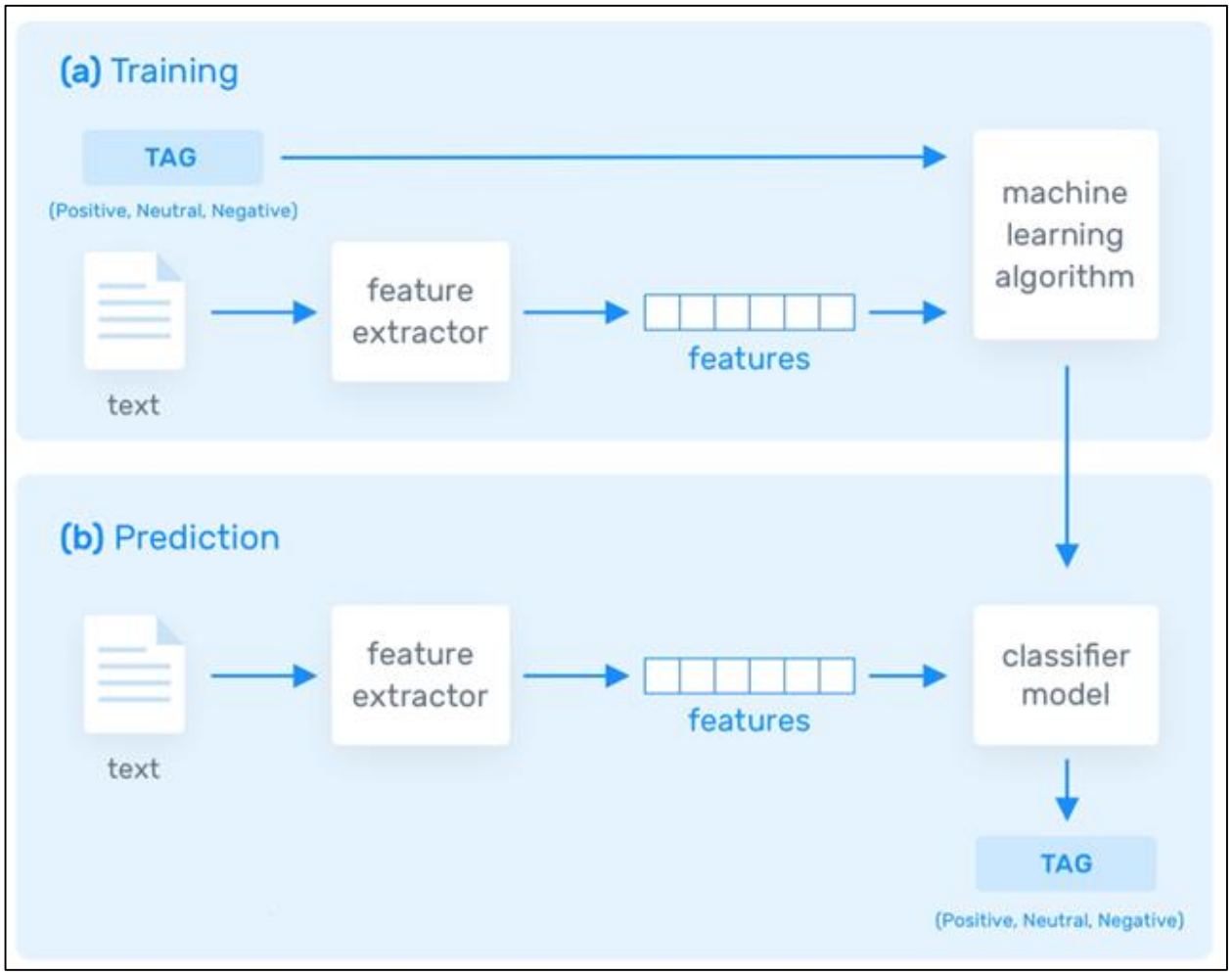

Рисунок 1.1 – Алгоритм роботи з класифікацією текстових даних

Отже, чим більше даних надсилають ці алгоритми NLP, тим більш точними будуть моделі аналізу тексту.

Одним із найпопулярніших завдань NLP є аналіз настроїв (показано на діаграмі вище) є, де моделі машинного навчання навчаються класифікувати текст за полярністю думок (позитивні, негативні, нейтральні та будь-які інші).

Однією з найбільших переваг моделей машинного навчання є їх здатність навчатися самостійно, без необхідності визначати правила вручну. Для цього просто потрібен набір відповідних навчальних даних із кількома прикладами для тегів, які потрібно проаналізувати. А за допомогою розширених алгоритмів глибокого навчання є можливість об’єднати кілька завдань обробки природної мови, як-от аналіз настроїв, виділення ключових слів, класифікація тем, виявлення намірів тощо, щоб працювати одночасно для надточних результатів.

Більшість завдань обробки природньої мови включають в себе синтаксичний і семантичний аналіз, що використовується для розбиття людської мови на частини, які машина зможе зчитати.

Синтаксичний аналіз визначає синтаксичну структуру тексту та зв’язки залежності між словами, представлені на діаграмі, яка називається деревом синтаксичного аналізу.

Семантичний аналіз фокусується на виявленні значення мови. Однак, оскільки мова багатозначна та неоднозначна, семантика вважається однією з найскладніших областей NLP.

Семантичні завдання аналізують структуру речень, взаємодію слів і пов’язані з ними поняття, намагаючись відкрити значення слів, а також зрозуміти тему тексту.

Далі розберемо деякі з основних підзавдань як семантичного, так і синтаксичного аналізу.

Токенізація є важливою задачею в обробці природної мови, яка використовується для розбиття рядка слів на семантично корисні одиниці, які називаються лексемами [2].

Токенізація речень розділяє речення в тексті, а токенізація слів розділяє слова в реченні. Як правило, лексеми слів відокремлюються пробілами, а лексеми речень – знаками. Однак ви можете виконати високорівневу токенізацію для більш складних структур, як-от слова, які часто йдуть разом, інакше відомі як сполучення (наприклад, Нью-Йорк). Принцип дії процесу токенізації розглянуто на рисунку 1.2. [3].

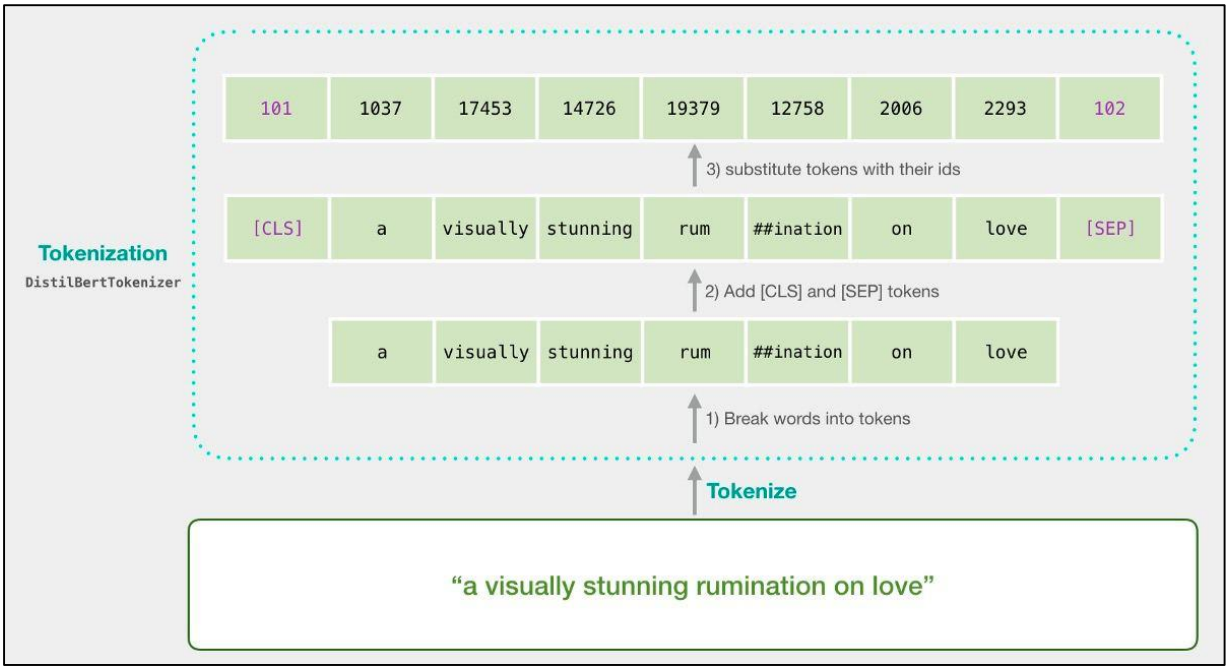

Рисунок 1.2 – Процес токенізації

Відмічання частини мови – ще один з етапів обробки тексту, який заключається в призначення кожній лексемі в тексті належної мітки частини мови, тобто синтаксичної функції слова(іменником, дієсловом, прикметником, прислівником тощо.)

Нормалізація – процес зведення подібних лексем до канонічної форми. наприклад, якщо вважатимемо, що Shop та shop мають однакове значення, то можемо нормалізувати текст, звівши обидва терміни у shop.

Також важливим етапом, при підготовці тексту до обробки моделями машинного навчання, це відфільтрування стоп-слів. Ці слова відфільтровуються перед будь-яким завданням попередньої обробки або

моделювання. Стоп-слова обираються з огляду на їх незначущість для поточного завдання NLP. Наприклад, список англійських стоп-слів у пакеті NLTK визначає для виключення такі загальні слова, як-то a, to, can.

Тепер розглянемо один з найвідоміших методів обробки природньої мови BERT та кілька його основних «послідовників».

BERT(Bidirectional Encoder Representations from Transformers) – цей метод базується на використання нейромереж нової архітектури для роботи з послідовностями, відомих як «трансформери». Ця технологія була розроблена компанією Google й пройшла попередню підготовку в корпусі Wikipedia. Проривом в роботі BERT є його здатність навчати мовні моделі на основі всього набору слів у додатку або запиті (двостороннє навчання), тоді як у традиційному навчанні аналізується упорядкована послідовність слів (зліва направо або справа наліво). BERT дозволяє мовній моделі розуміти контекст слова на основі оточуючих його слів, а не тільки того слова, яке йому передує або слідує відразу за ним. Схематично принцип роботи BERT відображений на рисунку нижче. (рис. 1.3) [4].

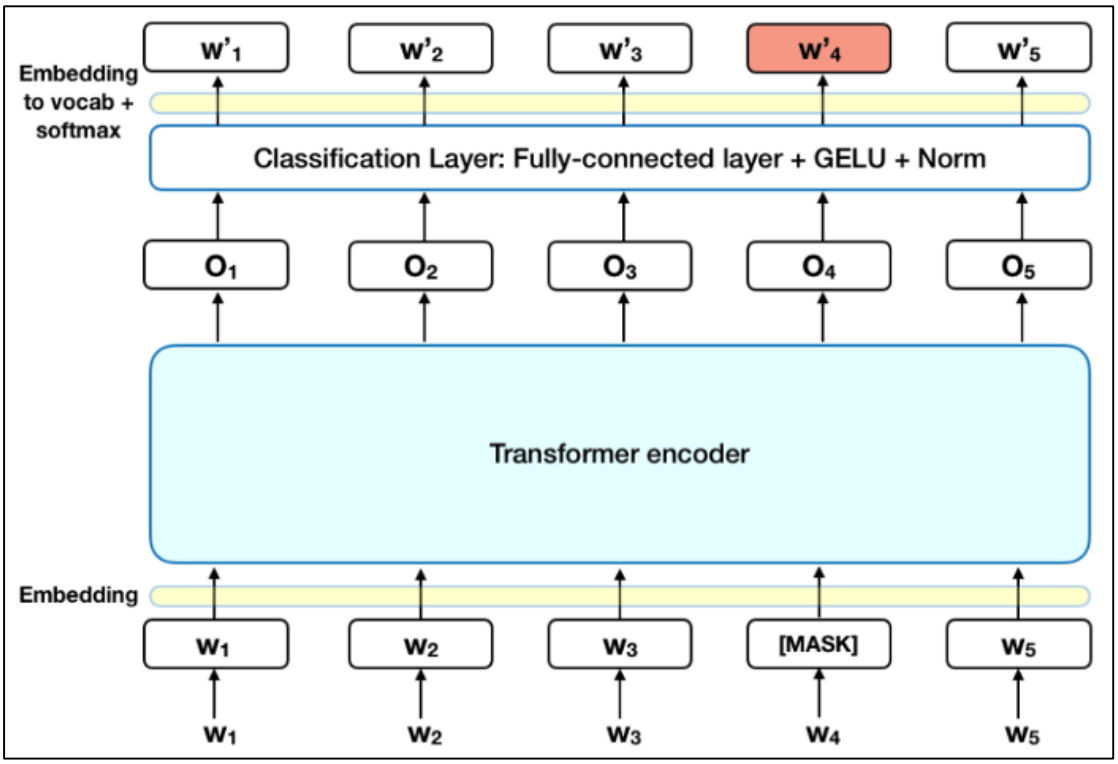

Рисунок 1.3 – Схема роботи BERT

Технологія BERT є у вільному доступі й містить в собі усі необхідні інструменти для аналізу, обробки, токенізації й розпізнаванню тексту, тому завдяки цьому будь-хто може використовувати цю технологію для навчання своєї власної системи обробки природної мови — для відповідей на запитання чи інших завдань.

DistilBERT – один з послідовників BERT, який націлений на оптимізацію навчання за рахунок зменшення розміру й збільшення швидкості BERT, но при цьому збереженні продуктивності. DistilBERT використовує BERT як «вчителя» своїй роботі. Дана модель важить на 40% менше й на 60% швидша чим оригінальна BERT – модель, і при тому зберігає 97% її функціональності. Такі результати були досягненні за рахунок використання лише 6 блоків енкодера, на відмінну від BERT, в якому їх 12. Ще одною важливою відмінністю DistilBERT є його алгоритм передтренування, воно відбувається лише через моделювання мови по масці. Спрощена схема роботи цієї моделі відображена на рисунку нижче (рис. 1.4) [5].

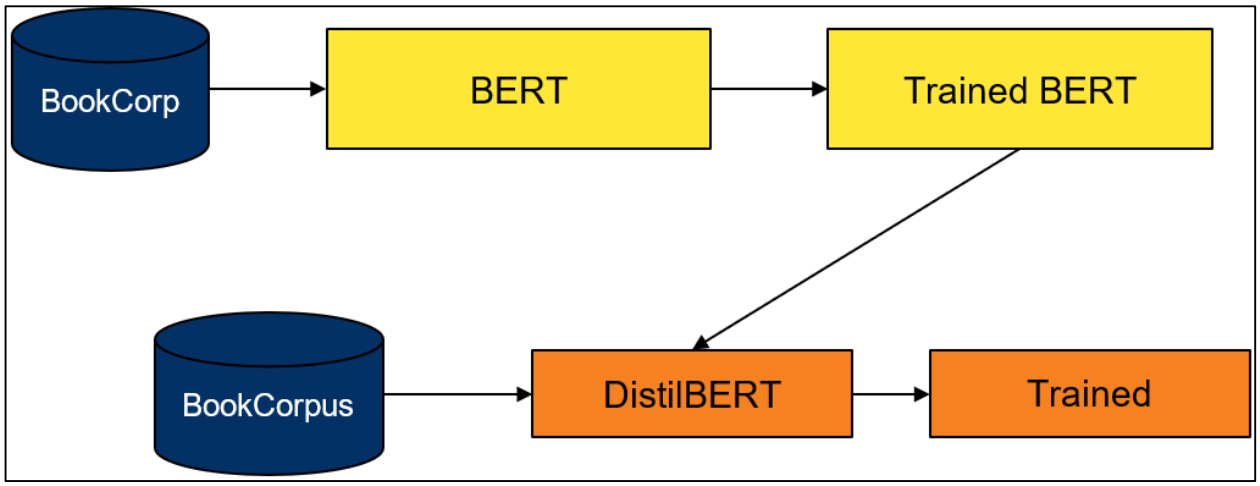

Рисунок 1.4 – Спрощена схема роботи DistilBERT

ALBERT – ще один послідовник BERT, який був представлений приблизно в той же час що і DistilBERT, його основними відмінностями від BERT це зменшення розміру моделі, приблизно в 18 раз менше параметрів, а

також сам процес навчання займає в 1,7 рази менше часу, але на відміну від DistilBERT, ALBERT не має переваги в продуктивності. Ця різниця полягає в тому що, він як і BERT навчається «з нуля», на відміну від DistilBERT, який використовує BERT як вчителя для процесу навчання. Схема роботи ALBERT показана на рисунку далі. (рис. 1.5) [6].

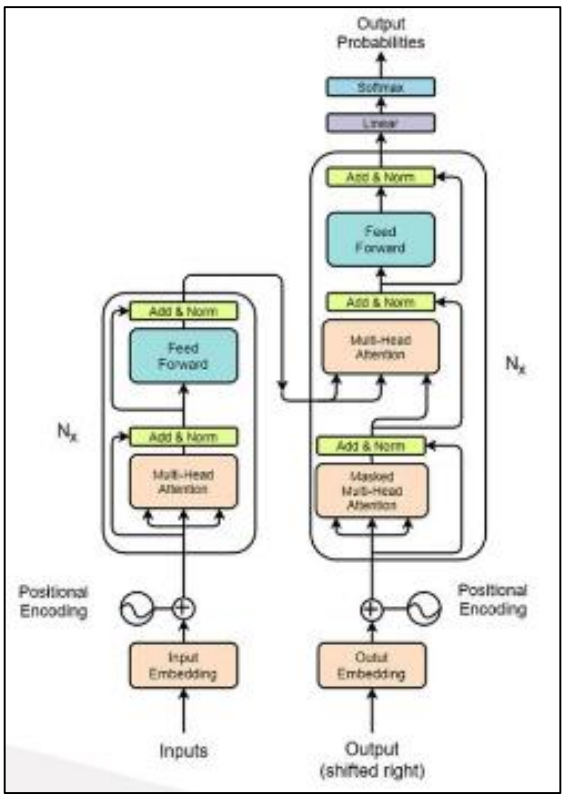

Рисунок 1.5 – Схема роботи ALBERT

Розглянувши вище описані методи обробки природньої мови, можемо сказати, що для вирішення поставленої задачі оптимальним буде використовувати DistilBERT, за рахунок його швидкості, продуктивності й зменшеної ваги моделі.

1.3 Аналіз моделей машинного навчання для класифікації ембеддингів

Класифікація тексту — це процес розуміння значення неструктурованого тексту та впорядкування його в попередньо визначені категорії (теги). Одним із найпопулярніших завдань класифікації тексту є аналіз настрою, мета якого класифікувати неструктуровані дані за настроєм.

Розглянемо кілька найпоширеніших моделей машинного навчання для класифікації, з якими працюватимемо в наступних розділах. Дані моделі входять до однієї з найвідоміших безкоштовних бібліотек для роботи з машинним навчанням – «scikit-learn».

Лінійна регресія – це контрольований алгоритм, який навчається моделювати залежну змінну, y, як функція деяких незалежних змінних (вона ж «особливості»), x_i, шляхом знаходження лінії (або поверхні), яка найкраще «відповідає» даним. Загалом, припускаємо y буди деяким числом і кожен x_i, може бути в основному що завгодно. Наприклад, прогнозування ціни будинку за кількістю кімнат у цьому будинку (y: ціна, x_1: кількість кімнат) або прогнозування ваги на основі зростання та віку (y: вага, x_1: висота, x_2: вік).

Загальне рівняння лінійної регресії має наступний вид [7] :

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{\text{стор}} x_{\text{стор}} + \epsilon$$

y – залежна змінна – те, що намагаємося передбачити;

x_i – незалежні змінні: функції, які використовує наша модель для моделювання y;

β_i – коефіцієнти (так звані «важелі») нашої моделі регресії; це – основа нашої моделі, саме їх наша модель «навчає» під час оптимізації;

ϵ – незнижувана помилка в нашій моделі; це – термін, який об’єднує всі немодельовані частини наших даних.

Лінійна регресія є простою та потужною моделлю для прогнозування числової відповіді на основі набору однієї чи кількох незалежних змінних. І

хоча може здатися, що лінійна регресія ігнорується в сучасному машинному навчанні, де постійно зростає складна архітектура нейронних мереж, алгоритм все ще широко використовується у великій кількості областей, оскільки він ефективний, його легко інтерпретувати та легко розширювати. Ключові ідеї лінійної регресії повторюються всюди, тому розуміння алгоритму є обов'язковим для міцної основи машинного навчання [7].

Random Forest є композицією (ансамблем) великої кількості дерев рішень, що дозволяє знизити проблему перенавчання та підвищити точність порівняно з одним деревом. Прогноз виходить у результаті агрегування відповідей безлічі дерев. Тренування дерев відбувається незалежно один від одного (на різних підмножинах), що не просто вирішує проблему побудови однакових дерев на тому самому наборі даних, але і робить цей алгоритм дуже зручним для застосування в системах розподілених обчислень. Взагалі, ідея бегінга, запропонована Лео Брейманом, добре підходить для розподілу обчислень. Для бегінга (незалежного навчання алгоритмів класифікації, де результат визначається голосуванням) є сенс використовувати велику кількість дерев рішень із досить великою глибиною. Під час класифікації фінальним результатом буде той клас, за який проголосувала більшість дерев, за умови, що одне дерево має один голос.

Random Forest (через незалежну побудову глибоких дерев) вимагає дуже багато ресурсів, а обмеження на глибину зашкодить точності (для вирішення складних завдань потрібно побудувати багато глибоких дерев). Можна зауважити, що час навчання дерев зростає приблизно лінійно їх кількості [8].

Звичайно, збільшення висоти (глибини) дерев не найкраще позначається на продуктивності, але підвищує ефективність цього алгоритму (хоча і разом з цим підвищується схильність до перенавчання). Занадто сильно боятися перенавчання не слід, оскільки це буде компенсовано числом дерев. Але й захоплюватись теж не слід. Всюди важливі оптимально підібрані параметри (гіперпараметри).

BaggingClassifier — це варіант алгоритму бегінгу (Bagging). У бегінгу видобуваємо з вихідного тренувального набору бутстрап-вибірки (вибираються випадкові випробування з поверненням), на яких навчаємо незалежні класифікатори, а результат класифікації визначається шляхом голосування. Звідси й назва терміна — bagging, тобто bootstrap aggregating (агрегування бутстрап-вибірок). У scikit-learn методи пакетування пропонуються як уніфікована BaggingClassifier (відповідно BaggingRegressor), яка приймає в якості вхідних даних визначену користувачем базову оцінку разом із параметрами, що визначають стратегію малювання випадкових підмножин. Зокрема, max_samples і max_features контролювати розмір підмножин (з точки зору зразків і ознак), а bootstrap й bootstrap_features контролювати, чи зразки і особливості малюються з заміною чи без неї. При використанні підмножини доступних зразків точність узагальнення можна оцінити для зразків із пакету, встановивши oob_score=True [9].

Multi-layer Perceptron (MLP) — це контрольований алгоритм навчання, який вивчає функцію $f(\cdot) : R^m \to R^o$ навчаючись на датасеті, де m – кількість вимірів для введення й o – це кількість вимірів для виведення. Дано набір ознак $X = x_1, x_2, \ldots, x_m$ і ціль y, він може вивчати апроксиматор нелінійної функції для класифікації або регресії. Вона відрізняється від логістичної регресії тим, що між вхідним і вихідним шарами може бути один або кілька нелінійних шарів, які називаються прихованими шарами. На рисунку показано одного прихованого шару MLP зі скалярним виходом (рис. 1.6) [10].

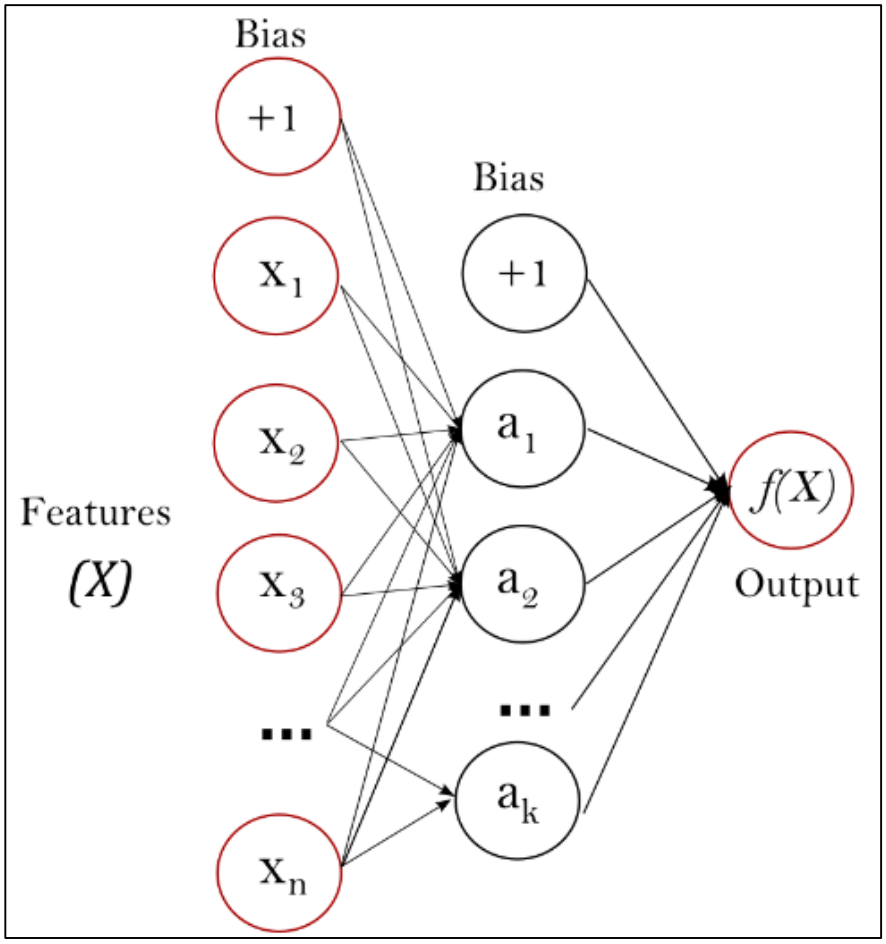

Рисунок 1.6 – Один з багатьох шарів MLP

XGBoost – це алгоритм машинного навчання, заснований на дереві пошуку рішень та використовує фреймворк градієнтного бустингу. У завданнях передбачення, які використовують неструктуровані дані (наприклад, зображення чи текст), штучна нейронна мережу перевершує й інші алгоритми чи фреймворки. Але коли справа доходить до структурованих або табличних даних невеликих розмірів, у першості виявляються алгоритми, що ґрунтуються на дереві пошуку рішень. На інфографіці можна переглянути еволюцію таких алгоритмів [11].

1.4 Висновки

В даному розділі розглянуто проблему автоматичної класифікації англомовної текстової інформації. Проаналізувавши проблему, можемо

зробити висновок, що ручний синтез тексту займає велику кількість ресурсів, тому цей процес краще автоматизувати шляхом розробки інтелектуальної інформаційної технологію.

Ознайомившись з доступними методами обробки природної мови (BERT, DistilBERT, ALBERT) прийнято рішення використовувати у даній роботи DistilBERT, оскільки він має переваги над іншими моделями завдяки своїй швидкості, більшому показнику продуктивності й зменшеній вазі моделі, що найкраще підходить під поставлену задачу.

Також, в цьому розділі проаналізовано кілька відомих моделей машинного навчання для класифікації тексту та вибрано оптимальні, які варто використовувати для розв’язання поставленої задачі. Усі вони – різні за своєю архітектурою та мають свої переваги в конкретних задачах, але поки невідомо яка з них краще підходить під класифікацію англомовного тексту про екологічні проблеми водних ресурсів.

2 ВИБІР ОПТИМАЛЬНИХ НАЛАШТУВАНЬ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ДЛЯ РОЗВ'ЯЗАННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ

2.1 Огляд датасету для тренування моделей

Для початку роботи з моделями, потрібно сформувати датасет й вручну класифікувати його, щоб на його основі тренувати їх й обрати оптимальну.

Для роботи використовується датасет «NLP : Reports & News Classification» (рис. 2.1), який був створений за моєї участі 2 роки тому.

	A	B	C	D	E	F
1	text	env_problems	pollution	treatment	climate	biomonitoring
2	In the Southern Bug basin wastewater is discharged from 341 organized sources (Map 5). In 2	1	1			
3	This amount is distributed as follows: urban and rural housing wastewater treatment facilitie	1	1			
4	Despite the similar volumes of discharged wastewater major part of pollutants comes with c	1	1			
5	They bring 84% of organic pollution 86% of phosphate ions and 84% of mineral nitrogen 91%	1	1			
6	The input of the industry is between 7вЂ“21% and agriculture has the lowest impact on wate	1	1			
7	Settlements that do not have sewage systems usually do not have central water supply Vod	1		1		
8	Equipment of WWTPs is very worn-out the amount of wastewater often exceeds their treatı	1		1		
9	The current state of sewage networks is a source of concern since about 24% of them due to	1		1		
10	Consequently in Ukraine on average 2 accidents occur at 1 km network per year which is muc	1	1			
11	The load of pollutants from accidental discharges cannot presently be estimated.	0				
12	The degree of connectivity of the population to WWTPs varies widely - from 11% to 95%. The	1		1		
13	In large cities with EH close or greater than 100000 EH connection to sewers ranges 85-92 80%	1		1		
14	Based on expert judgment taking that average degree of connectivity of the urban populatic	1	1	1		
15	The obtained results show that the main part (77%) of inorganic nitrogen comes to waste wa	1	1			
16	The majority of the wastewater treatment facilities in the Southern Bug basin are equipped	1		1		
17	This method allows extracting only a part of pollutants containing nitrogen and phosphorus.	1	1	1		
18	Technologies of deep extraction of these compounds are not used in Ukraine due to their hi	1		1		
19	The existing wastewater treatment facilities were constructed during the Soviet time they a	1	1	1		
20	Very often the volumes of wastewater exceed designed values which lead to additional pol	1	1	1		

Рисунок 2.1 – Вигляд датасету «NLP : Reports & News Classification»

Джерелом інформації послугували відкриті дані щодо екологічного стану річок басейну Південного Бугу.

1. Речення скопійовані з відкритих текстових даних англомовних джерел у форматі PDF разом із різними спец-символами (номерами сторінок тощо), щоб був певний «шум».

2. Деякі спеціальні символи були вибірково видалені, включаючи деякі коми та крапки.

3. Визначення цільових ознак виконано авторами цього набору даних, за моєї участі.

2.2 Розвідувальний аналіз

Для роботи моделі штучного інтелекту вихідними даними синтезу тексту, є датесет із такими базовими ознаками (рис. 2.2):

– «text» – речення із звітів чи новин;
– «env_problems» – наявність у тексті інформації про екологічну проблему;
– «pollution» – чи текст про забруднення навколишнього середовища?
– «treatment» – текст про очисні споруди чи екологічні технології;
– «climate» – чи текст про кліматичні показники?
– «biomonitoring» – текст про біологічний, біотичний моніторинг у воді чи в басейні річки.

	text	env_problems	pollution	treatment	climate	biomonitoring
0	In the Southern Bug basin wastewater is discha...	1	1	0	0	0
1	This amount is distributed as follows: urban a...	1	1	0	0	0
2	Despite the similar volumes of discharged wast...	1	1	0	0	0
3	They bring 84% of organic pollution 86% of pho...	1	1	0	0	0
4	The input of the industry is between 7–21% and...	1	1	0	0	0
...	...	...	...	...	...	...
95	Chemical composition of water bodies shall com...	0	0	0	0	0
96	According to the Directive provisions water bo...	0	0	0	0	0
97	Besides in the field of data processing during...	0	0	0	0	0
98	Use of these values was offered by the Monitor...	0	0	0	0	0
99	To determine confidence levels for assessment ...	0	0	0	0	0

Рисунок 2.2 – Ознаки датасету «NLP : Reports & News Classification»

Далі необхідно провести розвідувальний аналіз даних про очисні споруди чи екологічні технології (treatment problems).

Переглянемо середню довжину слова в тексті (рис. 2.3) [12].

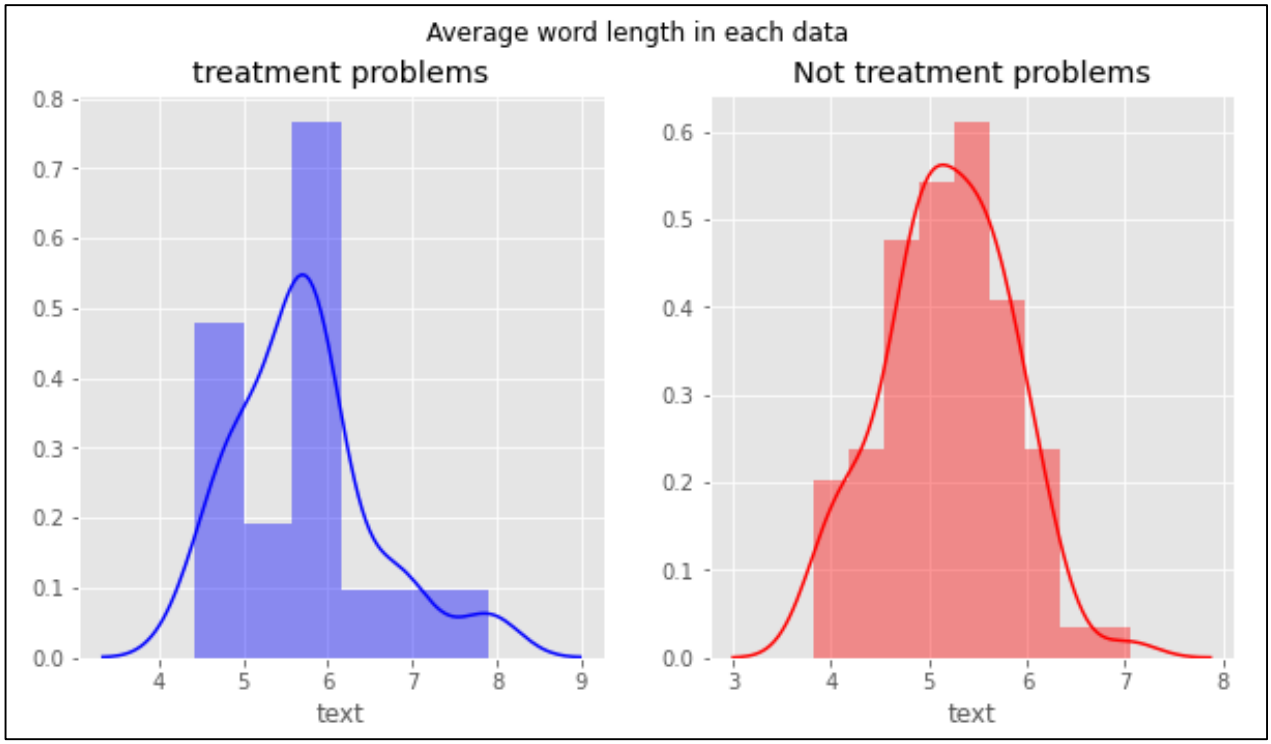

Рисунок 2.3 – Середня довжина слів

Найбільша кількість слів з 5 та 6 символами.

Далі переглянемо стоп-слова в тексті та їх кількість в «Treatment» та «Not treatment» даних (рис 2.4, 2.5) [12].

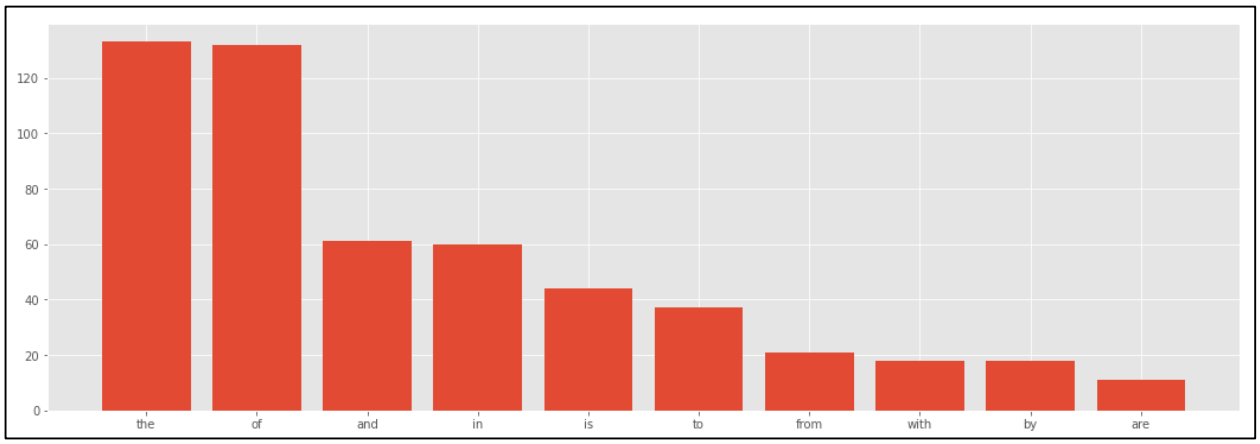

Рисунок 2.4 – Кількість стоп-слів в «Treatment» даних

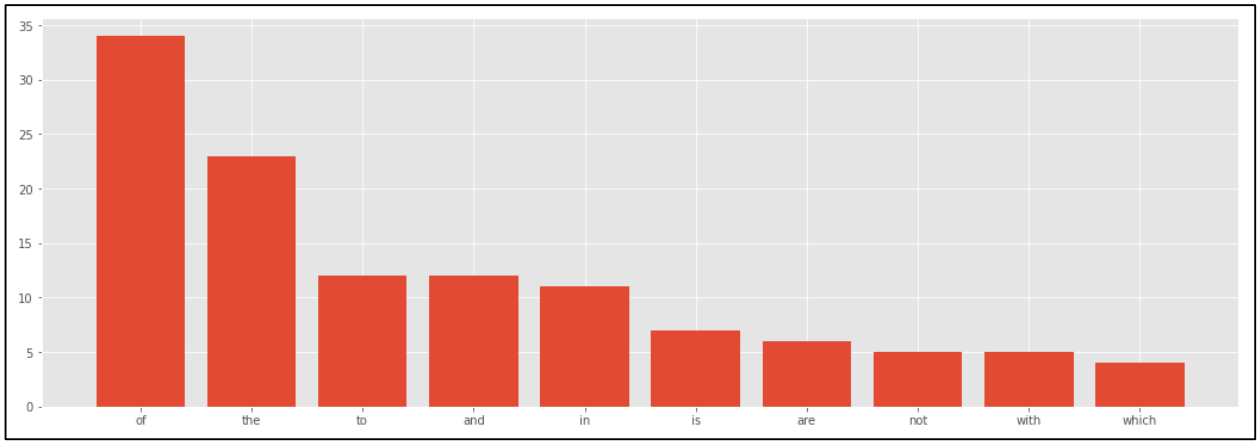

Рисунок 2.5 – Кількість стоп-слів в «Not treatment» даних

Серед стоп-слів найпоширенішими є слова «of» та «the».

Також важливим фактором є частота вживання кожного слова (рис. 2.6) [12].

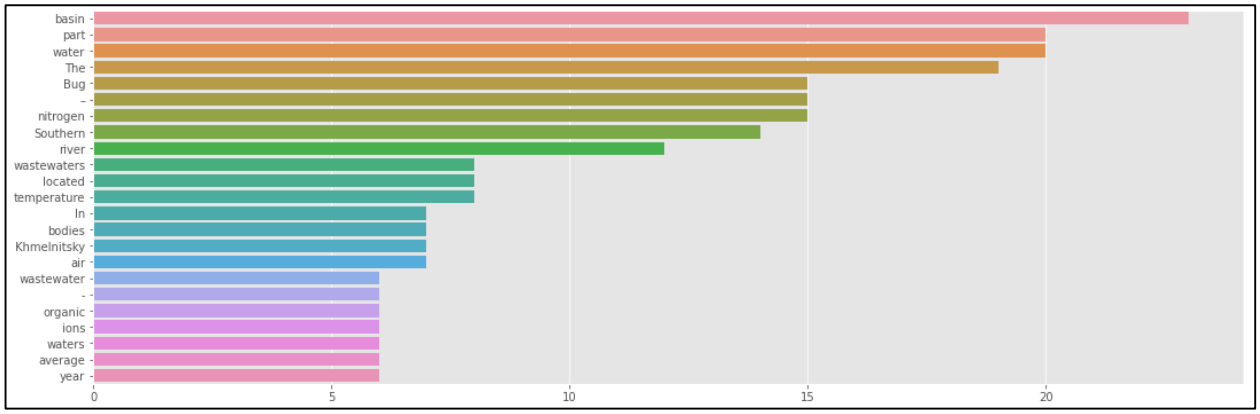

Рисунок 2.6 – Загальновживані слова

Найбільш вживаним є слово «basin» – басейн, що є закономірним, адже це – текст про річку «Південний Буг». Також є проблемою те, що одним з найпоширеніших є слово «The», тому цей текст потребує очищення.

Далі йде створення хмари слів із значеннями під кадром даних категорії, загальновживані в обох типах («Treatment» та «Not treatment») (рис. 2.7, 2.8) [12].

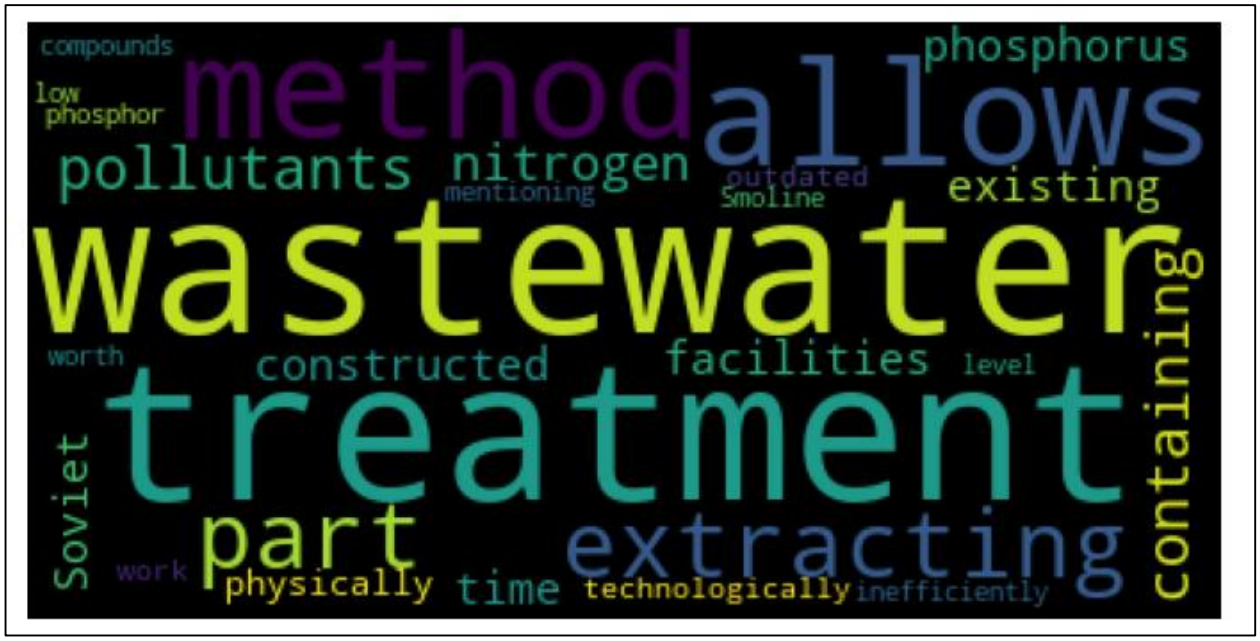

Рисунок 2.7 – Хмара слів у кластері «Treatment»

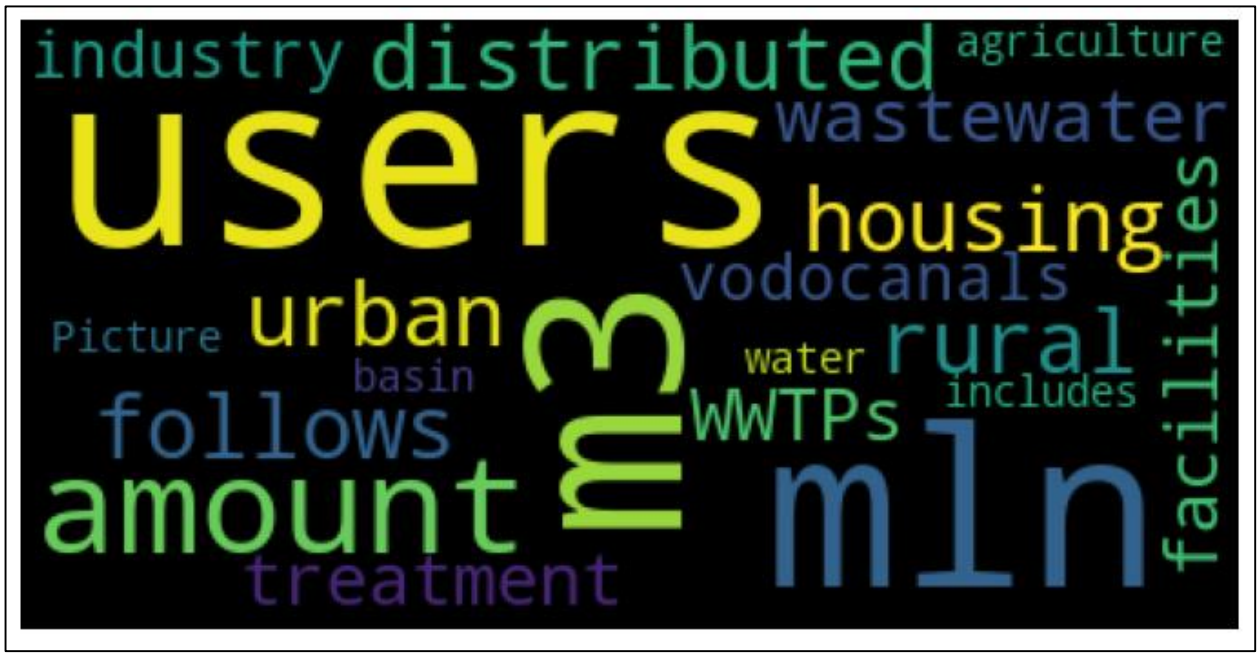

Рисунок 2.8 – Хмара слів у кластері «Not treatment»

Здійснимо візуалізацію вкладень (рис 2.9).

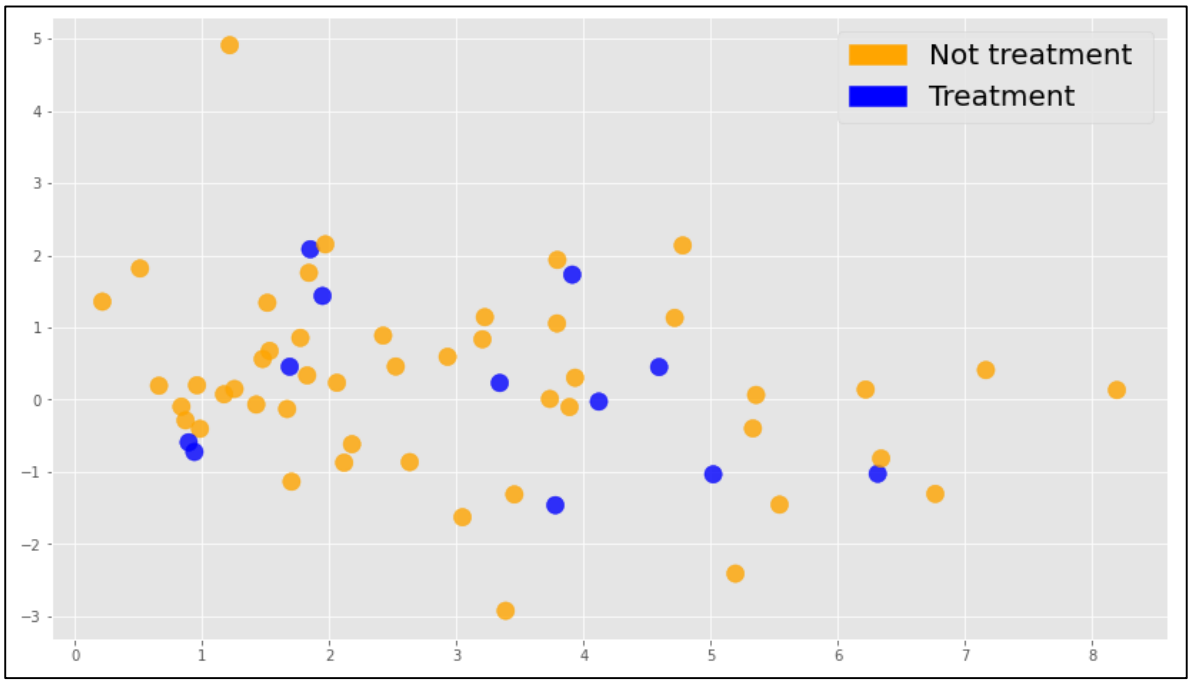

Рисунок 2.9 – Візуалізація вкладень

Вкладення виглядають не дуже чітко розділеними. Подивимося, чи є можливість їх кластеризувати краще (рис. 2.10) [12].

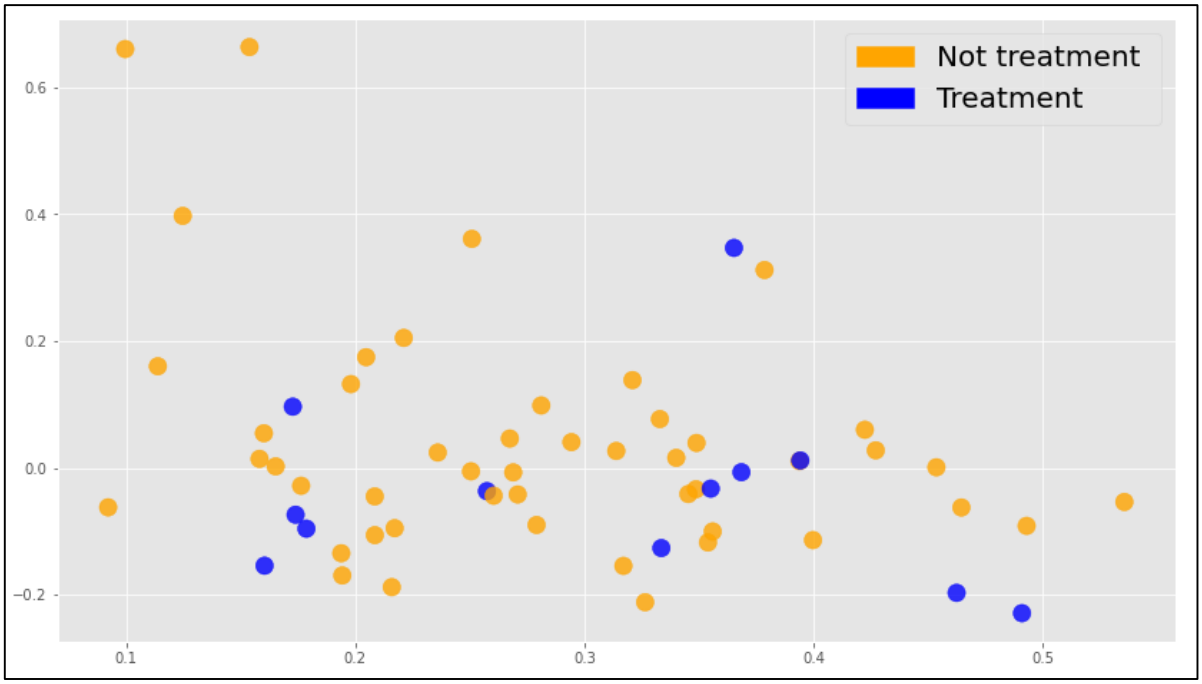

Рисунок 2.10 – Кластеризація даних

Ці дані непогано кластеризувались.

2.3 Створення передтренованих моделей для класифікації даних

Щоб використовувати інформаційну технологію, яка розробляється, потрібні передтреновані моделі-класифікатори. Алгоритм їх розробки наведений на рисунку 2.11.

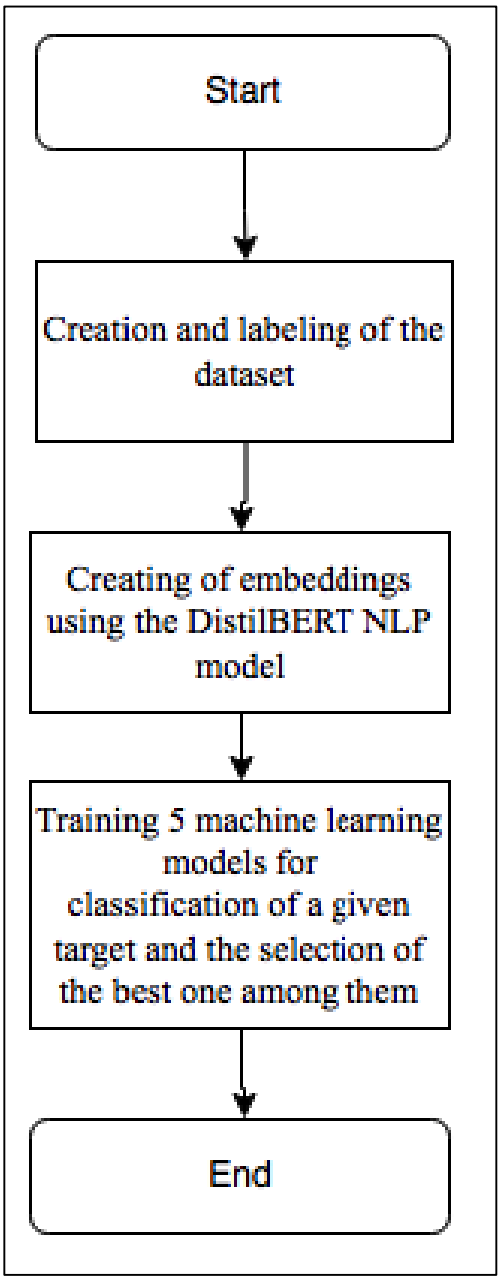

Рисунок 2.11 – Алгоритм розробки моделей-класифікаторів

Відповідно до цього алгоритму, першим етапом є створення та розмітка датасету, цей процес був описаний в підрозділі 2.1, саме цей датасет буде використовуватись. Наступним етапом є формування ембедингів, за допомогою методу обробки природньої мови DistilBERT, оскільки саме його обрано при аналізі огляду технологій для вирішення проблему в підрозділі 1.2. І фінальним кроком на цьому етапі буде тренування моделей, які були відібрані в підрозділі 1.3 та збирання їх результатів, на основі яких в кінці оберемо найкращі для подальшого використання на довільних датасетах.

Після усіх етапів описаних вище, приступаємо до безпосередньо тренування кожної моделі та їх подальшого порівняння (рис. 2.12) [13].

```
import numpy as np
import pandas as pd

import matplotlib
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score

# models
from sklearn.linear_model import LinearRegression, Perceptron, RidgeClassifier, SGDClassifier, LassoCV
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, ExtraTreesClassifier
from sklearn.ensemble import BaggingClassifier, AdaBoostClassifier, VotingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn import metrics
import xgboost as xgb
from xgboost import XGBClassifier

# NN models
import keras
from keras.models import Sequential
from keras.layers import Dense
from keras import optimizers
from keras.wrappers.scikit_learn import KerasClassifier

import torch
import transformers as ppb

import warnings
warnings.filterwarnings('ignore')
```

Рисунок 2.12 – Імпортуємо необхідні бібліотеки для роботи

Далі завантажимо підготовлений датасет для тренування моделей (рис. 2.13, 2.14) [13].

```
df = pd.read_csv('../input/nlp-reports-news-classification/water_problem_nlp_en_for_Kaggle_100.csv', delimiter=';', header=0)
df = df.fillna(0)

convert_dict = {'text': str,
                'env_problems': int,
                'pollution': int,
                'treatment': int,
                'climate': int,
                'biomonitoring': int}

df = df.astype(convert_dict)
df
```

	text	env_problems	pollution	treatment	climate	biomonitoring
0	In the Southern Bug basin wastewater is discha...	1	1	0	0	0
1	This amount is distributed as follows: urban a...	1	1	0	0	0
2	Despite the similar volumes of discharged wast...	1	1	0	0	0
3	They bring 84% of organic pollution 86% of pho...	1	1	0	0	0
4	The input of the industry is between 7–21% and...	1	1	0	0	0
...	...	...	...	...	...	...
95	Chemical composition of water bodies shall com...	0	0	0	0	0
96	According to the Directive provisions water bo...	0	0	0	0	0
97	Besides in the field of data processing during...	0	0	0	0	0
98	Use of these values was offered by the Monitor...	0	0	0	0	0
99	To determine confidence levels for assessment ...	0	0	0	0	0

100 rows × 6 columns

Рисунок 2.13 – Завантаження датасету

Виводимо інформацію про завантажений датасет (рис. 2.14).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   text           100 non-null    object
 1   env_problems   100 non-null    int64 
 2   pollution      100 non-null    int64 
 3   treatment      100 non-null    int64 
 4   climate        100 non-null    int64 
 5   biomonitoring  100 non-null    int64 
dtypes: int64(5), object(1)
memory usage: 4.8+ KB
```

Рисунок 2.14 – Інформація про завантажений датасет

Тепер завантажимо попередньо треновану модель DistilBERT для подальшої роботи з нею (рис. 2.15).

```
# For pre-trained DistilBERT:
model_class, tokenizer_class, pretrained_weights = (ppb.DistilBertModel, ppb.DistilBertTokenizer, 'distilbert-base-uncased')

# Other models: https://huggingface.co/transformers/pretrained_models.html

# Load pretrained model/tokenizer
tokenizer = tokenizer_class.from_pretrained(pretrained_weights)
model = model_class.from_pretrained(pretrained_weights)
```

Рисунок 2.15 – Завантаження DistilBERT

Тепер токенізуємо речення, тобто приведемо дані до зручного для BERT виду (рис. 2.16).

```
# Tokenization the sentences - break them up into word and subwords in the format BERT is comfortable with
tokenized = df['text'].apply((lambda x: tokenizer.encode(x, add_special_tokens=True)))

max_len = 0
for i in tokenized.values:
    if len(i) > max_len:
        max_len = len(i)

padded = np.array([i + [0]*(max_len-len(i)) for i in tokenized.values])
np.array(padded).shape
```

```
(100, 78)
```

```
# Creation variable to ignore (mask) the data padding
attention_mask = np.where(padded != 0, 1, 0)
print(attention_mask.shape)
attention_mask
```

```
(100, 78)
array([[1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 1, 1, 1],
       [1, 1, 1, ..., 0, 0, 0],
       ...,
       [1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0]])
```

Рисунок 2.16 – Токенізація й форматування даних для BERT

Наступний крок: на основі токенізованих даних й попередньо тренованої моделі DistilBERT створюємо модель для нашої задачі (рис. 2.17).

```
# Modeling
input_ids = torch.tensor(padded).to(torch.int64)
attention_mask = torch.tensor(attention_mask).to(torch.int64)

with torch.no_grad():
    last_hidden_states = model(input_ids, attention_mask=attention_mask)
```

```
# Last hidden states
features = last_hidden_states[0][:,0,:].numpy()
```

Рисунок 2.17 – Створення моделі DistilBERT

Задаємо параметри для налаштування роботи обраних моделей й зберігаємо середнє й максимальне значення результатів роботи (рис. 2.18, 2.19) [13].

```
# Set parameters of models
models = pd.DataFrame(columns = ['name', 'model', 'param_grid'])

# Linear Regression
n = 0
models.loc[n, 'name'] = 'Linear Regression'
models.at[n, 'model'] = LinearRegression()
models.at[n, 'param_grid'] = {}

# Random Forest Classifier
n = 1
models.loc[n, 'name'] = 'Random Forest Classifier'
models.at[n, 'model'] = RandomForestClassifier()
models.at[n, 'param_grid'] = {'n_estimators': [40, 50, 60, 100, 500],
                              'min_samples_split': [30, 40, 50, 100, 200],
                              'min_samples_leaf': [10, 12, 15, 20, 50],
                              'max_features': ['auto'],
                              'max_depth': [3, 4, 5, 6],
                              'criterion': ['gini'],
                              'bootstrap': [False]
                             }

# Bagging Classifier
n = 2
models.loc[n, 'name'] = 'Bagging Classifier'
models.at[n, 'model'] = BaggingClassifier()
models.at[n, 'param_grid'] = {'max_features': np.linspace(0.05, 0.8, 1),
                              'n_estimators': [3, 4, 5, 6],
                              'warm_start' : [False]
                             }

# XGB Classifier
n = 3
models.loc[n, 'name'] = 'XGB Classifier'
models.at[n, 'model'] = xgb.XGBClassifier(objective='binary:hinge')  # or binary:logistic
models.at[n, 'param_grid'] = {'n_estimators': [50, 70, 90],
                              'learning_rate': [0.01, 0.05, 0.1, 0.2],
                              'max_depth': [3, 4, 5]
                             }
```

Рисунок 2.18 – Параметри обраних моделей

```
# MLP Classifier
n = 4
models.loc[n, 'name'] = 'MLPClassifier'
models.at[n, 'model'] = MLPClassifier()
models.at[n, 'param_grid'] = {'hidden_layer_sizes': [i for i in range(2,10)],
                              'solver': ['sgd'],
                              'learning_rate': ['adaptive'],
                              'max_iter': [1000, 2000]
                             }

# Avg values
models.loc[5, 'name'] = 'Mean values'

# Max values
models.loc[6, 'name'] = 'Max values'
```

Рисунок 2.19 – Параметри обраних моделей (продовження)

Після того, як задано всі необхідні параметри, матимемо наступну картину (рис. 2.20).

	name	model	param_grid
0	Linear Regression	LinearRegression()	{}
1	Random Forest Classifier	RandomForestClassifier()	{'n_estimators': [40, 50, 60, 100, 500], 'min_...
2	Bagging Classifier	BaggingClassifier()	{'max_features': [0.05], 'n_estimators': [3, 4...
3	XGB Classifier	XGBClassifier(base_score=None, booster=None, c...	{'n_estimators': [50, 70, 90], 'learning_rate'...
4	MLPClassifier	MLPClassifier()	{'hidden_layer_sizes': [2, 3, 4, 5, 6, 7, 8, 9...
5	Mean values	NaN	NaN
6	Max values	NaN	NaN

Рисунок 2.20 – Відображення параметрів усіх моделей

Тепер, створимо функцію для тренування обраних моделей та обчислення отриманих результатів на підготовлених даних, а також розділимо датасет для майбутнього тюнінгу, тобто для тренування моделей (рис. 2.21-2.23) [13].

```python
def model_prediction(models, features, labels, test_size, verbose):
    # Models training and data prediction for all models from DataFrame models

    # Splitting train data for model tuning with cross-validation
    #cv_train = ShuffleSplit(n_splits=cv_n_split, test_size=0.2, random_state=random_state)
    train_features, test_features, train_labels, test_labels = train_test_split(features,
                                                                                labels,
                                                                                test_size=test_size,
                                                                                random_state=random_state)
    # Total list of prediction by all models
    total_train_pred = []
    total_test_pred = []

    # Results
    results = models[['name']].copy()
    results['acc_train'] = results['acc_test'] = 0

    for i in range(len(models)-2):
        # Training
        model = GridSearchCV(models.at[i, 'model'], models.at[i, 'param_grid'])
        model.fit(train_features, train_labels)

        # Prediction
        train_pred = model.predict(train_features).round(0).astype('int')
        total_train_pred.append(train_pred)
        test_pred = model.predict(test_features).round(0).astype('int')
        total_test_pred.append(test_pred)

        # Scoring
        acc_train = accuracy_score(train_labels, train_pred)
        acc_test = accuracy_score(test_labels, test_pred)
```

Рисунок 2.21 – Функція для тренування моделей – тюнінг моделей

```python
        if verbose:
            print(f"Model - {models.loc[i, 'name']}')
            print("Best parameters:", model.best_params_)
            print("Accuracy for training data: %0.3f" % acc_train)
            print("Accuracy for test data: %0.3f" % acc_test)
            print('\n')

        # Save results
        results.loc[i,'acc_train'] = acc_train
        results.loc[i,'acc_test'] = acc_test

    # MEAN values
    # Calc prediction
    mean_train_pred = (np.mean(np.array(total_train_pred), axis=0)).astype(int)
    mean_test_pred = (np.mean(np.array(total_test_pred), axis=0)).astype(int)

    # Scoring
    acc_train = accuracy_score(train_labels, mean_train_pred)
    acc_test = accuracy_score(test_labels, mean_test_pred)
    if verbose:
        print(f"Mean prediction values")
        print("Accuracy for training data: %0.3f" % acc_train)
        print("Accuracy for test data: %0.3f" % acc_test)

    # Save results
    n = len(results)-2
    results.loc[n,'acc_train'] = acc_train
    results.loc[n,'acc_test'] = acc_test
```

Рисунок 2.22 – Функція для тренування моделей – середні значення

```
    # MAX values
    # Calc prediction
    max_train_pred = (np.max(np.array(total_train_pred), axis=0))
    max_test_pred = (np.max(np.array(total_test_pred), axis=0))

    # Scoring
    acc_train = accuracy_score(train_labels, max_train_pred)
    acc_test = accuracy_score(test_labels, max_test_pred)
    if verbose:
        print(f"Maximum prediction values")
        print("Accuracy for training data: %0.3f" % acc_train)
        print("Accuracy for test data: %0.3f" % acc_test)

    # Save results
    n = len(results)-1
    results.loc[n,'acc_train'] = acc_train
    results.loc[n,'acc_test'] = acc_test

    return results
```

Рисунок 2.23 – Функція для тренування моделей – максимальні значення

І фінальним кроком є написання функції для передбачення класів даних по заданих таргетах, й відповідно збереження та відображення результатів та статистики (рис. 2.24) [13].

```
def target_prediction(models, df, features, target, test_size=0.2, verbose=True):
    # Text classification model and prediction for given feature "target" (with labels) in df

    # Target
    labels = df[target]

    # Extracting the number of examples of each class
    Relevant_len = df[df[target] == 1].shape[0]
    Not_len = df[df[target] == 0].shape[0]

    # Draw bar plot
    plt.rcParams['figure.figsize'] = (7, 5)
    plt.bar(10, Relevant_len, 3, label="Relevant", color='green')
    plt.bar(15, Not_len, 3, label="Not", color='red')
    plt.legend(loc='upper center')
    plt.ylabel('Number of examples')
    plt.title('Propertion of examples for ' + target)
    plt.show()

    # Models training, prediction and save results
    results = model_prediction(models, features, labels, test_size, verbose=verbose)
    results = results.sort_values(by=['acc_test', 'acc_train'], ascending=False)
    results.to_csv(f'{target}-models-scoring.csv', index=False)
    display(results)
```

Рисунок 2.24 – Функція для передбачення по таргетах

Тепер залишилось викликати створену функцію й передати їй усі необхідні параметри, які були підготовлені в процесі роботи. Результатом виконання цієї функції будуть статистичні дані роботи кожної моделі по окремих таргетах (рис. 2.25).

```
%%time
# Solving NLP Classification tasks
print('Solving NLP Classification tasks')
for col in cols:
    target_prediction(models, df, features, col, test_size=0.4, verbose=True)
```

Рисунок 2.25 – Виклик функції для тренування моделей по таргетах

Результати роботи моделей по різних таргетах (рис. 2.26 – 2.30) [13].

	name	acc_train	acc_test
4	MLPClassifier	0.983333	0.975
1	Random Forest Classifier	0.966667	0.950
0	Linear Regression	1.000000	0.900
5	Mean values	0.983333	0.900
2	Bagging Classifier	0.983333	0.875
3	XGB Classifier	0.983333	0.800
6	Max values	0.950000	0.775

Рисунок 2.26 – Результати роботи моделей по таргету «env_problems»

	name	acc_train	acc_test
6	Max values	0.950000	0.825
0	Linear Regression	1.000000	0.800
2	Bagging Classifier	0.966667	0.775
3	XGB Classifier	0.966667	0.775
4	MLPClassifier	1.000000	0.750
1	Random Forest Classifier	0.950000	0.725
5	Mean values	0.950000	0.725

Рисунок 2.27 – Результати роботи моделей по таргету «pollution»

	name	acc_train	acc_test
0	Linear Regression	1.000000	0.900
4	MLPClassifier	0.966667	0.900
2	Bagging Classifier	0.933333	0.900
1	Random Forest Classifier	0.866667	0.850
5	Mean values	0.866667	0.850
3	XGB Classifier	1.000000	0.775
6	Max values	1.000000	0.775

Рисунок 2.28 – Результати роботи моделей по таргету «treatment»

	name	acc_train	acc_test
0	Linear Regression	1.000000	1.000
4	MLPClassifier	1.000000	0.950
6	Max values	1.000000	0.950
1	Random Forest Classifier	1.000000	0.925
2	Bagging Classifier	0.983333	0.925
5	Mean values	0.983333	0.875
3	XGB Classifier	1.000000	0.850

Рисунок 2.29 – Результати роботи моделей по таргету «climate»

	name	acc_train	acc_test
2	Bagging Classifier	0.983333	0.950
1	Random Forest Classifier	0.900000	0.950
4	MLPClassifier	0.900000	0.950
5	Mean values	0.900000	0.950
0	Linear Regression	1.000000	0.900
3	XGB Classifier	0.983333	0.850
6	Max values	0.983333	0.825

Рисунок 2.30 – Результати роботи моделей по таргету «biomonitoring»

Відтворивши всі дії описані вище, маємо зручну для аналізу статистику про обрані моделі, на основі якої вже можемо підібрати оптимальні для досягнення мети.

Проаналізувавши результати на рисунках 2.26–2.30, видно, що оптимальною моделлю є лінійна регресія, але це означає, що дані сильно зашумлені і це заважає більше потужним моделям. Як правильно, модель лінійної регресії є кращою, коли вона «оверфітиться» під дані. Отже, пропонується як оптимальну використовувати дві:

- Лінійна регресія;
- Випадкові ліси (Random Forest).

2.4 Висновки

В даному розділі розглянуто побудову датасету, його розвідувальний аналіз даних та застосування мовних NLP-моделей і технологій для вирішення поставленого завдання.

Проведено підготовку тренувального датасету, який був класифікований вручну, для того, щоб він був максимально результативним і моделі були гарно навченими. Здійснено розвідувальний аналіз для того, щоб можна було робити висновки щодо того, наскільки створений датасет відповідає заданим вимогам. Потім, описано алгоритм розробки моделі класифікаторів, сформувано ембеддинги за допомогою методу оброблення природньої мови DistilBERT, оскільки саме його обрано під час аналізу технологій для вирішення проблеми, після чого відбулось тренування моделей, які були відібрані в підрозділі 1.3, та проведено аналіз результатів їх роботи.

Результатом цього розділу є статистика по роботі кожної з 5 моделей машинного навчання, які розглянуто в першому розділі. Провівши аналіз результатів, маємо, що моделі лінійної регресії та випадкових лісів (Random Forest) показали найкращий результат, а саме – точність 0.80 та 0.95 відповідно, тому на них і продовжимо класифікацію в наступному розділі.

3 СТВОРЕННЯ ІНТЕЛЕКТУАЛЬНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ КЛАСИФІКАЦІЇ АНГЛОМОВНОЇ ІНФОРМАЦІЇ ПРО МАСИВИ ВОД БАСЕЙНУ РІЧКИ ПІВДЕННИЙ БУГ

3.1 Алгоритм програмного забезпечення інформаційної технології

Розв’язок задачі пропонується розбити на 2 етапи:

Етап 1. Налаштування інформаційної технології та синтез оптимальних передтренованих моделей-класифікаторів (див. розділ 2).

Етап 2. Застосування оптимальних передтренованих моделей-класифікаторів на заданих даних.

Алгоритм етапу 1 наведено на рис. 2.11. Загальний алгоритм розробленої інформаційної технології на етапі 2 наведений на рис. 3.1.

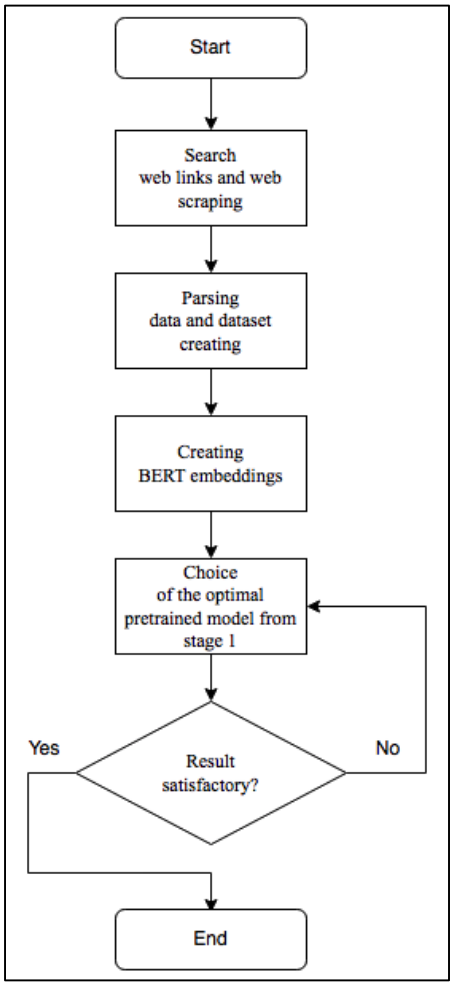

Рисунок 3.1 – Загальний алгоритм на етапі 2 технології, яка розробляється

Алгоритм дуже схожий до першого етапу своїми першими пунктами, оскільки також готується датасет для роботи з ним, але в цьому випадку не проводиться ручна класифікація та інші операції, а лише вони збираються з джерел та пропускаються через парсер, щоб отримати csv файл для роботи з моделями-класифікаторами. Після підготовки датасету, розпочинаємо тестування відібраних моделей, які вже треновані на попередньому датасеті з підрозділу 2.1, і якщо результат – задовільний, то закінчуємо програму, якщо ж ні – повертаємось до вибору моделі й беремо іншу.

3.2 Формування датасету

В якості джерела даних обрано англомовну монографію [2].

Для формування датасету розроблено парсер відповідно до вимог поставленої задачі [14]. Парсер містить в собі можливість вибору таких конфігурацій (рис. 3.2):

– мова тексту;
– тип джерела даних;
– мітка сегмента.

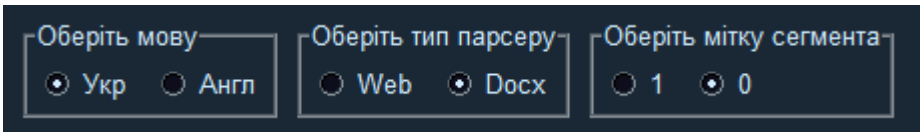

Рисунок 3.2 – Конфігурації парсера

При роботі з веб-документами, перш за все, потрібно провести самостійний аналіз веб-сторінки для знаходження ідентифікатора («id» або «class html» тега), на який буде орієнтуватись парсер. Далі можемо переходити до конфігурації парсера, для цього потрібно заповнити такі поля (рис. 3.3):

– URL – посилання на джерело;

– ID/class – ввести ідентифікатор пошуку, який був отриманий при самостійному аналізі;
– обрати тип ідентифікатора.

Рисунок 3.3 – Конфігурація для роботи з веб-документами

При роботі з текстовими документами у форматі docx алгоритм значно простіший, а саме в полі «Filepath» вказати шлях до файлу в системі (рис. 3.4).

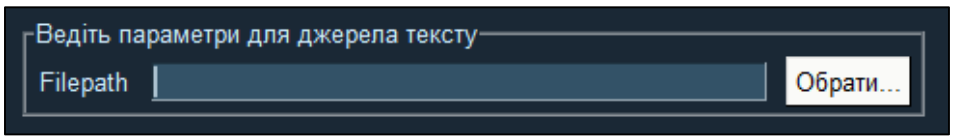

Рисунок 3.4 – Поле для шляху до джерела даних

Після парсингу даних з обраних джерел, щоб експортувати дані до csv-файлу, необхідно вказати його назву в полі «Output filename» та натиснути кнопку «Записати до CSV», згенерований файл буде знаходитись в тій же директорії, що й парсер (рис. 3.5).

Рисунок 3.5 – Поле для назви csv файлу

Після налаштування всіх необхідних конфігурацій та заповнення обов’язкових полів, програма матиме такий вигляд (рис. 3.6).

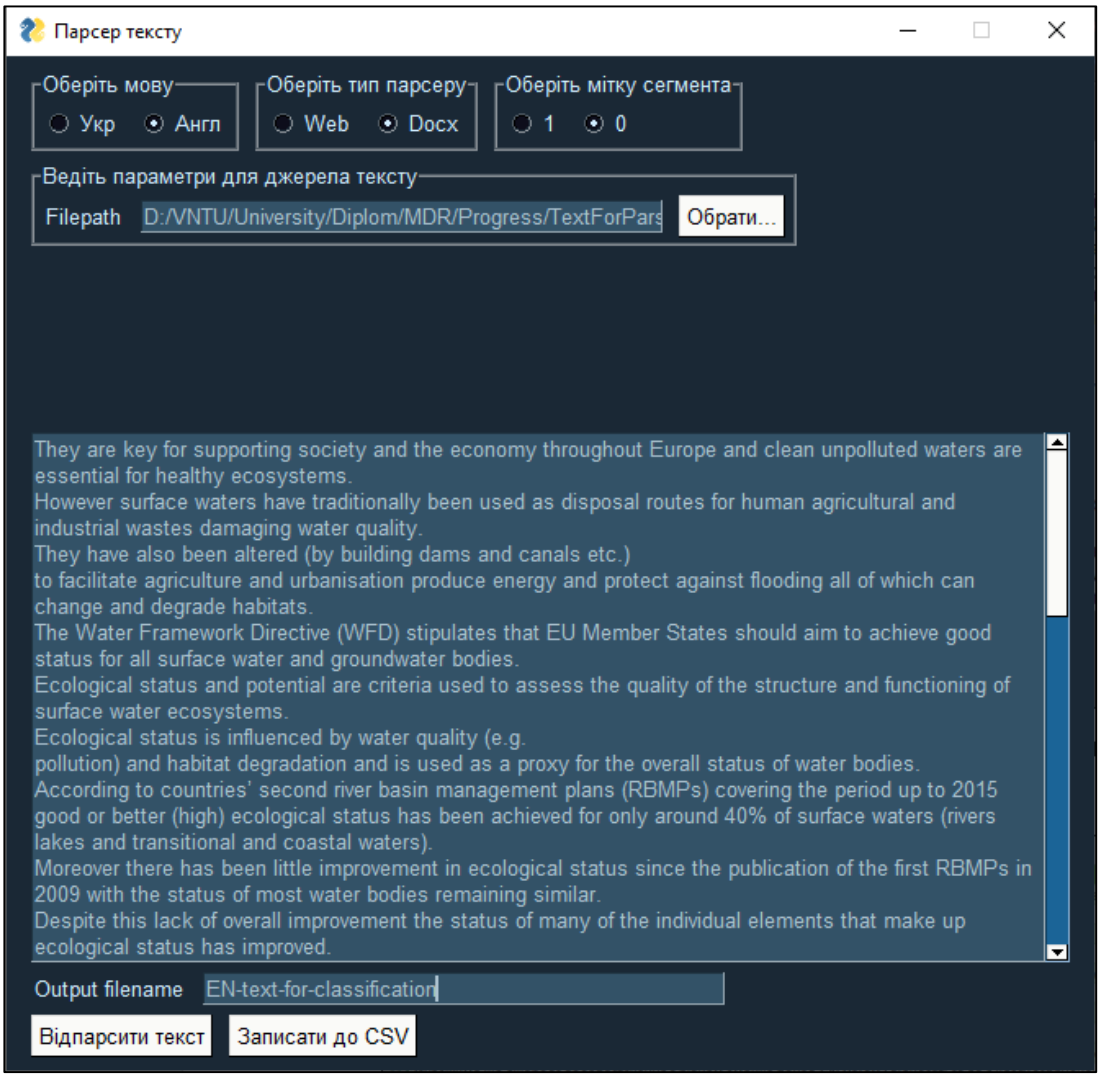

Рисунок 3.6 – Інтерфейс програми із заповненими даними

	text	env_problems	pollution	treatment	climate	biomonitoring
2	They are key for supporting society and the economy throughout Europe and clean unpollut	0	0	0	0	0
3	However surface waters have traditionally been used as disposal routes for human agricultu	0	0	0	0	0
4	They have also been altered (by building dams and canals etc.)	0	0	0	0	0
5	to facilitate agriculture and urbanisation produce energy and protect against flooding all of v	0	0	0	0	0
6	The Water Framework Directive (WFD)B stipulates that EU Member States should aim to ach	0	0	0	0	0
7	Ecological status and potential are criteria used to assess the quality of the structure and fun	0	0	0	0	0
8	Ecological status is influenced by water quality (e.g.	0	0	0	0	0
9	pollution) and habitat degradation and is used as a proxy for the overall status of water bodi	0	0	0	0	0
10	According to countriesвЂ™ second river basin management plans (RBMPs) covering the peri	0	0	0	0	0
11	Moreover there has been little improvement in ecological status since the publication of the	0	0	0	0	0
12	Despite this lack of overall improvement the status of many of the individual elements that	0	0	0	0	0
13	The main pressures on surface water bodies are pollution from point (e.g.	0	0	0	0	0
14	agriculture) sources and various hydromorphological pressures such as barriers (dams) and l	0	0	0	0	0
15	To improve and restore the approximately 60% of surface waters in the EU that still have les	0	0	0	0	0
16	Percentage of water bodies not in good ecological status or potential per river basic district	0	0	0	0	0
17	moderate poor or bad) ecological status varies between river basin districts (RBDs) across Eu	0	0	0	0	0
18	Surface water bodies in north-western Europe have the lowest status.	0	0	0	0	0
19	In Belgium (Flanders) northern Germany and the Netherlands the ecological status of more	0	0	0	0	0
20	Other problem areas include Czechia southern England (United Kingdom) northern France s	0	0	0	0	0
21	Northern countries particularly the northern Scandinavian region Iceland Ireland and Scotlar	0	0	0	0	0

Рисунок 3.7 – Згенерований csv файл

При розробці парсера, були використані бібліотеки, які забезпечують правильний поділ тексту на речення, і очищають його від зайвих символів, що значно спрощує роботу по токенізації речень для BERT і допомагає уникнути помилок при виконані цих процесів [15] [16].

3.3 Застосування технології до звітності Єврокомісії по стану водних ресурсів

Для того щоб протестувати розроблену технологію на практиці, були підібрані дані з сайтів Єврокомісій, де описуються екологічні стани водних ресурсів Європи. Згідно алгоритму, перш за все підготуємо датасет для роботи з ним.

Для парсингу використаємо програму створену й описану в розділі 3.2, а саме метод парсингу з docx файлу. Для початку зберемо усі дані до одного файлу (рис. 3.8).

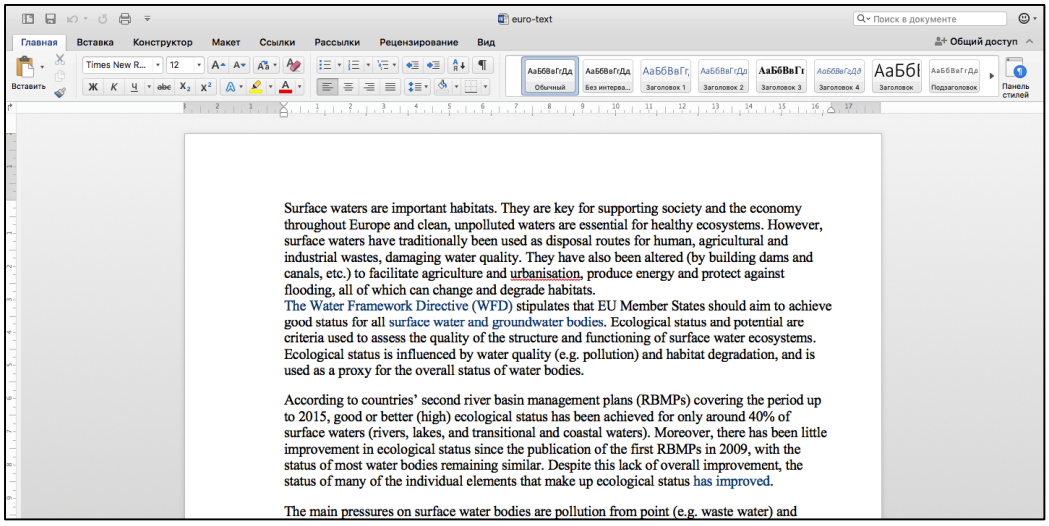

Рисунок 3.8 – Docx файл з даними для парсингу

Тепер завантажимо цей файл до програми і пропустимо через парсер й запишемо результат до файлу «euro-text-for-classification» (рис. 3.9).

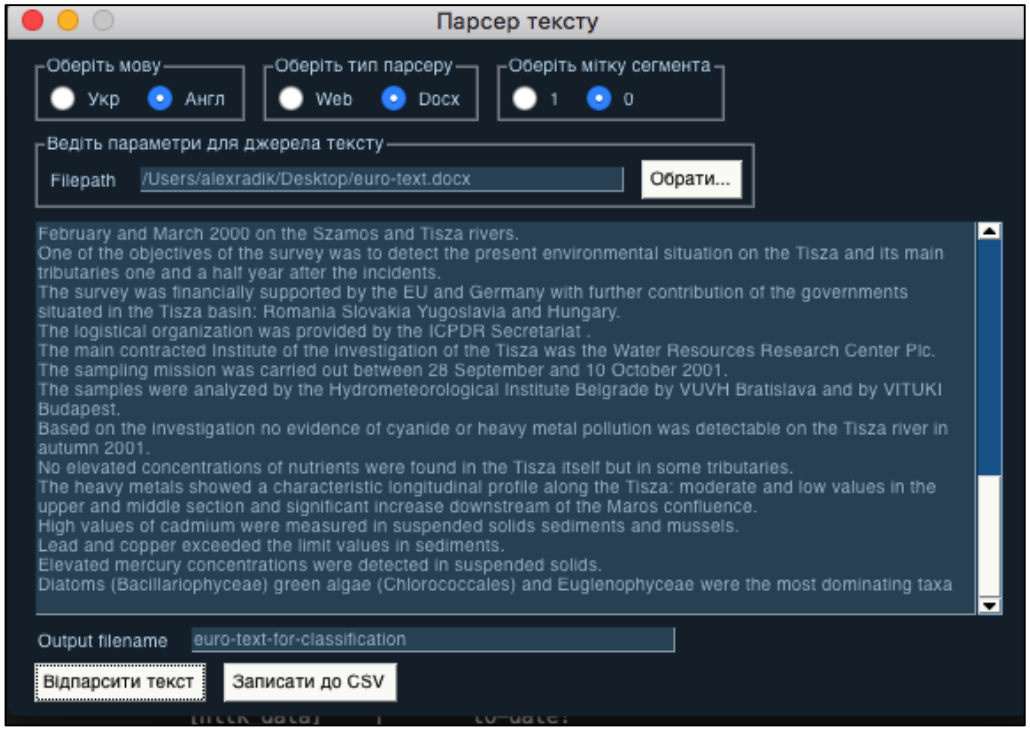

Рисунок 3.9 – Створення датасету за допомогою парсера

Після проведення усіх операцій описаних вище, матимемо готовий датасет для класифікації предтренованими моделями, обраними в розділі 2.3 (рис. 3.10).

	A	B	C	D	E	F
1	text	env_problem	pollution	treatment	climate	biomonitoring
2	They are key for supporting society and the eccnomy throu	0	0	0	0	0
3	However surface waters have traditionally beeı used as di	0	0	0	0	0
4	They have also been altered (by building dams and canals	0	0	0	0	0
5	to facilitate agriculture and urbanisation produce energy a	0	0	0	0	0
6	The Water Framework Directive (WFD)¬†stipuates that El	0	0	0	0	0
7	Ecological status and potential are criteria used to assess t	0	0	0	0	0
8	Ecological status is influenced by water quality (e.g.	0	0	0	0	0
9	pollution) and habitat degradation and is used as a proxy f	0	0	0	0	0
10	According to countriesвАЩ second river basin managemer	0	0	0	0	0
11	Moreover there has been little improvement in ecological	0	0	0	0	0
12	Despite this lack of overall improvement the status of mar	0	0	0	0	0
13	The main pressures on surface water bodies are pollution	0	0	0	0	0
14	agriculture) sources and various hydromorphological press	0	0	0	0	0
15	To improve and restore the approximately 60% of surface	0	0	0	0	0
16	The percentage of water bodies with less than good (i.e.	0	0	0	0	0
17	moderate poor or bad) ecological status varies between riv	0	0	0	0	0
18	Surface water bodies in north-western Europe have the lov	0	0	0	0	0
19	In Belgium (Flanders) northern Germany and the Netherla	0	0	0	0	0
20	Other problem areas include Czechia southern England (Ur	0	0	0	0	0
21	Northern countries particularly the northern Scandinavian	0	0	0	0	0
22	The investigation of the Tisza River was a follow-up projec	0	0	0	0	0
23	The objective of the survey was to investigate¬†the water	0	0	0	0	0
24	The countries participating at the survey include Romania	0	0	0	0	0

Рисунок 3.10 – Створений датасет для класифікації

Тепер застосуємо BERT та предтреновані моделі, вибрані у підрозділі 2.3: лінійну регресію та випадкові ліси (рис. 3.11, 3.12).

	text	env_problems	pollution	treatment	climate	biomonitoring
0	They are key for supporting society and the economy throughout Europe and clean unpolluted waters are essential for healthy ecosystems.	1	0	1	1	1
1	However surface waters have traditionally been used as disposal routes for human agricultural and industrial wastes damaging water quality.	1	1	1	1	0
2	They have also been altered (by building dams and canals etc.)	1	1	1	0	0
3	to facilitate agriculture and urbanisation produce energy and protect against flooding all of which can change and degrade habitats.	1	1	0	1	1
4	The Water Framework Directive (WFD) stipulates that EU Member States should aim to achieve good status for all surface water and groundwater bodies.	0	0	0	1	1
5	Ecological status and potential are criteria used to assess the quality of the structure and functioning of surface water ecosystems.	0	0	1	0	1
6	Ecological status is influenced by water quality (e.g.	1	1	1	0	1
7	pollution) and habitat degradation and is used as a proxy for the overall status of water bodies.	1	1	1	1	0

Рисунок 3.11 – Класифікація за допомогою моделі лінійної регресії

The result of classification using the model random-forest

	text	env_problems	pollution	treatment	climate	biomonitoring
0	They are key for supporting society and the economy throughout Europe and clean unpolluted waters are essential for healthy ecosystems.	1	0	0	0	0
1	However surface waters have traditionally been used as disposal routes for human agricultural and industrial wastes damaging water quality.	1	0	0	0	0
2	They have also been altered (by building dams and canals etc.)	1	0	0	0	0
3	to facilitate agriculture and urbanisation produce energy and protect against flooding all of which can change and degrade habitats.	1	0	0	0	0
4	The Water Framework Directive (WFD) stipulates that EU Member States should aim to achieve good status for all surface water and groundwater bodies.	1	0	0	0	0
5	Ecological status and potential are criteria used to assess the quality of the structure and functioning of surface water ecosystems.	1	0	0	0	0
6	Ecological status is influenced by water quality (e.g.	0	0	0	0	0
7	pollution) and habitat degradation and is used as a proxy for the overall status of water bodies.	1	0	0	0	0
8	According to countries' second river basin management plans (RBMPs) covering the period up to 2015 good or better (high) ecological status has been achieved for only around 40% of surface waters (rivers lakes and transitional and coastal waters).	1	0	0	0	0
9	Moreover there has been little improvement in ecological status since the publication of the first RBMPs in 2009 with the status of most water bodies remaining similar.	1	0	0	0	0
10	Despite this lack of overall improvement the status of many of the individual elements that make up ecological status has improved.	1	0	0	0	0

Рисунок 3.12 – Класифікація за допомогою моделі випадкові ліси (Random Forest)

Проаналізувавши рисунки наведені вище, можемо зробити висновок, що модель випадкових лісів (Random Forest) видала більш адекватний результат.

3.4 Застосування технології до українсько-шведського звіту про стан Південного Бугу

Для того, щоб додатково перевірити точність класифікації моделі випадкових лісів (Random Forest), обрану у підрозділі 3.3, підготуємо ще один датасет з монографії українсько-шведського звіту, який міститиме більше даних й відповідно більше шуму.

Відповідно усіх процесів, які вже описувались вище, підготували docx файл та пропустили його через парсер, щоб отримати готовий датасет, який можемо передати в роботу для моделі класифікації (рис. 3.13).

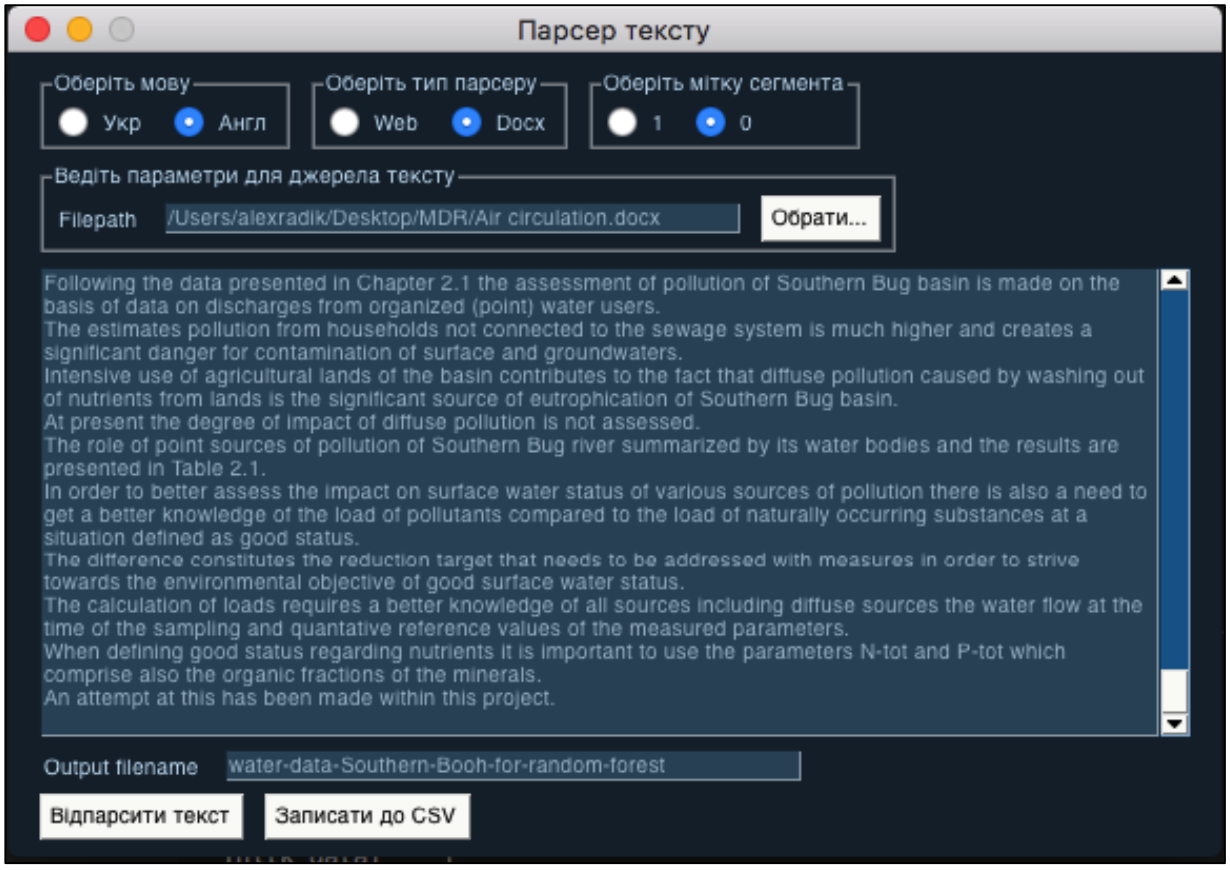

Рисунок 3.13 – Створення додаткового датасету

Після роботи парсера отримуємо датасет більш ніж на 200 елементів, готових до класифікації.

Тепер приступимо до класифікації створеного датасету, перш за все, імпортуємо бібліотеки, необхідні для роботи (рис. 3.14).

```
import numpy as np
import pandas as pd
import pickle

import matplotlib
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# models
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics

# NLP
import torch
import transformers as ppb

import warnings
warnings.filterwarnings('ignore')
```

Рисунок 3.14 – Імпортовані бібліотеки

Тепер встановимо модель, за допомогою якої будемо класифікувати датасет (рис. 3.15).

```
# Model choice
model_name = "random-forest"
# model_name = "linear-regression"
```

Рисунок 3.15 – Вибір моделі

Завантажимо треновану модель з попереднього ноутбука, відповідно вибору (рис. 3.16).

```
# Get notebook_models name
if model_name=="random-forest":
    notebook_models = "nlp-for-ua-bert-cls-with-rf"
elif model_name=="linear-regression":
    notebook_models = "nlp-for-ua-bert-cls-with-linreg"
```

Рисунок 3.16 – Вибір моделі з ноутбуку

Встановимо параметри для імпорту датасету (рис. 3.17).

```
# Set parameters
random_state = 0
notebook_data = "euroreportonwaterresources"
data_name = "water-data-Southern-Booh-for-random-forest.csv"
```

Рисунок 3.17 – Встановлення параметрів імпорту датасету

Завантажимо датасет по заданих параметрах (рис. 3.18, 3.19).

```
df = pd.read_csv(f'/kaggle/input/{notebook_data}/{data_name}', encoding='cp1251')
df = df.fillna(0)

convert_dict = {'text': str,
                'env_problems': int,
                'pollution': int,
                'treatment': int,
                'climate': int,
                'biomonitoring': int}

df = df.astype(convert_dict)
df
```

Рисунок 3.18 – Завантаження датасету

	text	env_problems	pollution	treatment	climate	biomonitoring
0	Air circulation with which air displacement from Atlantic Arctic and Mediterranean is connected plays an important role in climate formation in the Southern Bug basin.	0	0	0	0	0
1	In the upper and middle parts of the basin the climate is moderately continental.	0	0	0	0	0
2	Climate of southern districts is influenced by the Black Sea and in the downstream of the river gradually changes to the drought climate.	0	0	0	0	0
3	Considerable size of the basin from the north-west to the south-east causes noticeable differences in distribution of air temperature.	0	0	0	0	0
4	Average annual air temperature varies from 7.1o C to 10.0o C. Average multi-annual air temperature in the upper and middle parts of the basin varies within 7.1-8.1o C. Maximum temperature in summer reaches 39o C minimum temperature in cold winter days – up to -38o C. For this part of the basin winter season is characterized by precipitations in the form of snow and frequent fogs.	0	0	0	0	0
...	...	...	...	...	...	...
196	In order to better assess the impact on surface water status of various sources of pollution there is also a need to get a better knowledge of the load of pollutants compared to the load of naturally occurring substances at a situation defined as good status.	0	0	0	0	0

Рисунок 3.19 – Загальний вигляд датасету

Виведемо детальну інформацію про датасет разом з усіма типами даних й таргетами (рис. 3.20).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 201 entries, 0 to 200
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   text           201 non-null    object
 1   env_problems   201 non-null    int64 
 2   pollution      201 non-null    int64 
 3   treatment      201 non-null    int64 
 4   climate        201 non-null    int64 
 5   biomonitoring  201 non-null    int64 
dtypes: int64(5), object(1)
memory usage: 9.5+ KB
```

Рисунок 3.20 –Інформація про датасет

Після підготовки датасету, завантажимо попередньо треновану модель DistilBERT й токенізуємо речення, тобто зведемо дані до необхідного для BERT вигляду, далі на основі токенізованих даних і попередньо тренованої моделі DistilBERT створюємо модель необхідної для вирішення нашого завдання (рис. 3.21).

```
def BERT_modeling(df, text='text'):
    # Using DistilBERT for the dataframe df[text]

    model_class, tokenizer_class, pretrained_weights = (ppb.DistilBertModel, ppb.DistilBertTokenizer, 'distilbert-base-uncased')
    # Other models: https://huggingface.co/transformers/pretrained_models.html

    # Load pretrained model/tokenizer
    tokenizer = tokenizer_class.from_pretrained(pretrained_weights)
    model = model_class.from_pretrained(pretrained_weights)

    # Tokenization the sentences - break them up into word and subwords in the format BERT is comfortable with
    tokenized = df[text].apply((lambda x: tokenizer.encode(x, add_special_tokens=True)))

    max_len = 0
    for i in tokenized.values:
        if len(i) > max_len:
            max_len = len(i)

    padded = np.array([i + [0]*(max_len-len(i)) for i in tokenized.values])

    # Creation variable to ignore (mask) the data padding
    attention_mask = np.where(padded != 0, 1, 0)

    # Modeling
    input_ids = torch.tensor(padded).to(torch.int64)
    attention_mask = torch.tensor(attention_mask).to(torch.int64)

    with torch.no_grad():
        last_hidden_states = model(input_ids, attention_mask=attention_mask)

    # Last hidden states
    features = last_hidden_states[0][:,0,:].numpy()

    return features

features = BERT_modeling(df, 'text')
```

Рисунок 3.21 – Робота із BERT

Тепер, викличемо функцію для класифікації даних за заданими таргетами (рис. 3.22).

```
def model_prediction(df, test_features, target):
    # Models training and data prediction for all models from DataFrame models

    # Target
    labels = df[target]

    # Model loading
    with open(f'/kaggle/input/{notebook_models}/model-{model_name}-{target}.pkl','rb') as f:
        model = pickle.load(f)

    # Prediction
    test_pred = model.predict(test_features)
    test_pred = [int(x>0) for x in test_pred]

    return test_pred
```

Рисунок 3.22 – Класифікація даних по таргетам

Вкажемо список таргетів для роботи (рис. 3.23).

```
# List of the target features in df
cols = df.columns.tolist()[1:]
print('Target columns:', cols)
```

```
Target columns: ['env_problems', 'pollution', 'treatment', 'climate', 'biomonitoring']
```

Рисунок 3.23 – Таргети для роботи

Запуск функції для класифікації речень із датасету (3.24).

```
# Solving NLP Classification tasks
print('Solving NLP Classification tasks')
for col in cols:
    df[col] = model_prediction(df, features, col)
```

Рисунок 3.24 – Класифікація речень

Після класифікації отримуємо опрацьований датасет (3.25, 3.26).

```
# Results display
pd.set_option('max_colwidth',1000)
df.to_csv(f'{notebook_data}-{notebook_models}-classified.csv', index=False)
print(f"The result of classification using the model {model_name}")
display(df)
```

```
The result of classification using the model random-forest
```

Рисунок 3.25 – Код для візуалізації результату класифікації

	text	env_problems	pollution	treatment	climate	biomonitoring
0	Air circulation with which air displacement from Atlantic Arctic and Mediterranean is connected plays an important role in climate formation in the Southern Bug basin.	0	0	0	0	0
1	In the upper and middle parts of the basin the climate is moderately continental.	0	0	0	1	0
2	Climate of southern districts is influenced by the Black Sea and in the downstream of the river gradually changes to the drought climate.	0	0	0	1	0
3	Considerable size of the basin from the north-west to the south-east causes noticeable differences in distribution of air temperature.	0	0	0	1	0
4	Average annual air temperature varies from 7.1o C to 10.0o C. Average multi-annual air temperature in the upper and middle parts of the basin varies within 7.1-8.1o C. Maximum temperature in summer reaches 39o C minimum temperature in cold winter days – up to -38o C. For this part of the basin winter season is characterized by precipitations in the form of snow and frequent fogs.	0	0	0	1	0
...	...	...	...	...	...	...
196	In order to better assess the impact on surface water status of various sources of pollution there is also a need to get a better knowledge of the load of pollutants compared to the load of naturally occurring substances at a situation defined as good status.	1	0	0	0	0
197	The difference constitutes the reduction target that needs to be addressed with measures in order to strive towards the environmental objective of good surface water status.	0	0	0	0	0
198	The calculation of loads requires a better knowledge of all sources including diffuse sources the water flow at the time of the sampling and quantative reference values of the measured parameters.	1	1	0	0	0

Рисунок 3.26 – Класифікований датасет з використанням моделі випадкових лісів

Як видно на рис. 3.26, результат є задовільним, отже створена інформаційна технологія довела свою ефективність.

3.5 Висновки

Запропоновано архітектуру інформаційної технології, яка розробляється, та описано алгоритм її роботи.

Описано процес формування датасету із веб-ресурсів й парсингу docx файлів, які містять інформацію про екологічні проблеми річок і водойм басейну р. Південний Буг та звітності Єврокомісії.

Реалізовано класифікацію тестового датасету за допомогою моделей лінійної регресії та випадкових лісів. Як результат маємо, що модель випадкових лісів (Random Forest) краще впоралась із поставленим завданням.

Після класифікації на тестовому датасеті реалізовано класифікацію даних із доступної звітності Єврокомісії по стану водних ресурсів. Класифікація відбувалась за допомогою передтренованої моделі машинного навчання, яка показала кращий результат на попередньому етапі, тобто – модель випадкових лісів (Random Forest). Модель показала гарні результати, як на тестовому датасеті, так і на тренувальному, що означає, що інформаційна технологія виконує своє завдання успішно.

4 ЕКОНОМІЧНА ЧАСТИНА

4.1 Комерційний та технологічний аудит науково-технічної розробки

Метою даного розділу є проведення технологічного аудиту, в даному випадку нової інформаційної NLP-технології класифікації англомовної інформації про стан водних ресурсів. Особливістю розробки є підвищення точності класифікації текстової англомовної інформації про масиви вод басейну р. Південний Буг з використанням методів машинного навчання у порівнянні з аналогами. Розроблювана технологія матиме високий показник точності автоматичної класифікації.

За аналоги були взяті спам системи, в основі яких лежить також NLP, яке аналізує текст (вхідні листи) й класифікує їх чи спам це чи ні. Орієнтовна вартість такої системи, для комерційного сегменту, складає близько 50000$.

Для проведення комерційного та технологічного аудиту залучають не менше 3-х незалежних експертів. Оцінювання науково-технічного рівня розробки та її комерційного потенціалу рекомендується здійснювати із застосуванням п’ятибальної системи оцінювання за 12-ма критеріями, у відповідності із таблиці 4.1 [17].

Таблиця 4.1 – Рекомендовані критерії оцінювання комерційного потенціалу розробки та їх можлива бальна оцінка

Бали (за 5-ти бальною шкалою)					
Кри-терій	0	1	2	3	4
Технічна здійсненність концепції					
1	Достовірність концепції не підтверджена	Концепція підтверджена експертними висновками	Концепція підтверджена розрахунками	Концепція перевірена на практиці	Перевірено працездатність продукту в реальних умовах

Продовження табл. 4.1

Ринкові переваги					
2	Багато аналогів на малому ринку	Мало аналогів на малому ринку	Кілька аналогів на великому ринку	Один аналог на великому ринку	Продукт не має аналогів на великому ринку
3	Ціна продукту значно вища за ціни аналогів	Ціна продукту дещо вища за ціни аналогів	Ціна продукту приблизно дорівнює цінам аналогів	Ціна продукту дещо нижче за ціни аналогів	Ціна продукту значно нижче за ціни аналогів
4	Технічні та споживчі властивості продукту значно гірші, ніж в аналогів	Технічні та споживчі властивості продукту трохи гірші, ніж в аналогів	Технічні та споживчі властивості продукту на рівні аналогів	Технічні та споживчі властивості продукту трохи кращі, ніж в аналогів	Технічні та споживчі властивості продукту значно кращі, ніж в аналогів
5	Експлуатаційні витрати значно вищі, ніж в аналогів	Експлуатаційні витрати дещо вищі, ніж в аналогів	Експлуатаційні витрати на рівні експлуатаційних витрат аналогів	Експлуатаційні витрати трохи нижчі, ніж в аналогів	Експлуатаційні витрати значно нижчі, ніж в аналогів
Ринкові перспективи					
6	Ринок малий і не має позитивної динаміки	Ринок малий, але має позитивну динаміку	Середній ринок з позитивною динамікою	Великий стабільний ринок	Великий ринок з позитивною динамікою
7	Активна конкуренція великих компаній на ринку	Активна конкуренція	Помірна конкуренція	Незначна конкуренція	Конкурентів немає
Практик на здійсненність					
8	Відсутні фахівці як з технічної, так і з комерційної реалізації ідеї	Необхідно наймати фахівців або витрачати значні кошти та час на навчання наявних фахівців	Необхідне незначне навчання фахівців та збільшення їх штату	Необхідне незначне навчання фахівців	Є фахівці з питань як з технічної, так і з комерційної реалізації ідеї
9	Потрібні значні фінансові ресурси, які відсутні. Джерела фінансування ідеї відсутні	Потрібні незначні фінансові ресурси. Джерела фінансування відсутні	Потрібні значні фінансові ресурси. Джерела фінансування є	Потрібні незначні фінансові ресурси. Джерела фінансування є	Не потребує додаткового фінансування
10	Необхідна розробка нових матеріалів	Потрібні матеріали, що використовуються у військово-промисловому комплексі	Потрібні дорогі матеріали	Потрібні досяжні та дешеві матеріали	Всі матеріали для реалізації ідеї відомі та давно використовуються у виробництві

Продовження табл. 4.1

11	Термін реалізації ідеї більший за 10 років	Термін реалізації ідеї більший за 5 років. Термін окупності інвестицій більше 10-ти років	Термін реалізації ідеї від 3-х до 5-ти років. Термін окупності інвестицій більше 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій від 3-х до 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій менше 3-х років
12	Необхідна розробка регламентних документів та отримання великої кількості дозвільних документів на виробництво та реалізацію продукту	Необхідно отримання великої кількості дозвільних документів на виробництво та реалізацію продукту, що вимагає значних коштів та часу	Процедура отримання дозвільних документів для виробництва та реалізації продукту вимагає незначних коштів та часу	Необхідно тільки повідомлення відповідним органам про виробництво та реалізацію продукту	Відсутні будь-які регламентні обмеження на виробництво та реалізацію продукту

Усі дані по кожному параметру занесено в таблиці 4.2

Таблиця 4.2 – Результати оцінювання комерційного потенціалу розробки

Критерії оцінювання	ПІБ експертів		
	Експерт 1	Експерт 2	Експерт 3
	Бали		
Технічна здійсненність концепції	3	4	4
Наявність аналогів на ринку	3	3	4
Цінова політика	4	4	4
Технічні та споживчі властивості виробу	4	3	4
Експлуатаційні витрати	4	4	3
Ринок збуту	4	3	4
Конкурентоспроможність	3	4	3
Фахівці з технічної і комерційної реалізації	4	3	4
Фінансування	4	4	3
Матеріально-технічна база	3	3	3
Термін реалізації ідеї	4	4	4
Супровідна документація	2	3	3
Сума	42	42	43
Середньоарифметична сума балів	(42+42+43) / 3 = 42,33		

За даними таблиці 4.2 можна зробити висновок щодо рівня комерційного потенціалу даної розробки. Для цього доцільно скористатись рекомендаціями, наведеними в таблиці 4.3 [17].

Таблиця 4.3 - Рівні комерційного потенціалу розробки

Середньоарифметична сума балів, розрахована на основі висновків експертів	Рівень комерційного потенціалу розробки
0-10	Низький
11-20	Нижче середнього
21-30	Середній
31-40	Вище середнього
41-48	Високий

Як видно з таблиці, рівень комерційного потенціалу розроблюваного нового програмного продукту є високим, що досягається за рахунок того, що інформаційної NLP-технології класифікації англомовної інформації про стан водних ресурсів має вищу точність класифікації текстової англомовної інформації про масиви вод басейну р. Південний Буг з використанням методів машинного навчання у порівнянні з аналогами. Розроблювана технологія матиме високий показник точності автоматичної класифікації [17].

4.2 Прогнозування витрат на виконання науково-дослідної (дослідно-конструкторської) роботи

Основна заробітна плата розробників, яка розраховується за формулою:

$$З_о = \frac{M}{T_р} \cdot t \,, \tag{4.1}$$

де М – місячний посадовий оклад конкретного розробника (дослідника), грн;

$T_р$ – число робочих днів за місяць, 22 днів;

t – число днів роботи розробника (дослідника).

Результати розрахунків зведемо до таблиці 4.4.

Таблиця 4.4 – Основна заробітна плата розробників

Найменування посади	Місячний посадовий оклад, грн.	Оплата за робочий день, грн.	Число днів роботи	Витрати на заробітну плату, грн.
Керівник проекту	39000	1772,73	43	76227,273
Програміст	35000	1590,91	43	68409,091
Всього				144636,36

Так як в даному випадку розробляється програмний продукт, то розробник виступає одночасно і основним робітником, і тестувальником розроблюваного програмного продукту.

Додаткова заробітна плата розробників, які брати участь в розробці обладнання/програмного продукту[17].

Додаткову заробітну плату прийнято розраховувати як 10 % від основної заробітної плати розробників та робітників:

$$З_д = З_о \cdot 10\ \% / 100\ \% . \quad (4.2)$$

$З_д = (144636,36 \cdot 10\ \% / 100\ \%) = 14463,64$ (грн).

Згідно діючого законодавства нарахування на заробітну плату складають 22 % від суми основної та додаткової заробітної плати.

$$Н_з = (З_о + З_д) \cdot 22\ \% / 100\% . \quad (4.3)$$

$Н_з = (144636,36 + 14463,64) \cdot 22\ \% / 100\ \% = 35002,00$ (грн).

Оскільки для розроблювального пристрою не потрібно витрачати матеріали та комплектуючи, то витрати на матеріали і комплектуючі дорівнюють нулю.

Амортизація обладнання, що використовувалось для розробки в спрощеному вигляді розраховується за формулою:

$$A = \frac{Ц}{Тв} \cdot \frac{t_{вик}}{12} \text{ [грн]}, \quad (4.4)$$

де Ц – балансова вартість обладнання, грн.;

Т – термін корисного використання обладнання згідно податкового законодавства, років;

$t_{вик}$ – термін використання під час розробки, місяців.

Розрахуємо, для прикладу, амортизаційні витрати на комп'ютер балансова вартість якого становить 25000 грн., термін його корисного використання згідно податкового законодавства – 2 роки, а термін його фактичного використання – 1,95 міс.

$$A_{обл} = \frac{2500}{2} * \frac{1,95}{12} = 2035,99 \text{ (грн)}. \quad (4.5)$$

Аналогічно визначаємо амортизаційні витрати на інше обладнання та приміщення. Розрахунки заносимо до таблиці 4.5 [17].

Так як вартість ліцензійної ОС та спеціалізованих ліцензійних нематеріальних активів є безкоштовною, то $B_{нем.ак.}$ = 0 грн.

Таблиця 4.5 – Амортизаційні відрахування на матеріальні та нематеріальні ресурси для розробників

Найменування обладнання	Балансова вартість, грн.	Строк корисного використання, років	Термін використання обладнання, місяців	Амортиза-ційні відрахуван-ня, грн.
Комп'ютер та комп'ютерна периферія (ACER Aspire 5 A515-45 (NX.A82EU.00Z))	25000	2	1,95	2035,985
Офісне обладнання (меблі)	23000	4	1,95	936,553

Продовження табл. 4.5

Приміщення	980000	20	1,95	7981,061
Всього				10953,60

Тарифи на електроенергію для непобутових споживачів (промислових підприємств) відрізняються від тарифів на електроенергію для населення. При цьому тарифи на розподіл електроенергії у різних постачальників (енергорозподільних компаній), будуть різними. Крім того, розмір тарифу залежить від класу напруги (1-й або 2-й клас). Тарифи на розподіл електроенергії для всіх енергорозподільних компаній встановлює Національна комісія з регулювання енергетики і комунальних послуг (НКРЕКП). Витрати на силову електроенергію розраховуються за формулою:

$$В_е = В \cdot П \cdot Ф \cdot К_п, \quad (4.6)$$

де В – вартість 1 кВт-години електроенергії для 1 класу підприємства, В = 6,2 грн./кВт;

П – встановлена потужність обладнання, кВт. П = 0,4 кВт;

Ф – фактична кількість годин роботи обладнання, годин;

$К_п$ – коефіцієнт використання потужності, $К_п$ = 0,9.

$В_е = 0,9 \cdot 0,4 \cdot 8 \cdot 43 \cdot 6,2 = 767,808$ (грн).

До статті «Інші витрати» належать витрати, які не знайшли відображення у зазначених статтях витрат і можуть бути віднесені безпосередньо на собівартість досліджень за прямими ознаками. Витрати за статтею «Інші витрати» розраховуються як 50…100% від суми основної заробітної плати дослідників:

$$I_в = (З_о + З_р) \cdot \frac{Н_{ів}}{100\%}, \quad (4.7)$$

де $Н_{ів}$ – норма нарахування за статтею «Інші витрати».

$І_в$ = 144636,36 * 70% / 100% = 101245,5 (грн).

До статті «Накладні (загальновиробничі) витрати» належать: витрати, пов'язані з управлінням організацією; витрати на винахідництво та раціоналізацію; витрати на підготовку (перепідготовку) та навчання кадрів; витрати, пов'язані з набором робочої сили; витрати на оплату послуг банків; витрати, пов'язані з освоєнням виробництва продукції; витрати на науково-технічну інформацію та рекламу та ін. Витрати за статтею «Накладні (загальновиробничі) витрати» розраховуються як 100…150% від суми основної заробітної плати дослідників:

$$Н_{нзв} = (З_о + З_р) \cdot \frac{Н_{нзв}}{100\%}, \quad (4.8)$$

де $Н_{нзв}$ – норма нарахування за статтею «Накладні (загальновиробничі) витрати».

$Н_{нзв}$ = 144636,36 * 130 % / 100 % = 188027 (грн).

Сума всіх попередніх статей витрат дає загальні витрати на проведення науково-дослідної роботи:

$В_{заг}$ = 144636,36+14463,64+35002,00+10953,60+767,81+101245,5+
+188027 = 495096,13 (грн).

Розрахунок загальних витрат на науково-дослідну (науково-технічну) роботу та оформлення її результатів [17].

Загальні витрати на завершення науково-дослідної (науково-технічної) роботи та оформлення її результатів розраховуються за формулою:

$$ЗВ = \frac{B_{заг}}{\eta} \text{ [грн]}, \quad (4.9)$$

де η – коефіцієнт, який характеризує етап (стадію) виконання науково-дослідної роботи.

Так, якщо науково-технічна розробка знаходиться на стадії: науково-дослідних робіт, то η=0,1; технічного проектування, то η=0,2; розробки конструкторської документації, то η=0,3; розробки технологій, то η=0,4; розробки дослідного зразка, то η=0,5; розробки промислового зразка, то η=0,7; впровадження, то η=0,9. Оберемо η = 0,5, так як розробка, на даний момент, знаходиться на стадії дослідного зразка:

$ЗВ$ = 495096,13 / 0,5 = 990192 (грн).

4.3 Розрахунок економічної ефективності науково-технічної розробки за її можливої комерціалізації потенційним інвестором

В ринкових умовах узагальнювальним позитивним результатом, що його може отримати потенційний інвестор від можливого впровадження результатів тієї чи іншої науково-технічної розробки, є збільшення у потенційного інвестора величини чистого прибутку. Саме зростання чистого прибутку забезпечить потенційному інвестору надходження додаткових коштів, дозволить покращити фінансові результати його діяльності, підвищить конкурентоспроможність та може позитивно вплинути на ухвалення рішення щодо комерціалізації цієї розробки.

Для того, щоб розрахувати можливе зростання чистого прибутку у потенційного інвестора від можливого впровадження науково-технічної розробки необхідно:

а) вказати, з якого часу можуть бути впроваджені результати науково-технічної розробки;

б) зазначити, протягом скількох років після впровадження цієї науково-технічної розробки очікуються основні позитивні результати для потенційного інвестора (наприклад, протягом 3-х років після її впровадження);

в) кількісно оцінити величину існуючого та майбутнього попиту на цю або аналогічні чи подібні науково-технічні розробки та назвати основних суб'єктів (зацікавлених осіб) цього попиту;

г) визначити ціну реалізації на ринку науково-технічних розробок з аналогічними чи подібними функціями.

При розрахунку економічної ефективності потрібно обов'язково враховувати зміну вартості грошей у часі, оскільки від вкладення інвестицій до отримання прибутку минає чимало часу. При оцінюванні ефективності інноваційних проектів передбачається розрахунок таких важливих показників:

- абсолютного економічного ефекту (чистого дисконтованого доходу);
- внутрішньої економічної дохідності (внутрішньої норми дохідності);
- терміну окупності (дисконтованого терміну окупності).

Аналізуючи напрямки проведення науково-технічних розробок, розрахунок економічної ефективності науково-технічної розробки за її можливої комерціалізації потенційним інвестором можна об'єднати, враховуючи визначені ситуації з відповідними умовами [17].

Розробка чи суттєве вдосконалення програмного засобу (програмного забезпечення, програмного продукту) для використання масовим споживачем.

В цьому випадку майбутній економічний ефект буде формуватися на основі таких даних:

$$\Delta П_i = (\pm \Delta Ц_0 \cdot N + Ц_0 \cdot \Delta N)_i \cdot \lambda \cdot \rho \cdot (1 - \frac{\vartheta}{100}), \quad (4.10)$$

де $\pm\Delta Ц_о$ – зміна вартості програмного продукту (зростання чи зниження) від впровадження результатів науково-технічної розробки в аналізовані періоди часу;

N – кількість споживачів, які використовували аналогічний продукт у році до впровадження результатів нової науково-технічної розробки;

$Ц_о$ – основний оціночний показник, який визначає діяльність підприємства у даному році після впровадження результатів наукової розробки, $Ц_о = Ц_б \pm \Delta Ц_о$;

$Цб$ – вартість програмного продукту у році до впровадження результатів розробки;

ΔN – збільшення кількості споживачів продукту, в аналізовані періоди часу, від покращення його певних характеристик;

λ – коефіцієнт, який враховує сплату податку на додану вартість. Ставка податку на додану вартість дорівнює 20%, а коефіцієнт $\lambda = 0{,}8333$;

ρ – коефіцієнт, який враховує рентабельність продукту;

ϑ – ставка податку на прибуток, у 2022 році ϑ =18%.

Припустимо, що при прогнозованій ціні 85000 грн. за одиницю виробу, термін збільшення прибутку складе 3 роки. Після завершення розробки і її вдосконалення, можна буде підняти її ціну на *5000* грн. Кількість одиниць реалізованої продукції також збільшиться: протягом першого року – на 200 шт., протягом другого року – на 150 шт., протягом третього року на 50 шт. До моменту впровадження результатів наукової розробки реалізації продукту не було:

$\Delta П_1$ = (0*5000 + (85000 + 5000)*200)* 0,8333* 0,35) * (1 - 0,18) = 4065833,171 (грн).

$\Delta П_2$ = (0*5000 + (85000 + 5000)*(200+150)* 0,8333* 0,35) * (1 - 0,18) = 7533749,699 (грн).

$\Delta П_З = (0*5000 + (85000 + 5000)*(200+150+50)* 0{,}8333* 0{,}35) * (1 - 0{,}18) = 8609999{,}656$ (грн).

Отже, комерційний ефект від реалізації результатів розробки за три роки складе 20209582,52 грн.

Розраховуємо приведену вартість збільшення всіх чистих прибутків *ПП*, що їх може отримати потенційний інвестор від можливого впровадження та комерціалізації науково-технічної розробки:

$$ПП = \sum_{1}^{T} \frac{\Delta П_i}{(1+\tau)^t}, \qquad (4.11)$$

де $\Delta П_i$ – збільшення чистого прибутку у кожному із років, протягом яких виявляються результати виконаної та впровадженої науково-дослідної (науково-технічної) роботи, грн;

Т – період часу, протягом якою виявляються результати впровадженої науково-дослідної (науково-технічної) роботи, роки;

τ – ставка дисконтування, за яку можна взяти щорічний прогнозований рівень інфляції в країні, $\tau = 0{,}05...0{,}15$;

t – період часу (в роках).

Збільшення прибутку ми отримаємо, починаючи з першого року:

$$ПП = (4065833{,}171/(1+0{,}1)^1)+(7533749{,}699/(1+0{,}1)^2)+(8609999{,}656/(1+0{,}1)^3) = 3696211{,}97 + 6226239{,}42 + 6468820{,}177 = 16391271{,}57 \text{ (грн)}.$$

Далі розраховують величину початкових інвестицій *PV*, які потенційний інвестор має вкласти для впровадження і комерціалізації науково-технічної розробки. Для цього можна використати формулу:

$$PV = k_{інв} * ЗВ, \qquad (4.12)$$

де $k_{інв}$ – коефіцієнт, що враховує витрати інвестора на впровадження науково-технічної розробки та її комерціалізацію. Це можуть бути витрати на підготовку приміщень, розробку технологій, навчання персоналу, маркетингові заходи тощо; зазвичай $k_{інв}=2...5$, але може бути і більшим;
ЗВ – загальні витрати на проведення науково-технічної розробки та оформлення її результатів, грн.

$PV = 2 * 990192 = 1980384{,}54$ (грн).

Тоді абсолютний економічний ефект $E_{абс}$ або чистий приведений дохід *(NPV, Net Present Value)* для потенційного інвестора від можливого впровадження та комерціалізації науково-технічної розробки становитиме:

$$E_{абс} = ПП - PV, \tag{4.13}$$

$E_{абс} = 16391271{,}57 - 1980384{,}54 = 14410887{,}04$ (грн).

Оскільки $E_{абс} > 0$ то вкладання коштів на виконання та впровадження результатів даної науково-дослідної (науково-технічної) роботи може бути доцільним.

Для остаточного прийняття рішення з цього питання необхідно розрахувати внутрішню економічну дохідність або показник внутрішньої норми дохідності *(IRR, Internal Rate of Return)* вкладених інвестицій та порівняти її з так званою бар'єрною ставкою дисконтування, яка визначає ту мінімальну внутрішню економічну дохідність, нижче якої інвестиції в будь-яку науково-технічну розробку вкладати буде економічно недоцільно [17].

Розрахуємо відносну (щорічну) ефективність вкладених в наукову розробку інвестицій $E_в$. Для цього використаємо формулу:

$$E_в = \sqrt[T_ж]{1 + \frac{E_{абс}}{PV}} - 1, \tag{4.14}$$

$T_ж$ – життєвий цикл наукової розробки, роки.

$E_в$ = з $(1 + \sqrt{14410887,04/1980384,54} - 1$ = 1,023.

Визначимо мінімальну ставку дисконтування, яка у загальному вигляді визначається за формулою:

$$\tau = d + f, \tag{4.15}$$

де d – середньозважена ставка за депозитними операціями в комерційних банках; в 2022 році в Україні d = (0,09…0,14);

f – показник, що характеризує ризикованість вкладень; зазвичай, величина f = (0,05…0,5).

$$\tau_{\min} = 0,14 + 0,05 = 0,19. \tag{4.16}$$

Так як $E_в > \tau_{min}$, то інвестор може бути зацікавлений у фінансуванні даної наукової розробки.

Розрахуємо термін окупності вкладених у реалізацію наукового проекту інвестицій за формулою:

$$T_{ок} = \frac{1}{E_в}, \tag{4.17}$$

$T_{ок}$ = 1 / 1,023 = 0,98 р.

Оскільки $T_{ок}$ < 3-х років, а саме термін окупності рівний 0,98 роки, то фінансування даної наукової розробки є доцільним [17].

4.4 Висновки

Економічна частина даної роботи містить розрахунок витрат на розробку нового програмного продукту, сума яких складає 990192 гривень. Було спрогнозовано орієнтовану величину витрат по кожній з статей витрат. Також розраховано чистий прибуток, який може отримати виробник від реалізації нового технічного рішення, розраховано період окупності витрат для інвестора та економічний ефект при використанні даної розробки. В результаті аналізу розрахунків можна зробити висновок, що розроблений програмний продукт за ціною дешевший за аналог і є високо конкурентоспроможним. Період окупності складе близько 0,98 років.

ВИСНОВКИ

Дана магістерська робота присвячена розробці інтелектуальної інформаційної технології класифікації англомовної текстової інформації про масиви вод басейну річки Південний Буг. В ній описано загальну характеристику об’єкту досліджень. Здійснено постановку задачі та обрано оптимальні інформаційні технології для її вирішення.

Розглянуто важливість створення інтелектуальної інформаційної технології, яка має пришвидшити класифікацію інформації. Також, розглянуто методи вирішення подібних задач, зокрема – методи і технології оброблення природномовних даних (NLP) для перетворення тексту на ембеддинги, а також – моделі машинного навчання, які використовуються для подальшої класифікації тексту за цими ембеддингами.

Розглянуто етапи створення та аналізу датесету і застосування до нього NLP-моделей та моделей машинного навчання. Для роботи з цими технологіями необхідним є створення датасету із відповідним набором даних. Для полегшення цього завдання розроблено та описано автоматичний парсер для парсингу даних із різних інформаційних джерел різного формату.

Здійснено розроблення інтелектуальної інформаційної технології автоматичної класифікації англомовної екологічної текстової інформації. Також обрано дві можливі оптимальні моделі. Практичні випробування на даних з текстових даних з України та з Єврокомісії показали, що найкращою з них, тобто - оптимальною, є модель випадкових лісів (Random Forest) має точність класифікації 0.95, що є задовільною точністю як для таких задач.

Економічна частина даної роботи містить розрахунок витрат на розробку нового програмного продукту, сума яких складає 990192 гривень. Було спрогнозовано орієнтовану величину витрат по кожній з статей витрат. Також розраховано чистий прибуток, який може отримати виробник від реалізації нового технічного рішення, розраховано період окупності витрат для інвестора та економічний ефект при використанні даної розробки. В

результаті аналізу розрахунків можна зробити висновок, що розроблений програмний продукт за ціною дешевший за аналог і є високо конкурентоспроможним. Період окупності складе близько 0,98 роки.

За результатами даної роботи помітне підвищення точності класифікації текстових англомовних даних про водні ресурси, які синтезуються автоматично з використанням методів штучного інтелекту. Отже, завдання, поставлене у магістерській кваліфікаційній роботі, виконано в повному обсязі.

За результатами даної роботи написані тези доповіді, які були апробовані на Всеукраїнській науково-практичній інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи» (Вінниця, 2021-2022 рр.) та опубліковані у матеріалах цієї конференції.

Дана магістерська кваліфікаційна робота виконана на замовлення басейнової ради Південного Бугу, куди й передано її результати. Про них вже йшлось в обговоренні до п.4 порядку денного засідання ради (див. протокол № 12 від 07.12.2022 р. на сайті БУВР Південного Бугу).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. В. Б. Мокін, Є. М. Крижановський, М. А. Гораш, Д. В. Пасічнюк, О. В. Радецький, С. М. Хмарук. Інформаційні технології збирання та систематизування даних про стан та проблеми масивів вод на прикладі басейну річки Південний Буг. Молодь в науці: дослідження, проблеми, перспективи (МН-2022). Вінниця, 2022. URL: https://conferences.vntu.edu.ua/index.php/mn/mn2022/paper/viewFile/16236/13677

2. Обробка природної мови (NLP): що це таке і як це працює? URL: https://monkeylearn.com/natural-language-processing/

3. A Visual Guide to Using BERT for the First Time. URL: https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

4. BERT Explained: State of the art language model for NLP. URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

5. Finetune DistilBERT for multi-label text classsification task. URL: https://medium.com/analytics-vidhya/finetune-distilbert-for-multi-label-text-classsification-task-994eb448f94c

6. Top 10 Pre-Trained NLP Language Models for AI Application Building. URL: https://itechnolabs.ca/top-10-pre-trained-nlp-language-models-for-ai-application-building/

7. LINEAR REGRESSION. URL: https://mlu-explain.github.io/linear-regression/

8. Random forests. URL: https://scikit-learn.org/stable/modules/ensemble.html - random-forests

9. Перші кроки в NLP: розглядаємо Python-бібліотеку scikit-learn в реальному завданні. URL: https://dou.ua/lenta/articles/first-steps-in-nlp-scikit-learn/

10. Multi-layer Perceptron. URL: https://scikit-learn.org/stable/modules/neural_networks_supervised.html#multi-layer-perceptron

11. Classification Example with XGBClassifier in Python. URL: https://www.datatechnotes.com/2019/07/classification-example-with.html

12. ER-EnvProbl – NLP: Bag of Words, TF IDF, GloVe. URL: https://www.kaggle.com/code/sasharadeckiy/er-envprobl-nlp-bag-of-words-tf-idf-glove

13. NLP for EN: BERT CLS & 5 Classifiers. URL: https://www.kaggle.com/code/sasharadeckiy/nlp-for-en-bert-cls-5-classifiers

14. Python GUIs for Humans. URL: https://www.pysimplegui.org/en/latest/

15. Natural Language Toolkit. URL: https://www.nltk.org/

16. Tokenize UK. URL:https://github.com/lang-uk/tokenize-uk

17. Методичні вказівки до написання магістерської кваліфікаційної роботи для студентів. URL: http://pdf.lib.vntu.edu.ua/books/2022/Yaremchuk_2018_43.pdf

18. Савченко А.С., Синельніков О.О. Методи та системи штучного інтелекту: навч. посіб. Київ, 2021, 9 с.

19. Грицай Д.І. Розумні роботи. Щомісячний правовий часопис. 2020 № 1/2 С. 104

20. В. Б. Мокін, Є. М. Крижановський. Геоінформаційні системи в екології: навч. посіб / за ред. Є. М. Крижановського. Вінниця : ВНТУ, 2014. 61 с.

21. Мокін В. Б., Ю. Собко, Є. М. Крижановський, М.В. Дратований, Г. В. Горячев. Створення інформаційної системи моніторингу забруднення атмосферного повітря міста на основі технології «Інтернет речей», Вісник ВПІ. 2017. № 3. С. 49-58.

Додаток А

Міністерство освіти і науки України
Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації

ЗАТВЕРДЖУЮ
Завідувач кафедри САІТ
_____ д.т.н., проф. Мокін В. Б.
«_19_» _____09_____ 2022 р.

ТЕХНІЧНЕ ЗАВДАННЯ

на магістерську кваліфікаційну роботу

«ІНФОРМАЦІЙНА NLP-ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ АНГЛОМОВНОЇ ІНФОРМАЦІЇ ПРО СТАН ВОДНИХ РЕСУРСІВ»

08-53.МКР.001.02.000.ТЗ

Керівник: д.т.н., проф. каф. САІТ
_____ Мокін В. Б.
«_19_» _____09_____ 2022 р.

Розробив: студент гр. 2ІСТ-21м
_____ Радецький О.В.
«_19_» _____09_____ 2022 р.

Вінниця 2022

1. Підстава для проведення робіт

Підставою для виконання роботи є наказ № 203 по ВНТУ від «14» _09_ 2022 р., та індивідуальне завдання на МКР, затверджене протоколом № 3 засідання кафедри САІТ від «14» ___09____ 2022 р.

2. Джерела розробки:

- NLP for EN : BERT CLS & 10 Classifiers. URL: https://www.kaggle.com/code/vbmokin/nlp-for-en-bert-cls-10-classifiers
- NLP for UA : BERT CLS for the notebook result. URL: https://www.kaggle.com/code/vbmokin/nlp-for-ua-bert-cls-for-the-notebook-result
- NLP : Reports & News Classification. URL: https://www.kaggle.com/datasets/vbmokin/nlp-reports-news-classification
- План управління річковим басейном Південного Бугу: аналіз і заходи річкового басейну (англ.) / Мокін В. та ін. ; за ред. С. Афанасьєв, А. Петерс, В. Сташук, О. Ярошевич., Київ, 2014. 188 с.

3. Мета і призначення роботи:

Підвищення точності класифікації текстової англомовної інформації про масиви вод басейну річки Південний Буг шляхом створення відповідної інтелектуальної інформаційної технології.

4.Вихідні дані для проведення робіт:

Датасет Kaggle «NLP : Reports & News Classification» з даними про водні ресурси.

5.Методи дослідження:

Методи машинного навчання, регресійні моделі та моделі, які побудовані на основі дерев рішень.

6. Етапи роботи і терміни їх виконання:

1. Характеристика проблеми класифікації англомовної інформації про стан водних ресурсів..... 19.09 – 30.09
2. Вибір оптимальних налаштувань інформаційної технології для розв’язання поставленої задачі..... 01.10 – 16.10
3. Створення інтелектуальної інформаційна технології класифікації англомовної інформації про масиви вод басейну річки Південний Буг..... 17.10 – 13.11
4. Економічна частина ..... 14.11 – 23.11
5. Оформлення пояснювальної записки. ..... 24.11 – 29.11

7. Очікувані результати та порядок реалізації:

Точна класифікація текстових англомовних даних про водні ресурси, які синтезуються автоматично з використанням методів штучного інтелекту.

8. Вимоги до розробленої документації

Пояснювальна записка оформлена у відповідності до вимог «Методичних вказівок до виконання та оформлення магістерських кваліфікаційних робіт для

студентів спеціальності 126 – «Інформаційні системи та технології» денної форми навчання».

9. Порядок приймання роботи

Публічний захист ..... «_20_» _____12_____ 2022 р.

Початок розробки..... «_19_» _____09_____ 2022 р.

Граничні терміни виконання МКР ..... «_30_» _____11_____ 2022 р.

Розробив студент групи 2ІСТ-21м _____ Радецький О.В.

Додаток Б

Протокол перевірки кваліфікаційної роботи на наявність текстових запозичень

Назва роботи: «Інформаційна NLP-технологія класифікації англомовної інформації про стан водних ресурсів»
Тип роботи: магістерська кваліфікаційна робота
Підрозділ: кафедра САІТ
Науковий керівник: Мокін В. Б. д.т.н., проф. каф. САІТ

Показники звіту подібності Unicheck

Оригінальність	99 %
Схожість	1 %

Аналіз звіту подібності (відмітити потрібне):

✓ запозичення, виявлені у роботі, оформлені коректно і не містять ознак плагіату;
- виявлені у роботі запозичення не мають ознак плагіату, але їх надмірна кількість викликає сумніви щодо цінності роботи і самостійності її автора. Роботу направити на розгляд експертної комісії кафедри;
- виявлені у роботі запозичення є недобросовісними і мають ознаки плагіату та/або в ній містяться навмисні спотворення тексту, що вказують на спроби приховування недобросовісних запозичень.

Опис прийнятого рішення:

Робота допускається до захисту

Особа, відповідальна за перевірку Жуков С. О.

Ознайомлені з повним звітом подібності, який був згенерований системою Unicheck щодо роботи.

Автор роботи Радецький О. В.

Керівник роботи Мокін В. Б

Додаток В

Лістинг програми

```
import numpy as np
import pandas as pd
import pickle
import matplotlib
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# models
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics

# NLP
import torch
import transformers as ppb

import warnings
warnings.filterwarnings('ignore')

# Model choice
model_name = "random-forest"
# model_name = "linear-regression"

# Get notebook_models name
if model_name=="random-forest":
    notebook_models = "nlp-for-en-bert-cls-with-rf"
elif model_name=="linear-regression":
    notebook_models = "nlp-for-en-bert-cls-with-linreg"

# Set parameters
random_state = 0
```

```
notebook_data = "euroreportonwaterresources"
data_name = "water-data-Southern-Booh-for-random-forest.csv"

df = pd.read_csv(f'/kaggle/input/{notebook_data}/{data_name}')
df = df.fillna(0)
convert_dict = {'text': str,
                'env_problems': int,
                'pollution': int,
                'treatment': int,
                'climate': int,
                'biomonitoring': int}
df = df.astype(convert_dict)
df
df.info()
# BERT: Data prepairing and modeling
def BERT_modeling(df, text='text'):
    # Using DistilBERT for the dataframe df[text]
    model_class, tokenizer_class, pretrained_weights =
(ppb.DistilBertModel, ppb.DistilBertTokenizer, ' distilbert-
base-uncased')
    # Other models:
https://huggingface.co/transformers/pretrained_models.html
    # Load pretrained model/tokenizer
    tokenizer =
tokenizer_class.from_pretrained(pretrained_weights)
    model = model_class.from_pretrained(pretrained_weights)
    # Tokenization the sentences - break them up into word and
subwords in the format BERT is comfortable with
    tokenized = df[text].apply((lambda x: tokenizer.encode(x,
add_special_tokens=True)))
    max_len = 0
    for i in tokenized.values:
        if len(i) > max_len:
            max_len = len(i)
    padded = np.array([i + [0]*(max_len-len(i)) for i in
```

```
tokenized.values])
    # Creation variable to ignore (mask) the data padding
    attention_mask = np.where(padded != 0, 1, 0)
    # Modeling
    input_ids = torch.tensor(padded).to(torch.int64)
    attention_mask =
torch.tensor(attention_mask).to(torch.int64)
    with torch.no_grad():
        last_hidden_states = model(input_ids,
attention_mask=attention_mask)
    # Last hidden states
    features = last_hidden_states[0][:,0,:].numpy()
    return features
features = BERT_modeling(df, 'text')
# Text classification
def model_prediction(df, test_features, target):
    # Models training and data prediction for all models from
DataFrame models
    # Target
    labels = df[target]
    # Model loading
    with open(f'/kaggle/input/{notebook_models}/model-
{model_name}-{target}.pkl','rb') as f:
        model = pickle.load(f)
    # Prediction
    test_pred = model.predict(test_features)
    test_pred = [int(x>0) for x in test_pred]
   return test_pred
# List of the target features in df
cols = df.columns.tolist()[1:]
print('Target columns:', cols)
# Solving NLP Classification tasks
print('Solving NLP Classification tasks')
for col in cols:
    df[col] = model_prediction(df, features, col)
```

```
# Results display
pd.set_option('max_colwidth',1000)
df.to_csv(f'{notebook_data}-{notebook_models}-classified.csv',
index=False)
print(f"The result of classification using the model
{model_name}")
display(df)
```

Додаток Г

ІЛЮСТРАТИВНА ЧАСТИНА

ІНФОРМАЦІЙНА NLP-ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ АНГЛОМОВНОЇ ІНФОРМАЦІЇ ПРО СТАН ВОДНИХ РЕСУРСІВ

Виконав: студент гр. 2ІСТ-21м
_____ Радецький О.В.
«_01_» _____12_____ 2022 р.

Керівник: д.т.н., проф. каф. САІТ
_____ Мокін В. Б.
«_02_» _____12_____ 2022 р.

Нормоконтроль: к.т.н., доцент
_____ Жуков С. О.
«_02_» _____12_____ 2022 р.

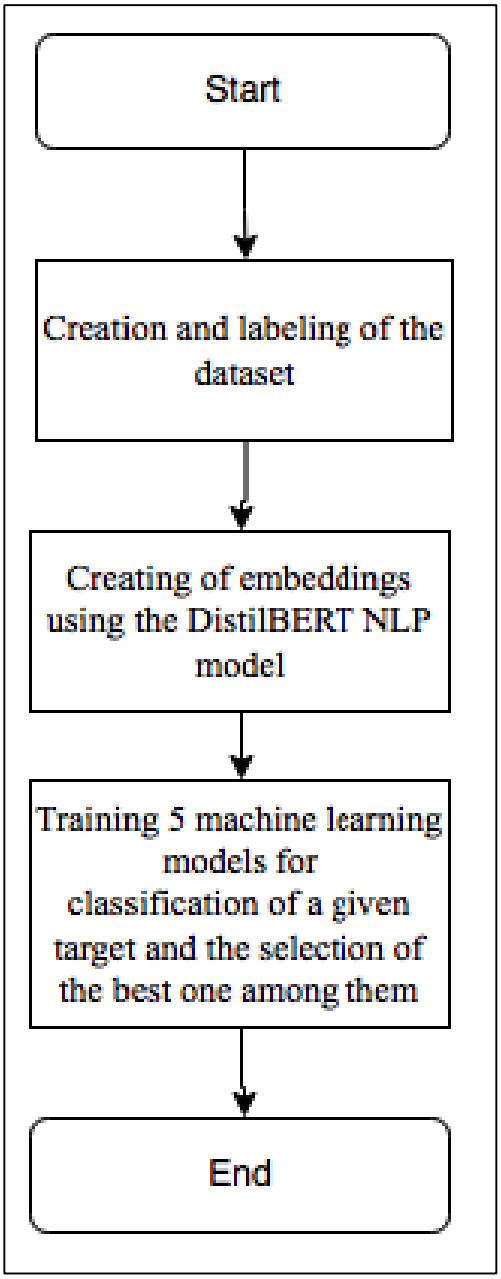

Рисунок Г.1 – Алгоритм розробки моделі класифікаторів

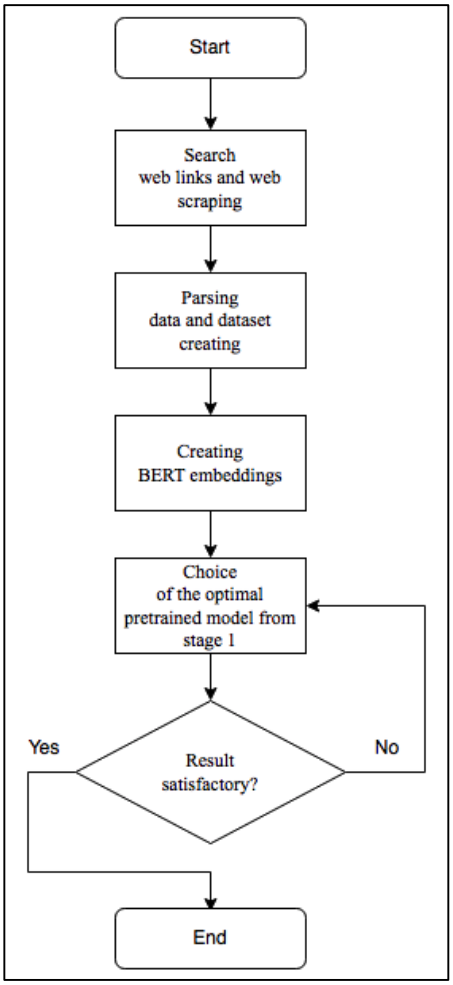

Рисунок Г.2 – Алгоритм інформаційної технології

	A	B	C	D	E	F
1	text	env_problem	pollution	treatment	climate	biomonitoring
2	They are key for supporting society and the ecconomy throu	0	0	0	0	0
3	However surface waters have traditionally beeɔ used as di	0	0	0	0	0
4	They have also been altered (by building dams and canals	0	0	0	0	0
5	to facilitate agriculture and urbanisation produce energy a	0	0	0	0	0
6	The Water Framework Directive (WFD)¬†stipu ates that El	0	0	0	0	0
7	Ecological status and potential are criteria used to assess t	0	0	0	0	0
8	Ecological status is influenced by water quality (e.g.	0	0	0	0	0
9	pollution) and habitat degradation and is used as a proxy f	0	0	0	0	0
10	According to countriesвАЩ second river basin managemer	0	0	0	0	0
11	Moreover there has been little improvement in ecological	0	0	0	0	0
12	Despite this lack of overall improvement the status of mar	0	0	0	0	0
13	The main pressures on surface water bodies are pollution	0	0	0	0	0
14	agriculture) sources and various hydromorphological press	0	0	0	0	0
15	To improve and restore the approximately 60% of surface	0	0	0	0	0
16	The percentage of water bodies with less than good (i.e.	0	0	0	0	0
17	moderate poor or bad) ecological status varies between riv	0	0	0	0	0
18	Surface water bodies in north-western Europe have the lov	0	0	0	0	0
19	In Belgium (Flanders) northern Germany and the Netherla	0	0	0	0	0
20	Other problem areas include Czechia southern England (Ur	0	0	0	0	0
21	Northern countries particularly the northern Scandinavian	0	0	0	0	0
22	The investigation of the Tisza River was a follow-up projec	0	0	0	0	0
23	The objective of the survey was to investigate¬†the water	0	0	0	0	0
24	The countries participating at the survey include Romania	0	0	0	0	0

Рисунок Г.3 – Сформований датасет для тестування моделей

	name	acc_train	acc_test
4	MLPClassifier	0.983333	0.975
1	Random Forest Classifier	0.966667	0.950
0	Linear Regression	1.000000	0.900
5	Mean values	0.983333	0.900
2	Bagging Classifier	0.983333	0.875
3	XGB Classifier	0.983333	0.800
6	Max values	0.950000	0.775

Рисунок Г.4 – Результат роботи відібраних моделей

	text	env_problems	pollution	treatment	climate	biomonitoring
0	They are key for supporting society and the economy throughout Europe and clean unpolluted waters are essential for healthy ecosystems.	1	0	1	1	1
1	However surface waters have traditionally been used as disposal routes for human agricultural and industrial wastes damaging water quality.	1	1	1	1	0
2	They have also been altered (by building dams and canals etc.)	1	1	1	0	0
3	to facilitate agriculture and urbanisation produce energy and protect against flooding all of which can change and degrade habitats.	1	1	0	1	1
4	The Water Framework Directive (WFD) stipulates that EU Member States should aim to achieve good status for all surface water and groundwater bodies.	0	0	0	1	1
5	Ecological status and potential are criteria used to assess the quality of the structure and functioning of surface water ecosystems.	0	0	1	0	1
6	Ecological status is influenced by water quality (e.g.	1	1	1	0	1
7	pollution) and habitat degradation and is used as a proxy for the overall status of water bodies.	1	1	1	1	0

Рисунок Г.5 – Класифікація за допомогою моделі лінійної регресії

The result of classification using the model random-forest

	text	env_problems	pollution	treatment	climate	biomonitoring
0	They are key for supporting society and the economy throughout Europe and clean unpolluted waters are essential for healthy ecosystems.	1	0	0	0	0
1	However surface waters have traditionally been used as disposal routes for human agricultural and industrial wastes damaging water quality.	1	0	0	0	0
2	They have also been altered (by building dams and canals etc.)	1	0	0	0	0
3	to facilitate agriculture and urbanisation produce energy and protect against flooding all of which can change and degrade habitats.	1	0	0	0	0
4	The Water Framework Directive (WFD) stipulates that EU Member States should aim to achieve good status for all surface water and groundwater bodies.	1	0	0	0	0
5	Ecological status and potential are criteria used to assess the quality of the structure and functioning of surface water ecosystems.	1	0	0	0	0
6	Ecological status is influenced by water quality (e.g.	0	0	0	0	0
7	pollution) and habitat degradation and is used as a proxy for the overall status of water bodies.	1	0	0	0	0
8	According to countries' second river basin management plans (RBMPs) covering the period up to 2015 good or better (high) ecological status has been achieved for only around 40% of surface waters (rivers lakes and transitional and coastal waters).	1	0	0	0	0
9	Moreover there has been little improvement in ecological status since the publication of the first RBMPs in 2009 with the status of most water bodies remaining similar.	1	0	0	0	0
10	Despite this lack of overall improvement the status of many of the individual elements that make up ecological status has improved.	1	0	0	0	0

Рисунок Г.6 – Класифікація за допомогою моделі випадкових лісів (Random Forest)