МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

УМАНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ САДІВНИЦТВА

Факультет економіки і підприємництва

Кафедра інформаційних технологій

Допущено до захисту

"_____" _____ 2023 р.

Зав. кафедри, к.т.н., доцент

_____ **Ліщук Р.І.**

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

РОЗРОБКА ІНФОРМАЦІЙНОЇ СИСТЕМИ ДЛЯ АНАЛІЗУ БІЗНЕС-ЗВІТІВ З ВИКОРИСТАННЯМ ІНСТРУМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ

Виконав – студент 2 курсу, групи 21м-кн
освітньої програми "Комп’ютерні науки"
спеціальності 122 Комп’ютерні науки
Вінявський Олександр Миколайович

Науковий керівник – к.т.н., доцент
Ліщук Роман Ігорович

Умань – 2023

Зміст

ВСТУП 3

РОЗДІЛ 1 АНАЛІЗ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧІ 5

1.1. Огляд сучасних методів оцінки підприємства 5

1.2. Аналіз існуючих аналогів 6

1.3. Постановка задачі 10

РОЗДІЛ 2 ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ ПІДСИСТЕМИ FaceAPI12

2.1. Методи дослідження 12

2.2. Вибір інструментів реалізації 13

2.3. Моделювання проєкту 17

2.4. Побудова моделі варіантів використання 19

2.5. Моделювання архітектури сервісу 22

РОЗДІЛ 3 РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ ДЛЯ АНАЛІЗУ БІЗНЕС-ЗВІТІВ З ВИКОРИСТАННЯМ ІНСТРУМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ 25

3.1. Реалізація моделей обробки даних методу 25

3.2. Реалізація веб-інтерфейсу методу 34

3.3. Опис алгоритму використання методу 35

3.4. Апробація результатів 37

ВИСНОВКИ 43

СПИСОК ЛІТЕРАТУРИ 45

ДОДАТКИ 49

ВСТУП

З розвитком інформаційних технологій та збільшенням обсягу доступних даних, підприємства постійно потребують нових можливостей щодо оптимізації діяльності та підвищення конкурентоспроможності. Використовуючи інтелектуальну обробки даних, аналіз даних ринку, конкурентів, споживачів та внутрішні ресурси підприємства підприємство має змогу оцінити власні ресурси, ринок та побудувати ефективні стратегії розвитку.

Актуальність. Використання штучного інтелекту для аналізу даних полягає в тому, що традиційні методи аналізу можуть виявитися неефективними при великих обсягах даних та змінних ринкових умовах. Штучний інтелект дозволяє здійснювати автоматичну обробку та аналіз даних, швидко знаходити закономірності та адаптуватися до нових умов, що сприяє підвищенню конкурентоспроможності підприємства. Прикладами використання штучного інтелекту є сучасні системи прогнозування попиту, сегментація клієнтів, оцінка конкурентів, управління ризиками, оптимізація виробничих процесів [1].

Мета. Розробка методу інтелектуальної обробки даних для оцінки конкурентоспроможності підприємства. Метод буде обробляти аудиторські тексти та визначати ESG критерії [2] формуючі звіт з реченнями, що стосуються ESG критеріїв та їх аналізу.

Мета складається з наступних задач:

- Аналізу предметної області, огляд аналогів;
- Постановка методу та встановлення методу апробації дослідження;
- Вибір інструментів реалізації методу;
- Структурно-функціональне моделювання методу, побудова діаграм використання методу та моделювання архітектури методу;
- Реалізація методу, навчання моделей штучного інтелекту, розробка вебінтерфейсу, опис алгоритму використання методу;
- Проведення апробації результатів методу.

Об'єкт дослідження. Інтелектуальний аналіз даних для оцінки конкурентоспроможності підприємства.

Предмет дослідження. Сукупність теоретичних та практичних аспектів методу обробки даних з використанням штучного інтелекту.

Наукова новизна. Новизна даного методу полягає у створенні методу який буде аналізувати текст саме на ESG критерії, відокремлюючи сутності, що відносяться до цих критеріїв та оцінюючі кожен з критеріїв.

Гіпотеза дослідження. Використання штучного інтелекту має пришвидшити та покращити точність аналізу тексту, що має не тільки зменшити час на обробку тексту, а і збільшити об’єми даних для аналізу.

Основні задачі. Дослідити предметну область, а саме оглянути сучасні методи оцінки даних підприємства, проаналізувати існуючі аналоги виявивши їх переваги та недоліки.

Встановити задачі дослідження, методи, що будуть використовуватися та обрати набір інструментів реалізації цих методів та задач.

Змоделювати за встановленими задачами та методами реалізації моделі IDEF0 [3] та Use case [4]

Практичне значення. Використання штучного інтелекту для аналізу тексту щодо ESG критеріїв пришвидшить процес оцінки та виявлення цих критеріїв що зменшить навантаження для аналітика.

РОЗДІЛ 1 АНАЛІЗ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧІ

1.1. Огляд сучасних методів оцінки підприємства

Згідно з досліджень використання штучного інтелекту у бізнесі підвищує його конкурентоспроможність, розглянемо одне з таких досліджень [5].

У досліджені розглянуто що за останні кілька років можна спостерігати появу великої кількості інтелектуальних продуктів та послуг, їх комерційну доступність та соціально-економічний вплив. Це породжує питання, чи є сучасне поширення ШІ просто тимчасовим явищем або методика використання ШІ дійсно має потенціал перетворити світ бізнесу. Стаття розглядає впливові академічні досягнення та інновації у галузі ШІ, їх вплив на підприємницьку діяльність та глобальний ринок.

Дослідження розглядає чотири основні області використання штучного інтелекту: комп'ютерний зір, текстовий аналіз, розпізнавання мови та гра. Аналіз 200 кращих сатрапів у галузі ШІ, показує вплив передових досліджень та інновацій у галузі ШІ на світовий ринок, хвиля ШІ триває і апетит до росту ШІ є експоненційним.

Сучасні методи оцінки також використовують штучний інтелект для аналізу відіграють величезну роль у рішенні різноманітних завдань. Від ефективного управління бізнесом до наукових досліджень, аналіз даних стає невід'ємною складовою успіху. Завдяки розвитку технологій та наукового прогресу, сучасні методи оцінки даних все більше використовують штучний інтелект для отримання точних показників витрат, продуктивності, потреб ринку та інших критичних параметрів.

Штучний інтелект забезпечує можливість аналізувати великі обсяги інформації швидко та ефективно, виявляючи закономірності та взаємозв'язки між даними, які можуть бути непомітними для людського ока. Також штучний інтелект може прогнозувати майбутнє ринка та компанії тому використання штучного інтелекту для аналізу підприємства відіграє велику роль у сучасному світі.

Використовують безліч технологій для оцінки підприємства проте їх функції можна представити у вигляді списку:

- Збір даних щодо діяльності підприємства
- Аналіз існуючих даних
- Прогнозування збільшення заданих критеріїв

1.2. Аналіз існуючих аналогів

Розглянемо основні існуючі аналоги, які використовуються для проведення аналізу діяльності компаній та визначення їх конкурентоспроможності, визначивши їх недоліки та переваги.

Sisens



Рисунок 1.1 - Вигляд проєкту у Sisense

До основних переваг платформи можна віднести:

- Легкість використання - дозволяє різним користувачам з будь-яким рівнем досвіду працювати з інструментом без значних навчальних зусиль та досвіду.

- Велика швидкість обробки даних - Sisense використовує технологію InChip [7], яка забезпечує високу швидкість обробки даних.

- Звіти та дашборди в реальному часі - Sisense дозволяє створювати інтерактивні звіти та дашборди, які відображають дані в реальному часі, що сприяє оперативному прийняттю рішень.

- Масштабованість: Sisense може легко масштабуватися, що дозволяє відповідати потребам різних компаній.

Основними недоліками є:

- Висока вартість - Sisense може бути досить коштовним рішенням, особливо для малого та середнього бізнесу.

- Базовий функціонал - деякі функціональні можливості, такі як розширені аналітичні інструменти або додаткові інтеграції, можуть вимагати додаткових модулів або розширень.

IBM Watson Analytics

IBM Watson Analytics [8] - це сервіс для аналізу даних, який використовує штучний інтелект та машинне навчання для виявлення закономірностей та тенденцій в даних. IBM розробила Watson Analytics (2014 рік). Ця платформа поєднує передбачувальний аналіз, обробку природної мови та можливості машинного навчання, щоб автоматизувати аналітичний процес. Сервіс допомагає компаніям проводити аналіз конкурентоспроможності, визначати слабкі місця та можливості для росту, отримувати індивідуальні рекомендації з удосконалення стратегій та бізнес-процесів (рис. 1.2).

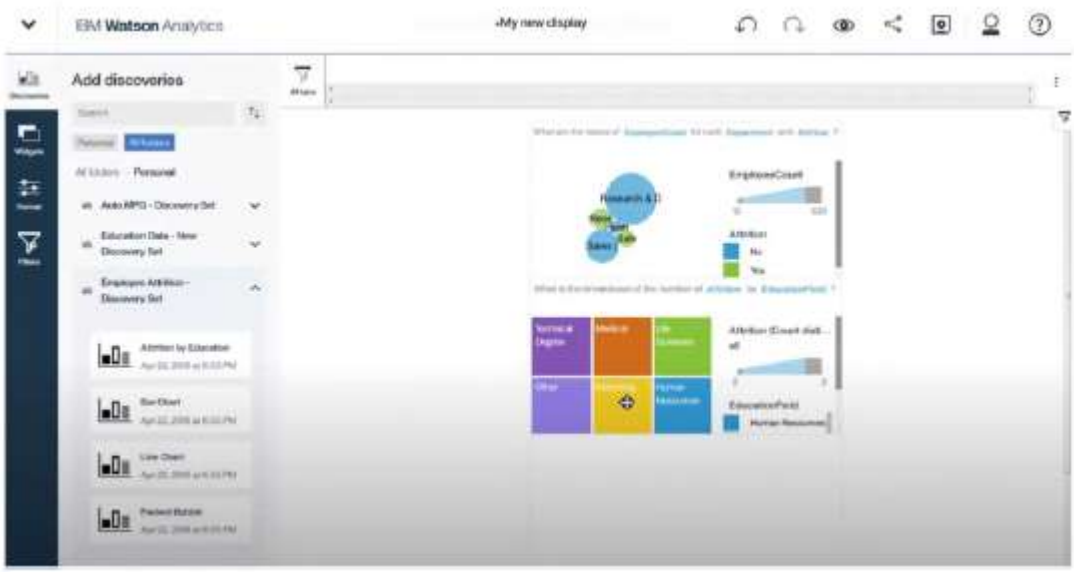


Рисунок 1.2 - Вигляд проєкту у IBM Watson Analytics

Основними перевагами даного сервісу є:

- Підтримка прийняття рішень - надає можливість задавати запити в натуральній мові, а також отримувати відповіді та рекомендації, засновані на даних, що сприяє ефективному прийняттю рішень.

- Інтеграція з різними джерелами даних - Watson Analytics підтримує інтеграцію з великою кількістю джерел даних, включаючи бази даних, хмарні сховища, API та інші.

- Безпека даних - забезпечує високий рівень безпеки даних, використовуючи шифрування, аутентифікацію та авторизацію для захисту інформації.

Недоліки сервісу:

- Недостатня гнучкість - відміну від інших інструментів аналітики, Watson Analytics може бути менш гнучким у налаштуванні та адаптації до конкретних потреб користувачів або специфічних вимог певної галузі.

- Потреба в додаткових модулях та розширеннях - деякі функціональні можливості, такі як розширені аналітичні інструменти або додаткові інтеграції.

DataRobot

DataRobot [9] - це автоматизована платформа машинного навчання, яка дозволяє компаніям легко будувати, тренувати та впроваджувати моделі машинного навчання для аналізу конкурентоспроможності. Завдяки просунутим алгоритмам штучного інтелекту та машинного навчання DataRobot може надавати глибокий аналіз даних, виявляти фактори, які впливають на конкурентоспроможність, та рекомендувати оптимальні дії для досягнення більш високої ефективності та прибутковості (рис. 1.3).

Перевагами цього сервісу є:

- Автоматизація машинного навчання - автоматизує процес підготовки даних, створення моделей та їх оптимізації.

- Широкий набір алгоритмів - підтримує велику кількість алгоритмів машинного навчання, що дозволяє користувачам вибирати найбільш відповідні моделі для своїх завдань.

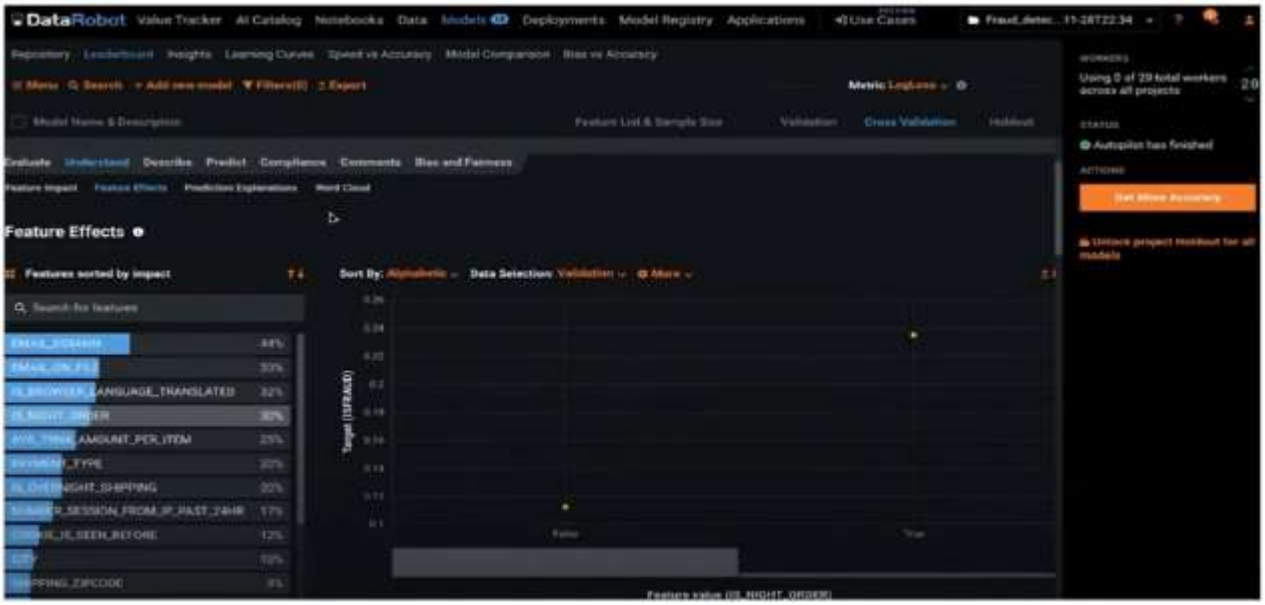


Рисунок 1.3 - Вигляд проєкту у Datarobot

Масштабованість та гнучкість дозволяє користувачам легко адаптувати платформу до своїх потреб та забезпечити високу продуктивність навіть при роботі з великими обсягами даних.

Серед основних недоліків можна виділити:

Залежність від провайдера - це хмарна платформа і користувачі повинні довіряти провайдеру для зберігання та обробки своїх даних.

Складний інтерфейс - користувачам може знадобитися додаткове навчання для того, щоб зрозуміти принципи роботи алгоритмів та як коректно інтерпретувати результати.

Для більшої наочності побудуємо таблицю (див. таблиця 1.1). порівняння аналогів:

Таблиця 1.1 - Огляд аналогів.

Критерій / Продукт	Легкість використання	Ціна	Безпека даних	Документація
Sisense	+	-	-	-/+
IBM Watson Analytics	+	+	+/-	+
Data robot	-	+	+	+

З огляду на проведений аналіз аналогів, ключовими факторами для майбутнього методу буде простота використання, модульність та документація до сервісу, на базі якого буде реалізований метод. Порівнявши аналоги, знаходимо найближчого конкурента - «IBM Watson Analytics», який навіть має допоміжний модуль AI, який за запитом користувача сформує таблицю даних, проте точність формування саме ESG критеріїв метода це не гарантує, велика вартість та громіздкість порівняно з єдиним сервісом, який існуватиме саме для аналізу ESG критеріїв.

1.3. Постановка задачі

Метою дослідження є створення методу інтелектуальної обробки даних для оцінки конкурентоспроможності підприємства.

Для досягнення цієї мети, будуть вирішені наступні завдання:

- Проведення аналізу предметної області, дослідження стану проблематики та причин створення методу, аналіз існуючих методів для визначення необхідних критеріїв методу.

- Постановка задачі, а саме: встановлення мети та задачі метода, вибір інструментів реалізації методу, визначити основні критерії роботи методу.

- Обрати інструменти реалізації поставлених завдань.

- Моделювання проєкту, побудова діаграми Use case [3] та IDEF0 [4].

- Реалізувати метод у вигляді сервісу та провести валідацію результатів.

- Апробація результатів методу.

Сформовано кроки щодо реалізації методу:

- Розробка алгоритму реалізації аналізу бізнес-звітів з використанням штучного інтелекту та Microsoft Azure [10].

- Визначення критеріїв «Екологія», «Соціум» та «Корпоративне управління» та їх опис.

- Розробка системи введення даних для аналізу, включаючи можливість завантаження документів у різних форматах.

- Розробка веб-інтерфейсу для відображення результатів аналізу звітів за встановленими критеріями.

- Підготовка документації для користувачів та технічної документації.

- Тестування та налаштування сервісу для забезпечення оптимальної продуктивності та точності аналізу.

- Проведення демонстрації та оцінка ефективності сервісу.

Очікуваний результат проєкту - створення простого та ефективного методу для аналізу бізнес-звітів за критеріями «Екологія», «Соціум» та «Корпоративне управління», що дозволить користувачам швидко та точно визначати рівень виконання цих критеріїв компаніями.

РОЗДІЛ 2 ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ ПІДСИСТЕМИ FaceAPI

2.1. Методи дослідження

Ідея методу полягає у використанні сучасних аналітичних технологій для обробки та аналізу великої кількості даних, пов'язаних з ESG критеріями [2] (екологічні, соціальні та корпоративне управління) підприємства. Це дозволить краще розуміти та оцінювати конкурентоспроможність компанії в умовах сучасного ринку та виявити можливі вектори розвитку для підвищення її сталості, а також сформувати короткий звіт щодо діяльності компанії з оцінкою ESG критеріїв.

Для аналізу ESG критеріїв метод передбачає використання NER (Named Entity Recognition) моделі [11], яка дозволяє відокремити сутності ESG критеріїв з текстових джерел. Це полегшує збір та обробку даних для подальшого аналізу. Далі, за допомогою сентимент-аналізу та регресивної моделі [12], проводиться оцінка ставлення до кожного з ESG критеріїв, що дозволить отримати оцінку за кожним ESG критерієм.

Щоб забезпечити гнучкість та можливість масштабування методу, буде використовуватися Azure ML Studio [10] як інструмент для розробки та впровадження моделей машинного навчання. Azure ML Studio дозволяє швидко створювати, тренувати та налаштовувати моделі, а також легко інтегрувати їх з іншими моделями, різними джерелами даних та сервісами. Завдяки можливостям Azure ML Studio, метод може бути масштабований та адаптований до різних сценаріїв та сфер діяльності підприємств.

На початку вхідні дані до методу переносяться у текстовий формат, у якому буде видалено порожні рядки, наступним етапом дані передаються до моделі для відокремлення сутностей що стосуються критеріїв, ці сутності передаються до наступної моделі яка формує оцінку критеріїв, з вихідних даних формується звіт у якому відображено речення які стосуються ESG критеріїв та оцінка за критерієм.

Загалом, метод інтелектуальної обробки даних для оцінки конкурентоспроможності підприємства забезпечує комплексний підхід до аналізу конкурентоспроможності, який враховує ESG критеріїв та використовує передові технології машинного навчання для отримання точних та актуальних результатів.

Для того щоб обрати метод дослідження результатів розглянемо методи дослідження:

- Експериментальний метод [13] - метод дозволяє отримати точні результати за допомогою проведення спеціально організованого експерименту, де відбувається маніпулювання змінними факторами і вимірювання результатів.

- Кореляційний метод [14] - метод дозволяє встановити залежність між двома або більше змінними, які можуть бути виміряні або підраховані. Він дає можливість зрозуміти, наскільки сильно залежні одна від одної змінні, тобто чи вони мають пряму або зворотню залежність.

- Анкетування та опитування [15] - метод полягає у зборі даних за допомогою заповнення анкет або проведення опитування серед відповідної аудиторії. Це дозволяє отримати більш об'єктивні результати та зрозуміти думки, погляди та переконання людей.

У даному випадку, якщо мета полягає в аналізі бізнес звітів за критеріями «Екологія», «Соціум» та «Корпоративне управління», більш доцільно використати експериментальний метод, оскільки це дозволить свій метод виявлення точності показників, враховуючі додаткові фактори.

2.2. Вибір інструментів реалізації

Наступним етапом необхідно обрати інструменти реалізації. Необхідні інструменти для будування моделі штучного інтелекту та її навчання. На сьогодні існує багато додатків для аналізу тексту, проте незначна кількість з них спрямовані на проведення аналізу бізнес-критеріїв або бізнес-звітів за

певними критеріями. Серед додатків для аналізу сутностей тексту можна виділити такі, як: «Google Cloud Natural Language API» [16], «IBM Watson Natural Language Understanding» [17], «TextRazor» [18], «Stanford Named Entity Recognizer (NER)» [19] та «Microsoft Azure» [10], які можна порівняти за їхніми проєктами.

Порівняємо 5 продуктів для аналізу текстів за такими критеріями:

- Вартість.
- Мовні можливості.
- Функціональні можливості аналізу тексту.
- Доступність та інтеграція.
- Швидкість та максимальний об’єм обробки тексту.

Вартість:

- Google Cloud Natural Language API: базовий план з безкоштовним обсягом до 5 000 одиниць обробки на місяць, після чого плата за кожну додаткову одиницю обробки; є також платні плани зі збільшеним обсягом обробки та додатковими функціями.

- IBM Watson Natural Language Understanding: безкоштовний тестовий період на 30 днів, після чого плата за кожну одиницю обробки; є також платні плани з різними обсягами обробки та функціями.

- TextRazor: базовий план з безкоштовним обсягом до 5 000 запитів на місяць, після чого плата за кожний додатковий запит; є також платні плани зі збільшеним обсягом запитів та додатковими функціями.

- Stanford Named Entity Recognizer (NER): безкоштовний та відкритий джерело.

-Microsoft Azure: безкоштовний тестовий період на 7 днів, після чого плата за кожну одиницю обробки; є також платні плани з різними обсягами обробки та функціями.

Мовні можливості:

- Google Cloud Natural Language API, IBM Watson Natural Language Understanding, TextRazor та Microsoft Azure підтримують більшість мов, включаючи англійську, іспанську, французьку, німецьку, італійську, португальську, російську та інші. Деякі з цих сервісів також підтримують менш поширені мови, такі як грузинська, угорська та інші.

- Stanford Named Entity Recognizer (NER) підтримує обробку англійської, німецької та китайської мов.

Функціональні можливості аналізу тексту:

- Google Cloud Natural Language API, Microsoft Azure та IBM Watson Natural Language Understanding мають дуже широкі функціональні можливості для аналізу тексту, включаючи визначення сутностей, відношень між ними, категоризацію тексту та аналіз настроїв.

- TextRazor та Stanford Named Entity Recognizer (NER) більше спеціалізуються на визначенні сутностей та їхнього класифікації.

Доступність та інтеграція:

- Всі перераховані продукти мають велику доступність та підтримку з боку своїх розробників. У порівнянні з іншими сервісами, Google Cloud Natural Language API, IBM Watson Natural Language Understanding та Microsoft Azure мають більш гнучкі та зручні інструменти інтеграції з різними платформами.

Швидкість та максимальний об'єм обробки тексту:

- Google Cloud Natural Language API та IBM Watson Natural Language Understanding можуть обробляти великі обсяги тексту з високою швидкістю, що робить їх підходящими для великих проєктів.

- TextRazor та Stanford Named Entity Recognizer (NER) мають менші обсяги обробки тексту, але їхні можливості досить точні.

- Microsoft Azure пропонує потужні інструменти для обробки тексту з великим обсягом, але швидкість обробки може відрізнятися залежно від обраних продуктів.

Сформуємо таблицю 2.1 порівняння за переліченими вище критеріями:

Таблиця 2.1 - Порівняння існуючих програм аналізу тексту.

Критерії	Назва продуктів				
	Google Cloud Natural Language API	IBM Watson Natural Language Understanding	TextRazor	Stanford Named Entity Recognizer (NER)	Microsoft Azure Services
Вартість	+	+			
Мовні можливості	+	+	+		+
Функціональні можливості аналізу тексту	+				+
Доступність та інтеграція	+/-	+/-	+/-	+	+
Швидкість та максимальний об'єм обробки тексту	+	+	+	-	+

Загалом, Google Cloud Natural Language API та IBM Watson Natural Language Understanding мають більш розширену функціональність та здатні визначати більше характеристик тексту, але вони також є комерційними продуктами, тоді як: TextRazor та Stanford Named Entity Recognizer (NER) є безкоштовними, але можуть бути обмеженими у своїй функціональності, найбільш переваг має сервіс Microsoft Azure, тому його було обрано для створення методу оцінки конкурентоспроможності підприємства з використанням ESG критеріїв.

2.3. Моделювання проєкту

Ідея методу полягає у створенні алгоритму для аналізу бізнес-текстів, який оцінює їх за ESG критеріями, використовуючи NER модель для відокремлення речень, що стосуються ESG, та проведення сентимент-аналізу з допомогою регресивної моделі. Метод включає обробку тексту, сентимент-аналіз та генерацію звіту з оцінкою критеріїв та прикладами речень. Сервіс спрощує аналіз великих аудиторських текстів для користувача.

Для структурно функціонального моделювання зробимо опис параметрів використовуючи контекстної діаграму IDEF0 [4]. Отже, ця контекстна діаграма містить узагальнений блок «Процес функціонування методу оцінки конкурентоспроможності підприємства з використанням штучного інтелекту». Відповідно до методології функціонального моделювання IDEF0, для виконання роботи необхідно встановити вхідні дані, результативні дані, дані керування та дані механізмів, які представлені на діаграмі у вигляді стрілок:

Вхідні дані:

- Текст бізнес-звітів, що містять інформацію про ESG критерії.

Управління:

- Опис речення ESG критеріїв

- Microsoft Azure для розгортання та реалізації алгоритмів, зберігання даних та обробки результатів

- Метод генерації звіту

- Опис як оцінювати речення з ESG критерієм

Механізми:

- NER модель для відокремлення речень що стосуються ESG критеріїв

- Інструмент для сентиментального аналізу речення

- Веб-сервіс для надання доступу користувачам до звіту

Вихідні дані:

- Короткий звіт щодо ESG критеріїв у тексті

- Оцінка ESG критеріїв

Графічне зображення діаграми IDEF0 зображено на рисунку 2.1.



Рисунок 2.1 - Контекстна діаграма в нотації IDEF 0

Для зображення процесів було розроблено діаграму декомпозиції IDEF0 [20]. Особливість IDEF0 полягає у її акценті на ієрархічному відображенні об'єктів, що суттєво спрощує розуміння предметної області. У IDEF0 розглядаються логічні зв'язки між роботами, а також відображаються всі сигнали керування.

При декомпозиції виявлено наступні 3 кроки:

- Відокремлення речень що містять ESG критерії
- Сентиментальний аналіз речення
- Поєднання результатів двох моделей у єдиний звіт

Нижче зображена діаграма яка описує вхідні параметри для цих кроків (рис. 2.2)



Рисунок 3.2 - Діаграма декомпозиції першого рівня IDEF0

На початку з тексту який необхідно проаналізувати будуть відокремлені речення, які місять словосполучення, слова, що відносяться до одного з ESG критеріїв, наприклад з речення «A new energy management system that enabled us to reduce our energy consumption by 15% compared to the previous year» (переклад: новий енергетичний менеджмент зменшив споживання електроенергії на 15%) буде відокремлено «reduce our energy consumption by 15%» (переклад: споживання електроенергії зменшено на 15 %) як екологічний показник компанії.

Наступним кроком це речення передається на аналіз сентиментальної моделі яка визначає вплив цього показника на загальну оцінку екологічного критерію, також буде враховуватися числові показники наприклад «reduce our energy consumption by 15%» буде впливати менше на оцінку екології ніж «reduce our energy consumption by 30%».

Останнім етапом результати двох моделей формують звіти, звіт з оцінкою критеріїв з тексту та звіт з зображенням які речення з ESG критеріями було знайдено.

2.4. Побудова моделі варіантів використання

Дослідивши варіанти використання методу аналізу ESG критеріїв, було визначено такі основні операції застосування:

- Додавання даних на аналіз
- Перегляд проаналізованих даних

Саме за такими операціями було складено діаграму User Case [3]. Діаграма Use Case використовується для зображення сценаріїв використання додатку користувачами системи або акторами, які використовують функції додатку. Актором в системі є Користувач, який передає вхідні дані на аналіз та отримує підсумок, другим актором є сервіси Azure які аналізуються дані та формують звіти, діаграму User Case наведено на рисунку 2.3.

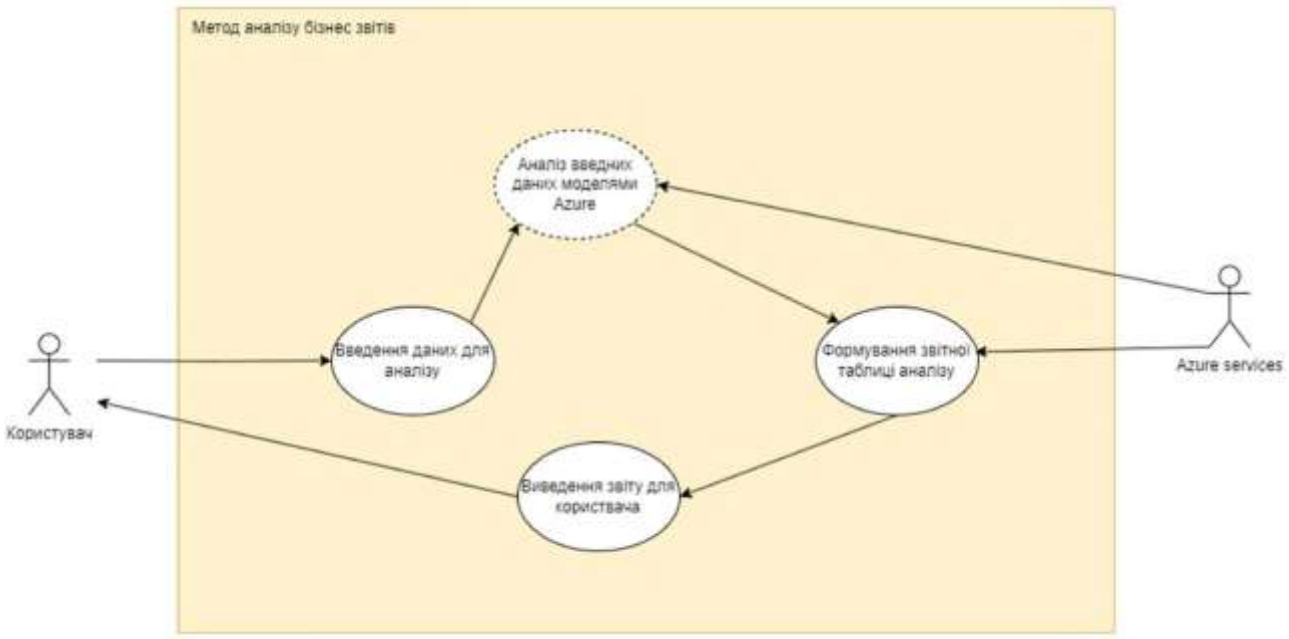


Рисунок 2.3 - Діаграма Use Case

Для кожного варіанту використання побудуємо діаграми послідовностей у яких буде описано дії, актори та обмін даними.

Введення даних для аналізу - дає змогу користувачу завантажити файл текстового формату для аналізу (рис 2.4).

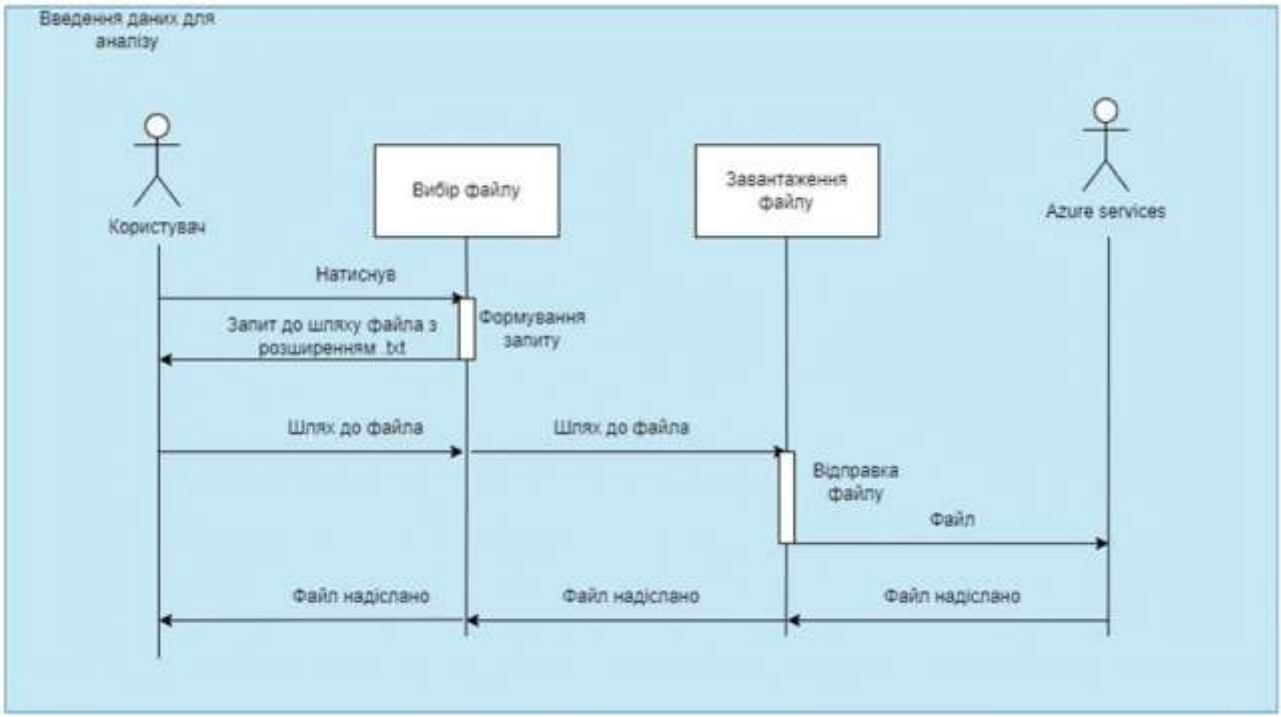


Рисунок 2.4 - Діаграма введення даних для аналізу

Аналіз введених даних моделями Azure - Обробляє вхідний текст відокремлюючи речення що стосуються ESG критеріїв та надає оцінку критеріям (рис. 2.5).

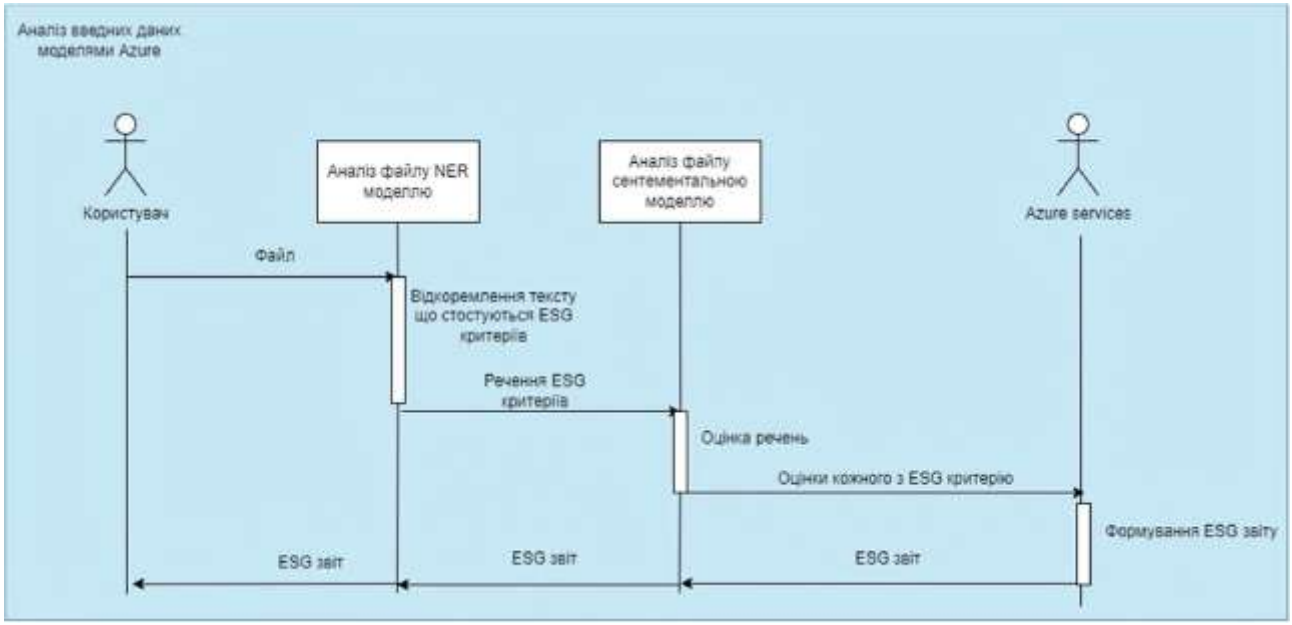


Рисунок 2.5 - Діаграма введення даних для аналізу

Формування звітної таблиці аналізу - З даних наданими моделей формує таблицю звіту у якій зображено критерії, їх оцінка та речення які стосуються цих критеріїв (рис. 2.6).

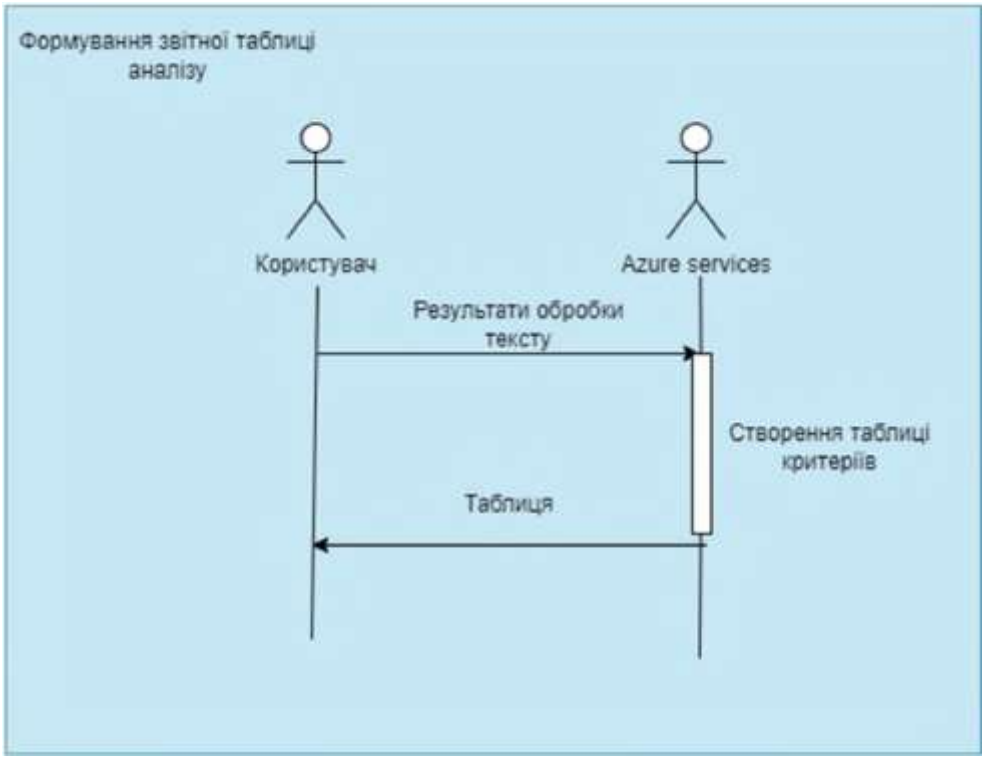


Рисунок 2.6 - Діаграма введення даних для аналізу

Виведення звіту користувачу - Сторінка яка відображує сформовану таблицю (рис. 2.7).

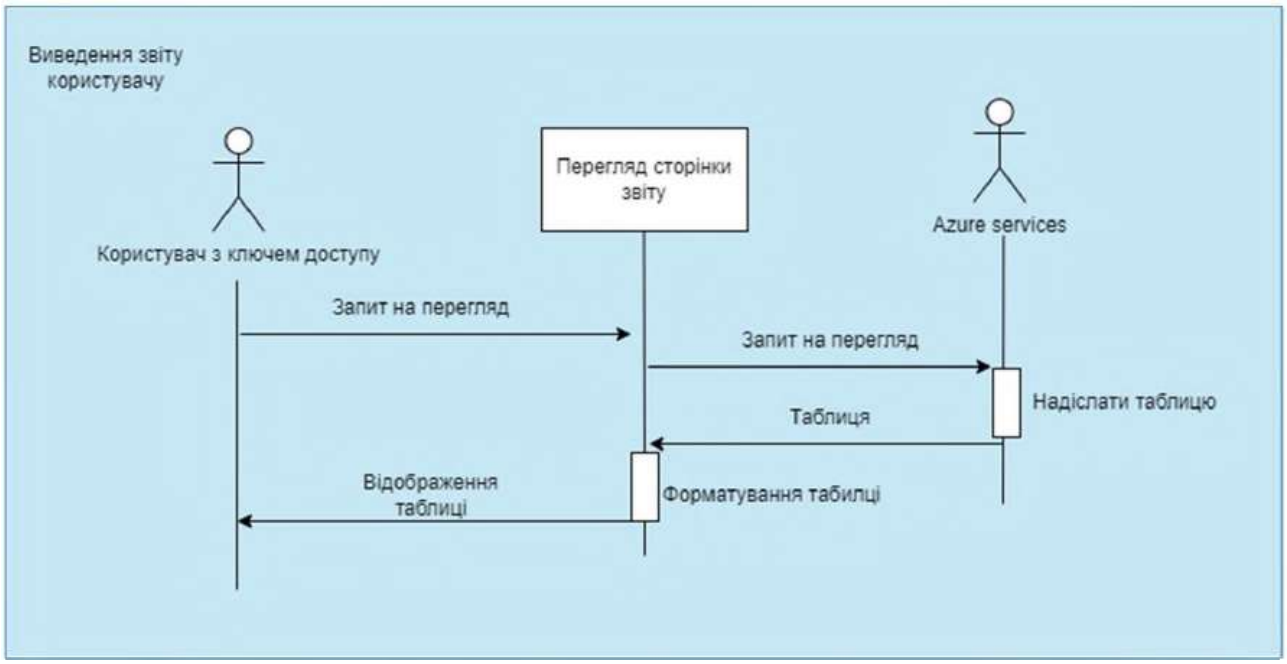


Рисунок 2.7 - Діаграма введення даних для аналізу

2.5. Моделювання архітектури сервісу

За допомогою моделі архітектури сервісу буде зображено компоненти сервісу та їх взаємозв’язок. Сам сервіс який реалізовує метод, використовує «клієнт-сервіс» [21] архітектура яка складається з серверу що обробляє вхідні дані та повертає клієнту та клієнта що надсилає та відображує дані, діаграму архітектурної системи «клієнт-сервер» наведено на рисунку 2.8

Основними компонентами системи є:

- Сторінка для завантаження даних.
- Модель NER.
- Модель сентиментального аналізу.
- Модуль для поєднання результатів моделей.
- Сторінка з результатами аналізу.

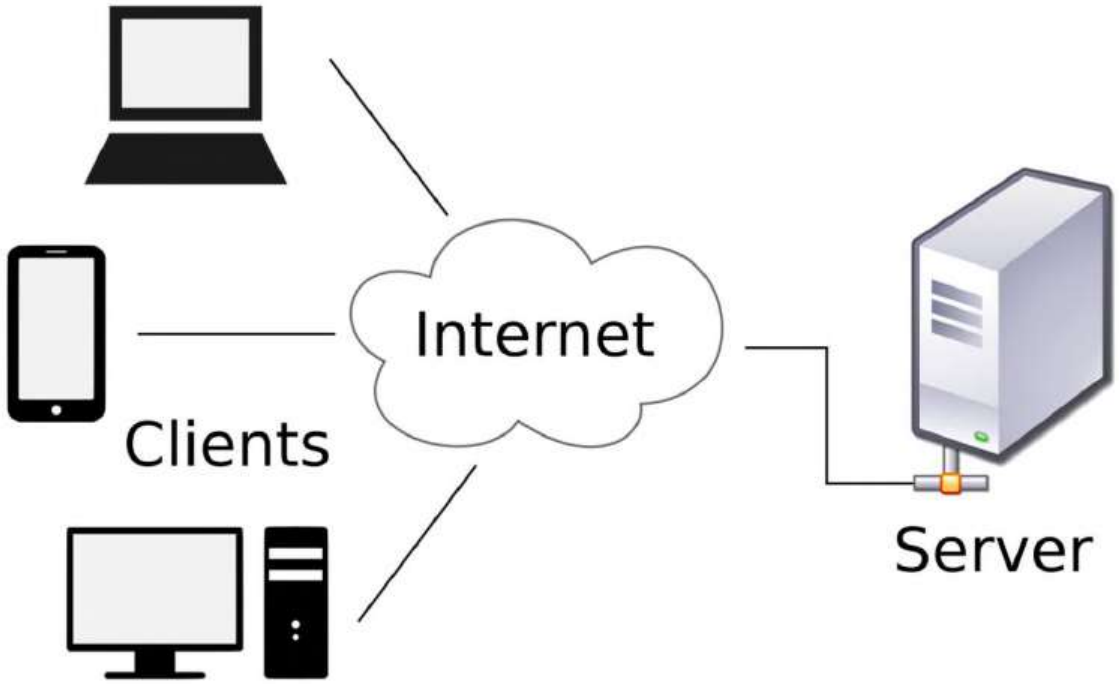


Рисунок 2.8 - Діаграма клієнт-сервер

Побудуємо діаграму компонентів взаємозв’язків компонентів та до якої з частин «клієнт-сервер» архітектури вони належать (рис. 2.9).

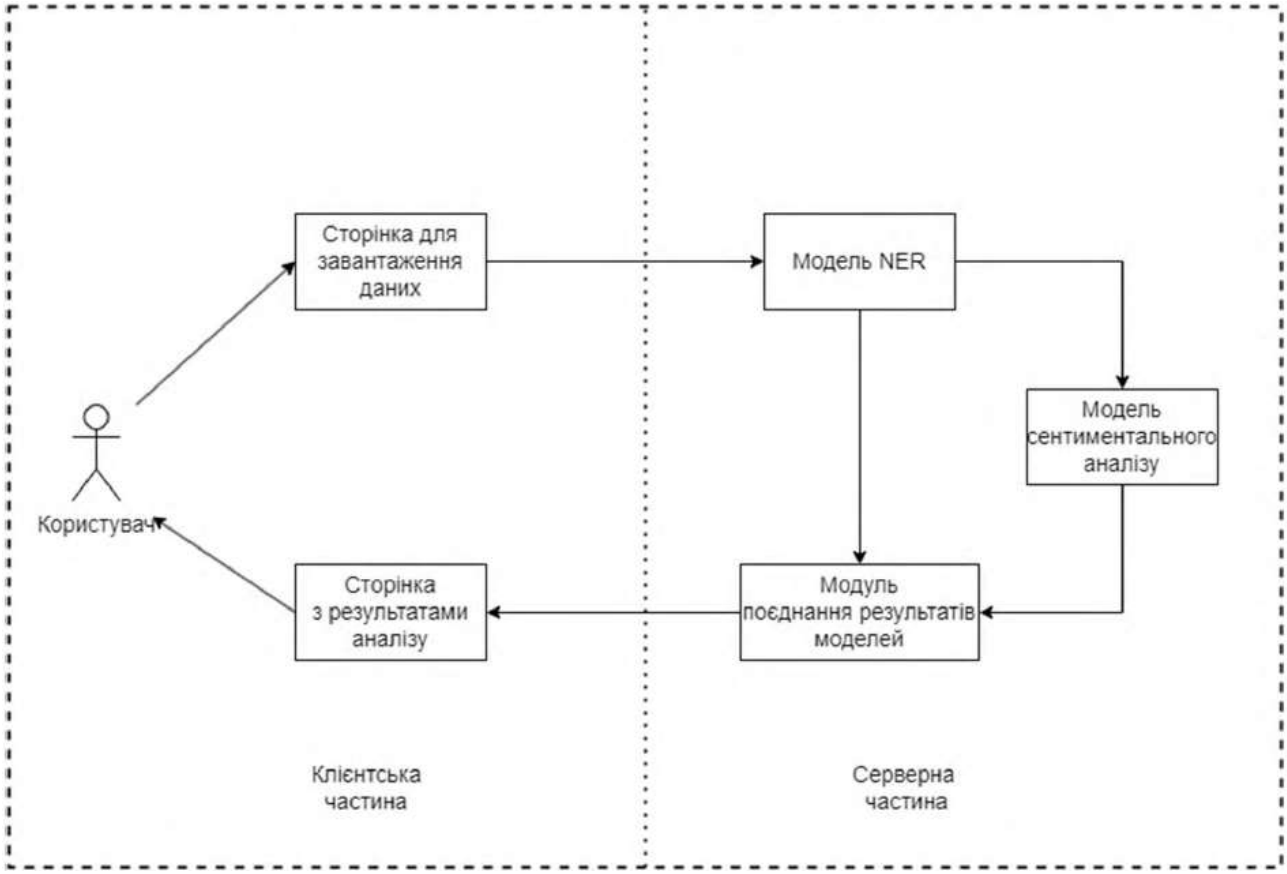


Рисунок 2.9 - Діаграма взаємодії компонентів системи

Користувач надає дані для аналізу, модель NER відокремлює речення які стосуються ESG критеріїв та передає їх у модель сентиментального аналізу, де на основі цих речень визначається оцінка критеріїв, далі формуються результати за цими двома моделей та відображуються користувачу.

РОЗДІЛ 3 РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ ДЛЯ АНАЛІЗУ БІЗНЕС-ЗВІТІВ З ВИКОРИСТАННЯМ ІНСТРУМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ

3.1. Реалізація моделей обробки даних методу

Для аналізу тексту використовуються навчанні моделі машинного навчання одна з яких NER (named entity regoniztion) модель що визначає сутності з тексту та друга сентиментальна модель (sentimental analysis model) яка встановлює оцінку речень за певним критерієм.

NER модель - це модель машинного навчання, яка використовується для виявлення та класифікації іменованих сутностей (Named Entities) у тексті. Named Entities можуть включати імена людей, організації, місця, дати, географічні місця, числа та інші сутності, які мають певну семантичну цінність в контексті тексту [11].

Модель має класифікувати до яких з трьох ESG критеріїв належить речення або словосполучення, для навчання будь-якої моделі необхідні вхідні дані для навчання та тренування NER моделі це набір символів які належать певній категорії, які можна представити у такому форматі (рис. 3.1)

text	label
Carbon emissions have been reduced by 15% of energy-efficient technologies and practices	Ecology
Water consumption was reduced by 20%	Ecology
The company has identified and assessed climate-related risks and opportunities	Social

Рисунок 3.1 – Приклад вхідних даних для NER моделі

Також існують більш складні формати представлення даних для NER моделей, один з яких CoNLL формат [22]. З використанням інструменту Azure ML Data labeling [23] створимо файл CoNLL вручну відокремлюючи з тексту словосполучення, речення які належать до одного з критеріїв «Ecology», «Social» та «Government» (рис. 3.2).

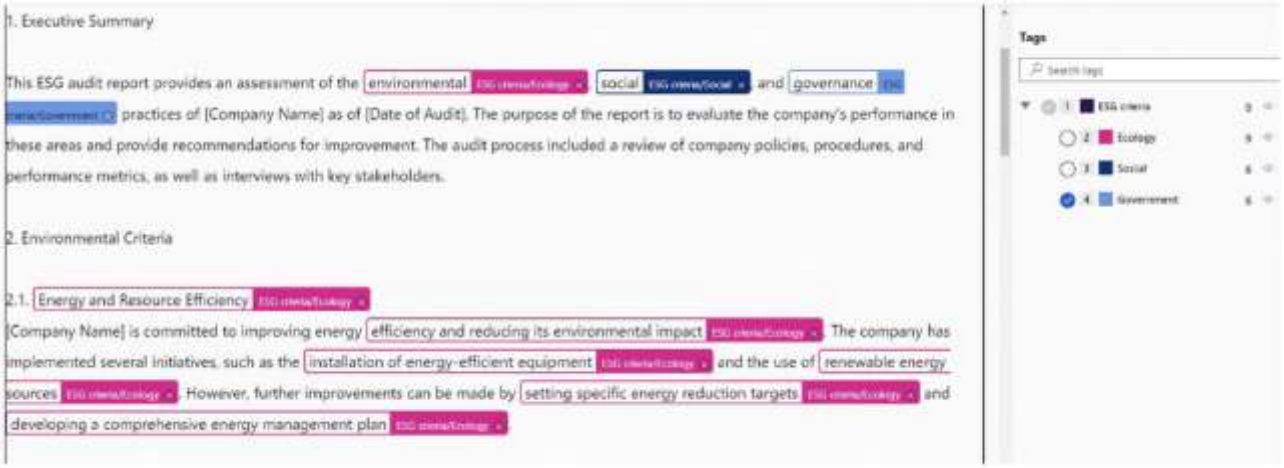


Рисунок 3.2 - Приклад вхідних даних для NER моделі

Після чого у CoNLL форматі речення «The company has identified and assessed climate-related risks and opportunities» (переклад «Компанія ідентифікувало ключові ризики та можливості які пов’язані зі зміною клімату» буде представлено у вигляді (рис. 3.3).

```
The,O
company,B-Ecology
has,I-Ecology
identified,I-Ecology
and,I-Ecology
assessed,I-Ecology
climate-related,I-Ecology
risks,I-Ecology
and,I-Ecology
opportunities,I-Ecology
integrating,O
them,O
into,O
its,O
business,O
strategy,O
```

Рисунок 3.3 - Приклад формату CoNLL

Біля кожного слова встановлюється мітка, мітка «O», позначає що слово ніяк не стосується критерію ESG, «B-» позначають початок сутності та мітка «I» визначає належність до мітки «B», після чого йде вже назва критерію

«Ecology», «Social», «Government», завдяки цьому модель навчається не тільки знаходити схожі слова, а розуміє як повинні починатися та що може містити у собі словосполучення яке стосується критерію.

Наступним етапом є визначення компонентів для створення NER моделі, за допомогою компонента spaCy [24]. Згідно з документаціє процес навчання моделі поділений на такі кроки:

- Токенізація [25]
- Частини мови тегування та аналіз залежностей [26]
- Розпізнавання іменованих сутностей [27]
- Навчання моделі [28]

Токенізація:

Процес розділення тексту на окремі слова, фрази, символи або інші значимі елементи, які називаються токенами, рисунок 3.4 наводить приклад тонізації.

0	1	2	3	4	5	6	7	8	9	10
Apple	is	looking	at	buying	U.K.	startup	for	$	1	billion

Рисунок 3.4 - Приклад тонізації.

Спочатку речення розподіляється на слова та нумерується (кожне слово це токен), spaCy розуміє абрівеатури та скорочення наприклад як бачимо у прикладі (рис. 4.3) «U.K» це окремий токен коли «$1» було розбито на два токени «$» та «1»

Частини мови тегування та аналіз залежностей:

Використовуючи статичні моделі які навчені на великих масивах даних spaCy проводить морфологічний аналіз речення та визначає залежності між словами нижче представлено приклад аналізу (рис. 3.5).

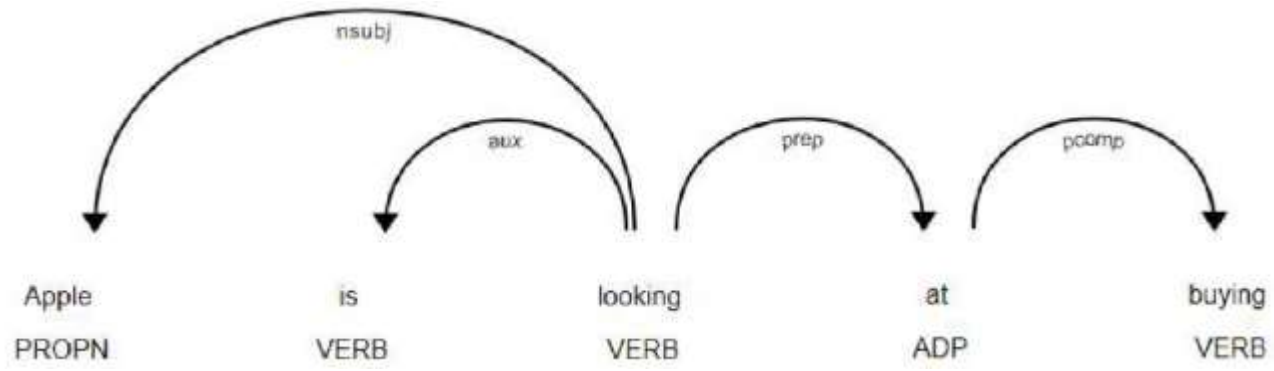


Рисунок 3.5 - Приклад аналізу частин мови та аналіз залежностей

Розпізнавання іменованих сутностей:

Надаючі моделі приклади у форматі CoNLL, модель використовуючи попередні наведені методи будує патерни які містять опис послідовностей частин мов, кожен з патернів належить до певної сутності описаної у CoNLL файлі (рис. 3.6).



Рисунок 3.6 - Приклад визначення сутностей з тексту

Навчання моделі:

Навчання є ітеративним процесом, в якому прогнози моделі порівнюються з посиланнями на анотації для оцінки градієнта втрат. Градієнт втрат потім використовується для розрахунку градієнта ваги за допомогою зворотного розповсюдження. Градієнти вказують, як слід змінити вагові значення, щоб прогнози моделі ставали більш схожими на посилання на мітки з часом.

Під час навчання моделі створюється теорія як визначати певні сутності з тексту. Вхідні дані поділені на дані для тестування та оцінки. Використовують ітеративне навчання у якому після навчання модель відправляють на оцінку. Під час оцінки проводять експеримент де моделі надають приклад тексту де вже знають відповідь як цей текст має визначитися, очікувані результати порівнюють з результати моделі, завдяки чому визначається точність моделі. Якщо точність недостатня вагу нейронів моделей змінюють, так с кожною

ітерацією встановляється залежність між зміною ваги конкретних нейронів до точності оцінки.

Побудуємо діаграму яка описує процес створення моделі NER (рис. 3.7)

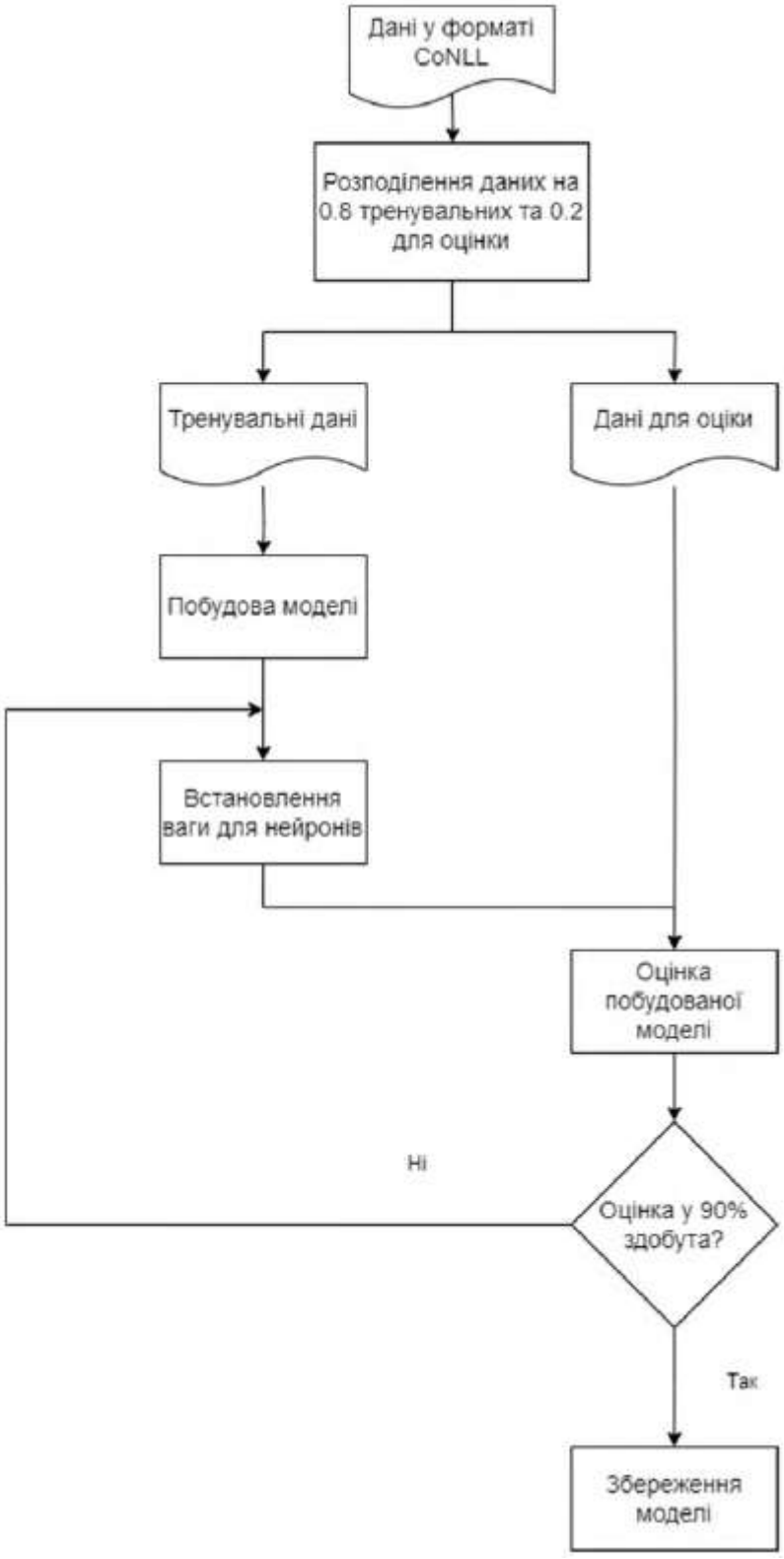


Рисунок 3.7 - Приклад визначення сутностей з тексту

Спочатку вхідні дані розподіляються на дані для навчання та тренування, завдяки даним для навчання визначаються патерни для аналізу та будується модель, встановляється вага нейронів моделі, після чого модель відправляють на тестування де визначають її оцінку, якщо точність моделі нижче 90 відсотків, тоді встановляються нові показники нейронів, так визначають яка вага нейронів впливає на результати моделі, функція навчання моделі наведена у додатку А.

Результат реалізації моделі, вхідні дані та результати аналізу зображено на рисунку 3.8.

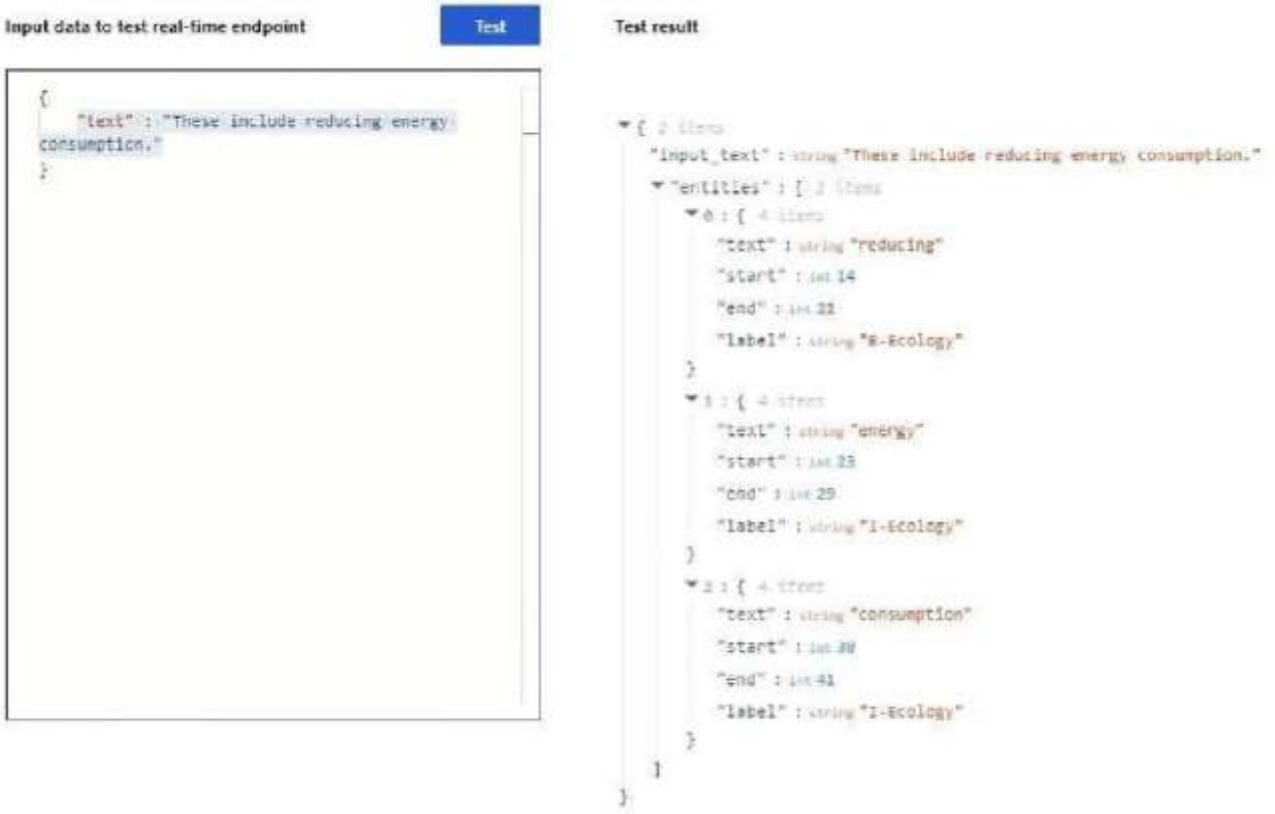


Рисунок 3.8 - Приклад аналізу тексту моделлю NER

Зліва зображено вхідний текст, а саме «These include reducing energy consumption», та результати аналізу кожного слова з його мітками «B-» та «I-». Речення містить екологічну сутність.

Реалізація моделі сентиментального аналізу

Принцип роботи цієї моделі схожий за NER, особливо це стосується навчання моделі проте замість токенізації тексту ця модель використовує техніку векторизації слів [29], завдяки чому кожному слову надається вектор, а

сукупність цих векторів у реченні формує сентиментальну оцінку речення. У даному випадку використовується векторизатор Tf-idf (Term Frequency-Inverse Document Frequency) [30]. Також для цієї моделі у зв'язку з тим що речення треба оцінити у числовому представлені, використовується інший методу розрахунку точності моделі, а саме середньоквадратична помилка «MSE» [31], середню абсолютна помилка «MAE» та «RA2» [32] для навчальних та валідаційних даних.

Модель має визначати вагу речень які характеризують критерії ESG, тому для навчальних даних використаємо вихідні дані моделі NER та згрупуємо їх у формат таблиці (рис. 3.9).

text	label	score
Carbon emissions have been reduced by 15% of energy-efficient technologies and practices	Ecology	0.4
Water consumption was reduced by 20%	Ecology	0.7
donates 20% of its annual profits to philanthropic causes	Social	0.3

Рисунок 3.9 - Приклад навчальних даних сентиментального аналізу

Таблиця містить 3 колонки, перші дві з яких текст та назва критерію, беруться з результатів попередньої моделі, третя колонка містить оцінку цього речення яка визначена за допомогою аналізу та нормалізації даних, оцінка може бути від -1 до 1. Ця модель працює у парі з векторизатором навпроти NER, тому що для визначення речення використовується навчений векторизатор, дані про вагу слів що стосуються критеріїв та самої лінійна модель регресії яка робить передбачення.

Лінійна регресійна модель - це статистична модель, яка використовується для прогнозування залежної змінної на основі однієї або кількох незалежних змінних. Вона використовує лінійну функцію для опису відносин між змінними, функція представлена у вигляді:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta n*xn + \varepsilon$$

- *y* – це залежна змінна, який прогнозує модель, тобто оцінку речення;

- β0 – Точка перетину значення;

- β1 до βn – це коефіцієнти регресії, вони показують, наскільки у (оцінка) змінюється при зміні x (векторів слів);

- x1 до xn - це незалежні змінні (параметри, на основі яких ми прогнозуємо у, а саме вектори слів);

- ε - це випадкова помилка (різниця між реальним і прогнозованим значеннями у).

Графічне представлення формули з урахуванням що «X_vectors» це векторні значення слів згенерований за допомогою векторизатора «Tf-idf» та «Y_sentiment_score», сентиментального значення вхідних даних (від -1 до 1), наведено на рисунку 3.10.

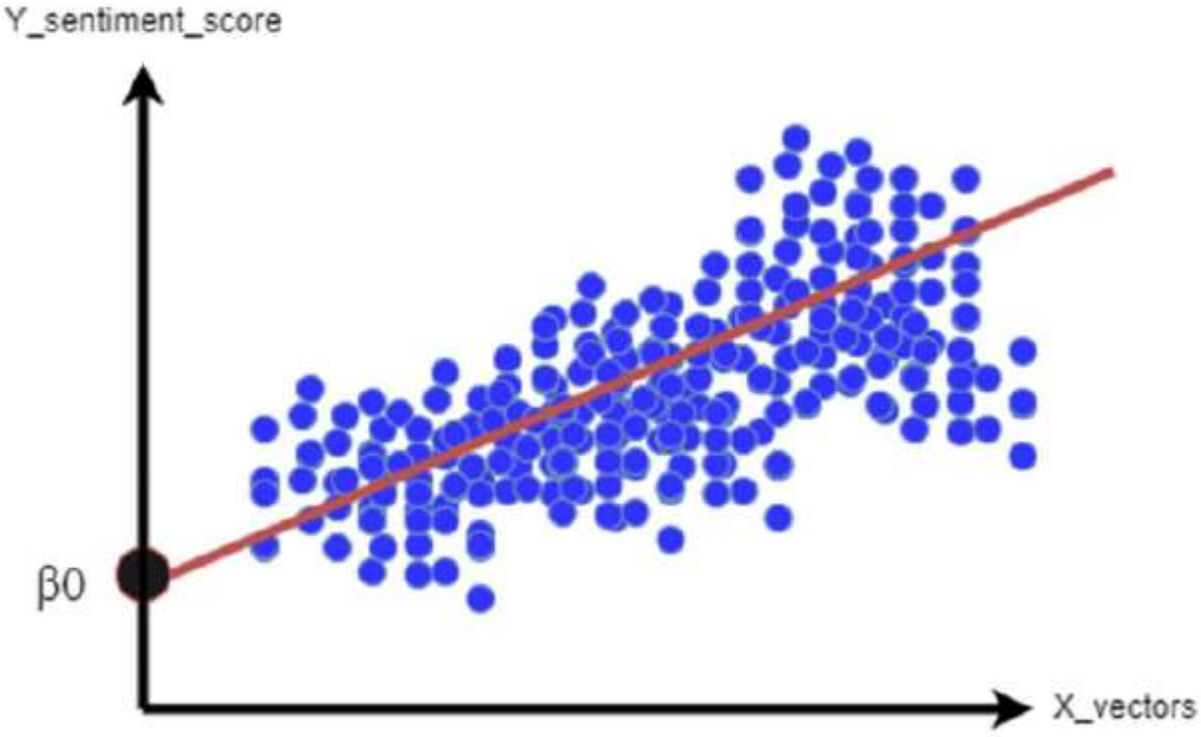


Рисунок 3.10 - Приклад навчальних даних сентиментального аналізу

Зобразимо сам алгоритм навчання моделі на рисунку 3.11, цей алгоритм послідовний.

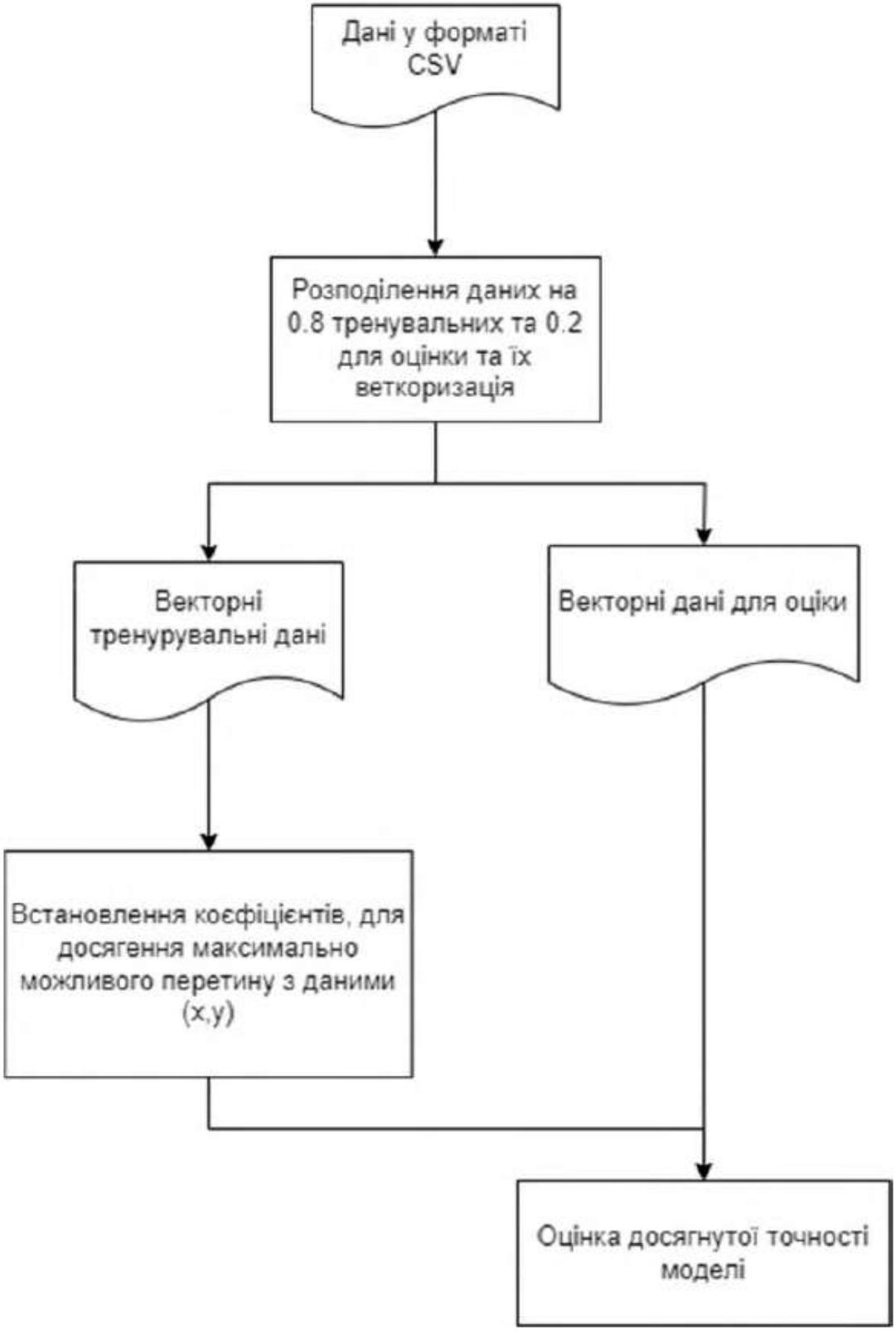


Рисунок 4.11 - Алгоритм навчання моделі оцінювання.

Алгоритм складається з компонентів бібліотек навчання штучного інтелекту, лістинг коду наведено у додатку А.

3.2. Реалізація веб-інтерфейсу методу

Веб-інтерфейс методу має завантажувати текст до NER моделі та оцінювати вихідні речення за допомогою моделі сентиментального аналізу, результати обох моделей поєднати та відобразити користувачеві.

Розгортання серверу здійснено за допомогою Flask [33], фреймворку для розгортання веб-додатків на мові Python [34], акцент на мові Python зроблено у зв'язку простоти розгортання моделей штучного інтелекту. Для початку завантажимо навчені моделі на сервер (рис. 3.12).

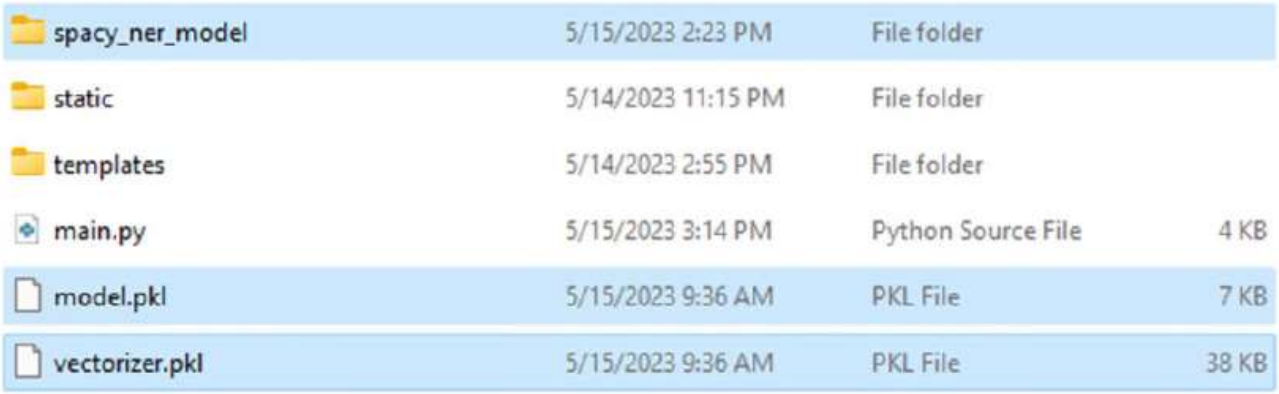


Рисунок 3.12 - Моделі методу у папці сервера.

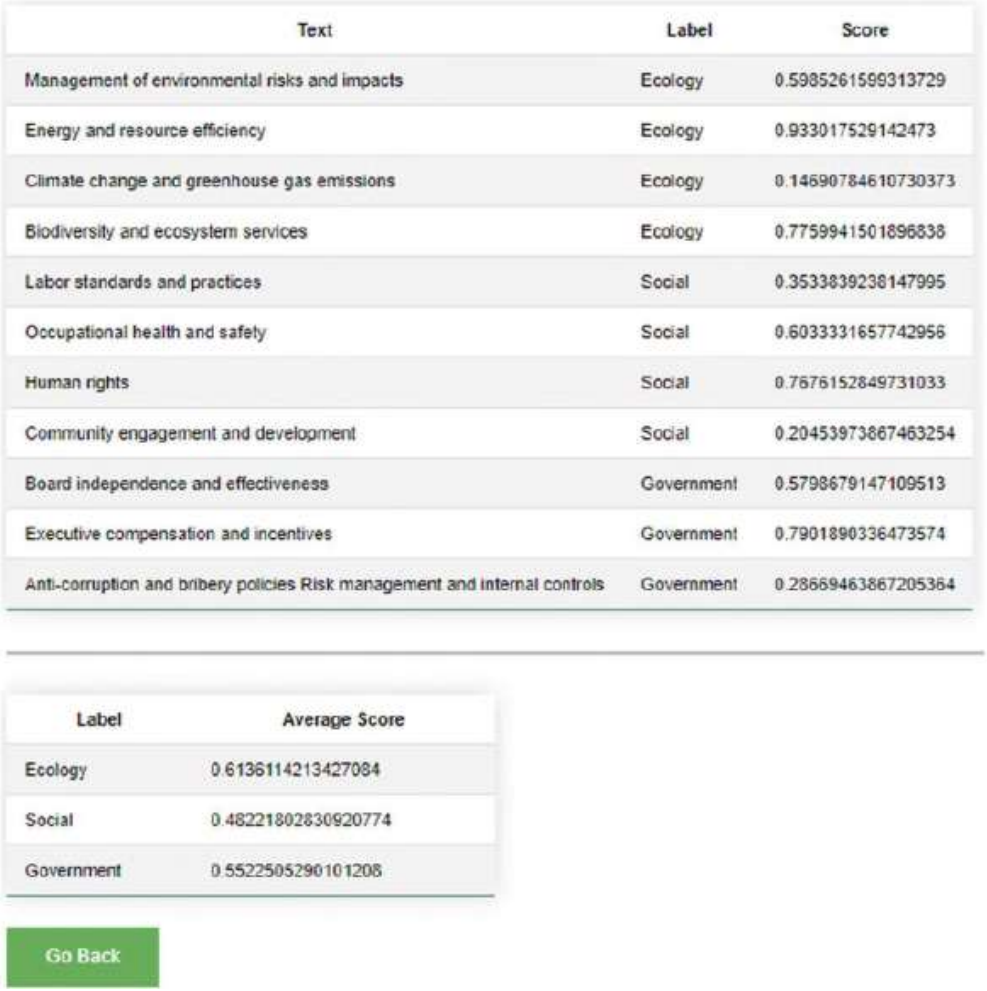
Далі створимо сторінку для завантаження файлу upload.html (рис. 3.13).



Рисунок 3.13 - Вигляд сторінки для завантаження файлу

Сторінка має Drag & Drop зону [35] для завантаження файлу, листування коду сторінки наведено у додатку А. Наступним етапом створимо сторінку для відображення даних, вона має містити 2 таблиці таблиця з реченнями та їх

аналізом та таблицю з підсумками за критеріями, кнопку для повернення до сторінки завантаження вигляд display.html наведено на рисунку 3.14.



Text	Label	Score
Management of environmental risks and impacts	Ecology	0.5985261599313729
Energy and resource efficiency	Ecology	0.933017529142473
Climate change and greenhouse gas emissions	Ecology	0.14690784610730373
Biodiversity and ecosystem services	Ecology	0.7759941501896838
Labor standards and practices	Social	0.3533839238147995
Occupational health and safety	Social	0.6033331657742956
Human rights	Social	0.7676152849731033
Community engagement and development	Social	0.20453973867463254
Board independence and effectiveness	Government	0.5798679147109513
Executive compensation and incentives	Government	0.7901890336473574
Anti-corruption and bribery policies Risk management and internal controls	Government	0.28669463867205364

Label	Average Score
Ecology	0.6136114213427084
Social	0.48221802830920774
Government	0.5522505290101208

Go Back

Рисунок 3.14 - Вигляд сторінки для відображення файлу

Сторінки передають дані через сесію, при завантажені файлу зчитується його текст та передається до моделей, після обробки моделлю відбувається перехід до сторінки відображення.

3.3. Опис алгоритму використання методу

Для зручного завантаження та відображення даних методу використовується веб-інтерфейс, який складається з двох сторінок сторінки завантаження та сторінки відображення. Після завантаження файлу відразу зображується результат аналізу файлу (рис. 3.15).



Рисунок 3.15 - Сторінка завантаження файлу

Відразу після завантаження та обробки файлу відкривається наступна сторінка display.html, яка відображує проаналізовані дані (рис. 3.16).

Text	Label	Score
Management of environmental risks and impacts	Ecology	0.5985261599313729
Energy and resource efficiency	Ecology	0.933017529142473
Climate change and greenhouse gas emissions	Ecology	0.14690784610730373
Biodiversity and ecosystem services	Ecology	0.7759941501896838
Labor standards and practices	Social	0.3533839238147995
Occupational health and safety	Social	0.6033331657742956
Human rights	Social	0.7676152849731033
Community engagement and development	Social	0.20453973867463254
Board independence and effectiveness	Government	0.5798679147109513
Executive compensation and incentives	Government	0.7901890336473574
Anti-corruption and bribery policies Risk management and internal controls	Government	0.28669463867205364

Label	Average Score
Ecology	0.6136114213427084
Social	0.48221802830920774
Government	0.5522505290101208

Go Back

Рисунок 3.16 - Сторінка відображення даних

Сторінка містить кнопку «Go Back» для повторного завантаження нового файлу, після натискання кнопки користувач повертається до сторінки завантаження. Увесь алгоритм взаємодії з веб-інтерфейсу методу представлено на рисунку 3.17.



Рисунок 3.17 - Алгоритм взаємодії з веб-інтерфейсом

3.4. Апробація результатів

Метод дослідження був обраний експериментальний метод (пункт 2.2). Опишемо критерії щодо експерименту, він повинен досліджувати точність знаходження речень які стосуються критеріїв ESG та точність їх оцінки. Показники мають бути дослідженими при різних об’ємах тексту, щоб порівняти як розмір файлу впливає на точність знаходження та оцінки речень методом.

Результати експерименти мають змогу виявити не тільки точність критеріїв, а й якість навчання моделей щодо певного критерію. Експеримент буде проходити за схожим алгоритмом що і оцінка моделей, проте на цей раз буде оцінюватися не показники однієї моделі, результати обробки обох моделей.

Приготуємо вхідні дані для експерименту.

Візьмемо два файли текстових файлів різного розміру, що містять ESG критерії та створимо таблицю очікуваних результатів експерименту.

Перший файл містить невеличку кількість тексту та прості речення, файл містить 92 слова (рис. 3.18).

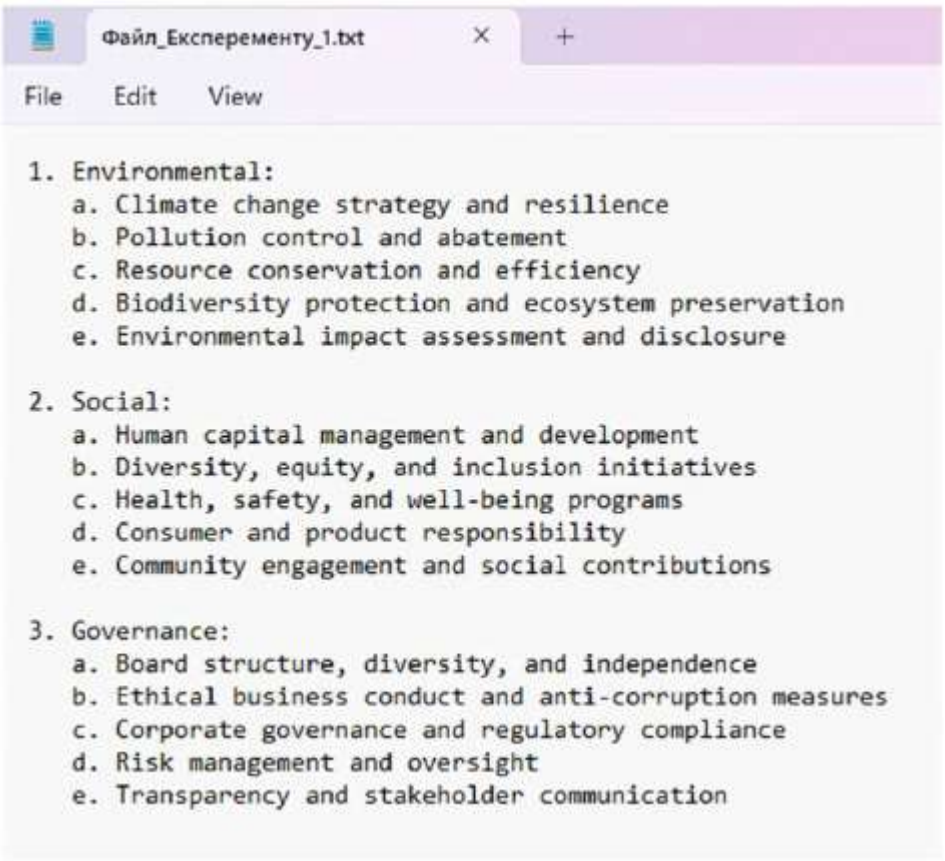


Рисунок 3.18 - Зміст першого файлу вхідних даних

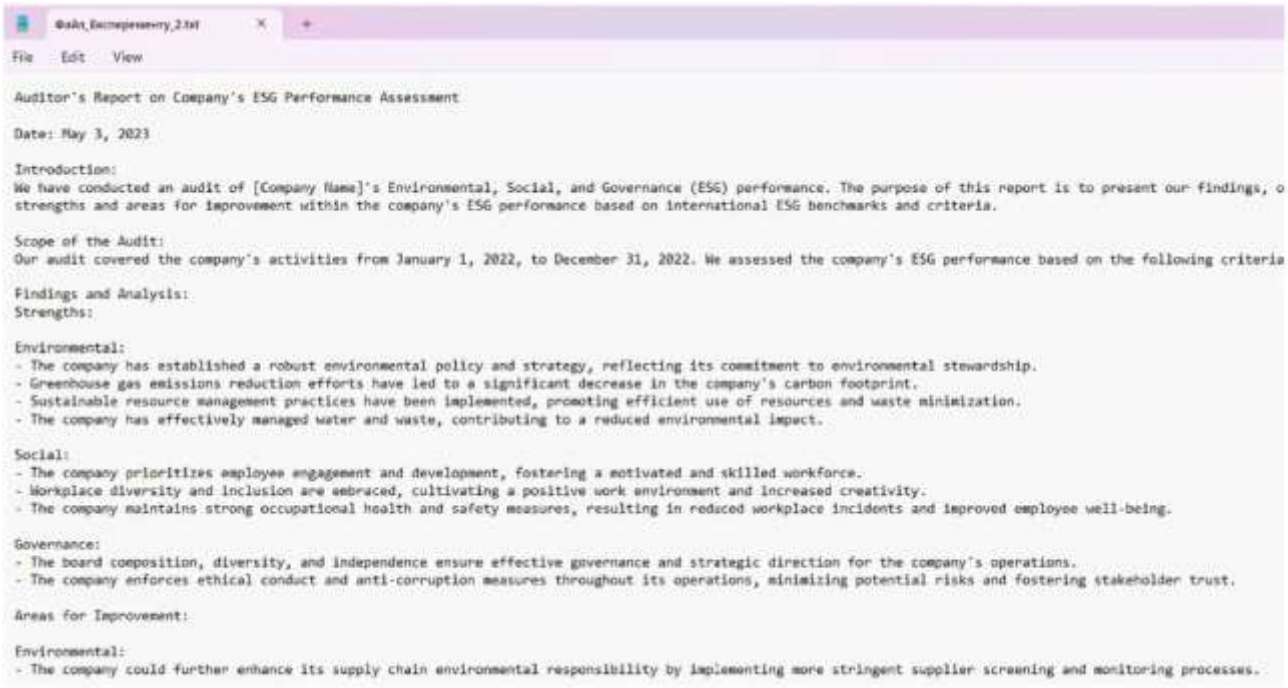
За цим файлом створимо таблицю очікуваних результатів (рис. 3.19)

text	label	score
Climate change strategy and resilience	Ecology	0.3
Pollution control and abatement	Ecology	0.7
Resource conservation and efficiency	Ecology	0.5
Biodiversity protection and ecosystem preservation	Ecology	0.45
Environmental impact assessment and disclosure	Ecology	0.2
Human capital management and development	Social	0.4
Diversity, equity, and inclusion initiatives	Social	0.8
Health, safety, and well-being programs	Social	0.5
Consumer and product responsibility	Social	0.7
Community engagement and social contributions	Social	0.55
Board structure, diversity, and independence	Government	0.5
Ethical business conduct and anti-corruption measures	Government	0.75
Corporate governance and regulatory compliance	Government	0.55
Risk management and oversight	Government	0
Transparency and stakeholder communication	Government	0.52

Ecology	0.43
Social	0.491666667
Government	0.464

Рисунок 3.19 - Таблиця очікуваних результатів першого файлу

Наступний файл буде відрізнятися розміром та складністю речень він містить 396 слів (рис. 3.20)



File Edit View

Auditor's Report on Company's ESG Performance Assessment

Date: May 3, 2023

Introduction:
We have conducted an audit of [Company Name]'s Environmental, Social, and Governance (ESG) performance. The purpose of this report is to present our findings, o
strengths and areas for improvement within the company's ESG performance based on international ESG benchmarks and criteria.

Scope of the Audit:
Our audit covered the company's activities from January 1, 2022, to December 31, 2022. We assessed the company's ESG performance based on the following criteria

Findings and Analysis:
Strengths:

Environmental:
- The company has established a robust environmental policy and strategy, reflecting its commitment to environmental stewardship.
- Greenhouse gas emissions reduction efforts have led to a significant decrease in the company's carbon footprint.
- Sustainable resource management practices have been implemented, promoting efficient use of resources and waste minimization.
- The company has effectively managed water and waste, contributing to a reduced environmental impact.

Social:
- The company prioritizes employee engagement and development, fostering a motivated and skilled workforce.
- Workplace diversity and inclusion are embraced, cultivating a positive work environment and increased creativity.
- The company maintains strong occupational health and safety measures, resulting in reduced workplace incidents and improved employee well-being.

Governance:
- The board composition, diversity, and independence ensure effective governance and strategic direction for the company's operations.
- The company enforces ethical conduct and anti-corruption measures throughout its operations, minimizing potential risks and fostering stakeholder trust.

Areas for Improvement:

Environmental:
- The company could further enhance its supply chain environmental responsibility by implementing more stringent supplier screening and monitoring processes.

Рисунок 3.20 - Зміст другого файлу з великою кількістю тексту

Створимо таблицю очікуваних результатів (рис. 3.21)

text	label	score
The company has established a robust environmental policy and strategy, reflecting its commitment to environmental stewardship.	Ecology	0.85
Greenhouse gas emissions reduction efforts have led to a	Ecology	0.35
significant decrease in the company's carbon footprint.	Ecology	0.6
Sustainable resource management practices have been implemented, promoting efficient use of resources and waste minimization.	Ecology	0.78
The company has effectively managed water and waste, contributing to reduced environmental impact	Ecology	0.5
The company could further enhance its supply chain environmental responsibility by implementing more stringent supplier screening and monitoring processes	Ecology	0.34
The company prioritizes employee engagement and development, fostering a motivated and skilled workforce.	Social	0.35
Workplace diversity and inclusion are embraced, cultivating a positive work environment and increased creativity.	Social	0.55
Consumer satisfaction and product responsibility could be strengthened by increasing focus on product quality, safety, sustainability.	Social	0.18
The company could deepen its commitment to philanthropy and community investment through more targeted initiatives and partnerships.	Social	0.5
The company maintains strong occupational health and safety measures, resulting in reduced workplace incidents and improved employee well-being	Government	0.7
The board composition, diversity, and independence ensure effective governance and strategic direction for the company's operations.	Government	0.4
The company enforces ethical conduct and anti-corruption measures throughout its operations, minimizing potential risks and fostering stakeholder trust.	Government	0.6
Legal compliance and regulatory adherence could be improved by establishing a more comprehensive compliance management system.	Government	0.8
Risk management and internal controls could be refined to better identify, assess, and mitigate emerging risks.	Government	0.6
Shareholder rights and engagement could be further enhanced by facilitating more open and transparent communication channels.	Government	0.7
Ecology		0.57
Social		0.395
Government		0.75

Рисунок 3.21 - Таблиця очікуваних результатів другого файлу

Далі знайдемо точність моделі порівнявши очікуваний результат з дійсним та визначимо залежність швидкості обробки даних. Заносимо вхідні файли до моделі (рис. 3.22).

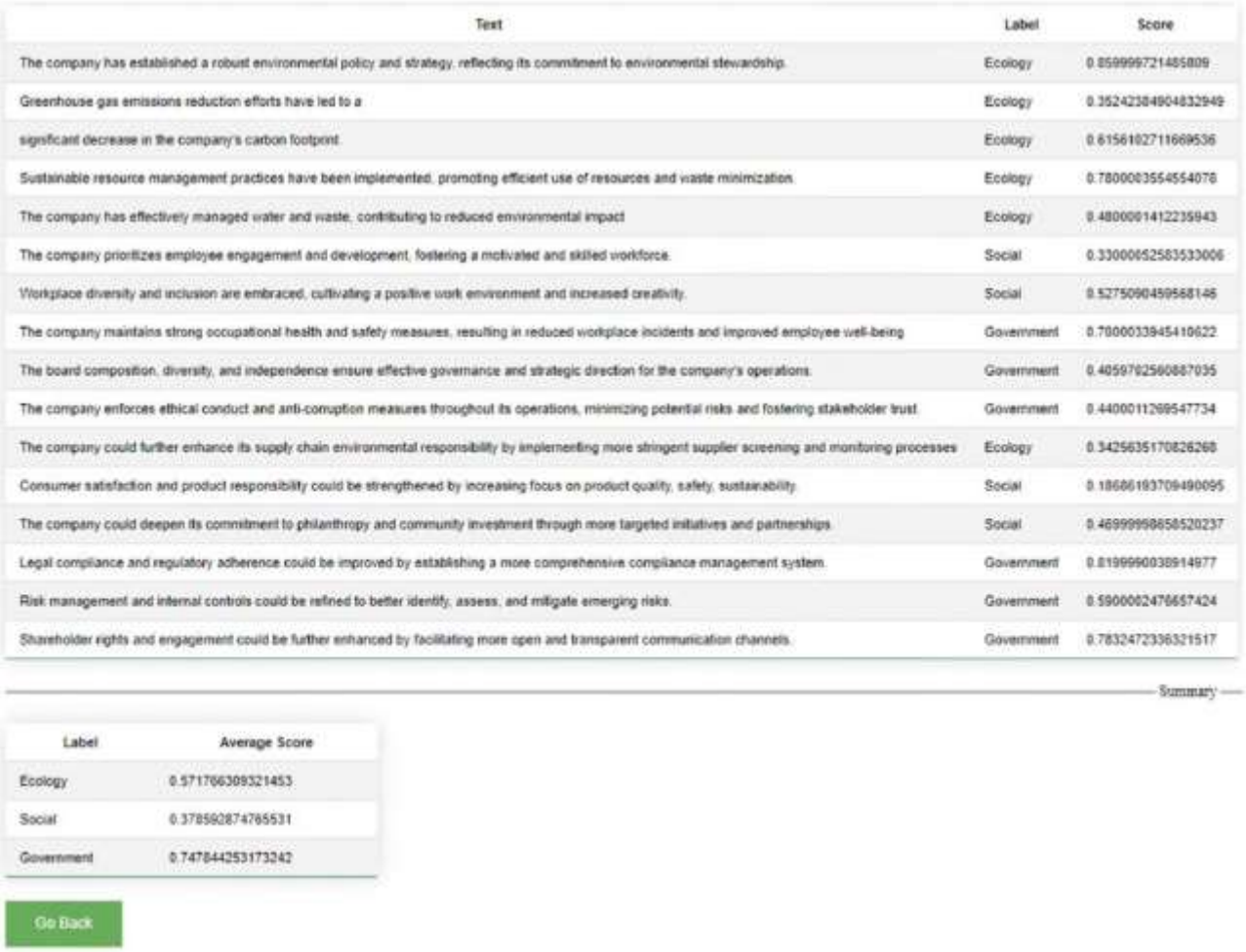
Text	Label	Score
Climate change strategy and resilience	Ecology	0.3899994727513324
Pollution control and abatement c.	Ecology	0.71503279326006915
Resource conservation and efficiency d.	Ecology	0.51772742162279016
Biodiversity protection and ecosystem preservation e.	Ecology	0.39000041870777225
Environmental impact assessment and disclosure	Ecology	0.12831237914156302
Human capital management and development	Social	0.3973023580078604
Diversity, equity,	Social	0.8007125667428762
Health, safety, and well-being programs	Social	0.492124325255214
d. Consumer and product responsibility	Social	0.6922424553275212
Community engagement and social contributions	Social	0.49119855409186537
Board structure, diversity, and independence b.	Government	0.4785814451203696
Ethical business conduct and anti-corruption measures	Government	0.7572643927868278
Corporate governance and regulatory compliance d.	Government	0.555841118262865
Risk management and oversight	Government	0.05474916828354038
e. Transparency and stakeholder communication	Government	0.53412145721533426

Label	Average Score
Ecology	0.42821449731443213
Social	0.44375881867965376
Government	0.476111516214235112

Go Back

Рисунок 3.22 - Таблиця результатів моделі першого файлу.

Наступним етапом завантажимо другий файл який більший у 4 рази (рис. 3.23)



Text	Label	Score
The company has established a robust environmental policy and strategy, reflecting its commitment to environmental stewardship.	Ecology	0.859999721485809
Greenhouse gas emissions reduction efforts have led to a	Ecology	0.3524238490483294
significant decrease in the company's carbon footprint.	Ecology	0.6156102711669536
Sustainable resource management practices have been implemented, promoting efficient use of resources and waste minimization.	Ecology	0.7800003554554076
The company has effectively managed water and waste, contributing to reduced environmental impact.	Ecology	0.4800001412235943
The company prioritizes employee engagement and development, fostering a motivated and skilled workforce.	Social	0.33000652583533006
Workplace diversity and inclusion are embraced, cultivating a positive work environment and increased creativity.	Social	0.5275090459568146
The company maintains strong occupational health and safety measures, resulting in reduced workplace incidents and improved employee well-being.	Government	0.7000033945416622
The board composition, diversity, and independence ensure effective governance and strategic direction for the company's operations.	Government	0.4059702560887035
The company enforces ethical conduct and anti-corruption measures throughout its operations, minimizing potential risks and fostering stakeholder trust.	Government	0.4400011269547734
The company could further enhance its supply chain environmental responsibility by implementing more stringent supplier screening and monitoring processes.	Ecology	0.3425635170826268
Consumer satisfaction and product responsibility could be strengthened by increasing focus on product quality, safety, sustainability.	Social	0.18686193709490095
The company could deepen its commitment to philanthropy and community investment through more targeted initiatives and partnerships.	Social	0.46999998858520237
Legal compliance and regulatory adherence could be improved by establishing a more comprehensive compliance management system.	Government	0.8199990038914977
Risk management and internal controls could be refined to better identify, assess, and mitigate emerging risks.	Government	0.5900002476657424
Shareholder rights and engagement could be further enhanced by facilitating more open and transparent communication channels.	Government	0.7832472336321517

Summary

Label	Average Score
Ecology	0.571766309321453
Social	0.378592874765531
Government	0.747844253173242

Go Back

Рисунок 3.23 - Таблиця результатів моделі другого файлу.

Першим чином треба перевірити кожну з таблиць на однакову кількість знайдених сутностей - таблиці результатів моделей та очікуваних результатів не відрізняються за кількістю сутностей. Наступним етапом порівняємо середнє значення оцінки кожного показника (рис. 3.24).

Label	Average Score		
Ecology	0.428214497		
Social	0.443758819		
Government	0.476111516		
		Модуль різниці точності	Різниця точності у відсотках
Ecology	0.43	0.001785503	0.1786%
Social	0.491666667	0.047907848	4.7908%
Government	0.464	0.012111516	1.2112%

Рисунок 3.24 - Порівння середніх показників

Найбільша різниця точності становить для критерію «Social» у розмірі 4.7%, на другому місці критерій «Government» 1.2%, останнє місце «Ecology» 0.17%.

Порівняємо значення критеріїв для другого файлу (рис 3.25).

Label	Average Score		
Ecology	0.571766309		
Social	0.378592875		
Government	0.747844253		
		Модуль різниці точності	Різниця точності у відсотках
Ecology	0.57	0.001766309	0.1766%
Social	0.395	0.016407125	1.6407%
Government	0.76	0.012155747	1.2156%

Рисунок 3.25 - Порівння середніх показників

У другому файлі такий самий порядок з точності показників, де найбільш точно оцінено критерій «Ecology».

Порівнявши результати обох файлів можна дійти висновку що розмір файлу не впливає на точність результатів, а впливає наявність схожих речень завдяки яким моделі проходили навчання, також завдяки експерименту було визначено що показник «Ecology» має найбільш високу точність порівняно з іншими показниками, тому для подальшого масштабування методу, а саме збільшення якості обробки інформацій потрібно зосередитися на показниках «Social» та «Government».

ВИСНОВКИ

У кваліфікаційній роботі було розроблено метод інтелектуальної обробки даних для оцінки конкурентоспроможності підприємства. Метод обробляє аудиторські тексти, відокремлюючи речення що стосуються ESG критеріїв та наводячи оцінку кожному з критерію.

Для реалізації методу було проаналізовано найближчі аналоги сервісу, визначено їх переваги та недоліки серед яких найкращим аналогом для порівняння став IBM Watson Analytics», який має допоміжний модуль який за запитом аналізує дані, цей аналіз аналогів відокремив основні моменти, на які необхідно звернути увагу, які технології використовувати в процесі проведення розробки методу.

Обрано метод дослідження даних аналізів, а саме кореляційний метод, за допомогою якого встановлюється залежність між змінами та їх вага.

Також було обрано інструмент реалізації методу, а саме платформа Microsoft Azure, яка має багато сервісів для створення моделей штучного інтелекту, навчання, керування та інтеграції з іншими сервісами.

Побудована діаграма IDEF0, в якій було визначено зв’язки між проведенням робіт, необхідними для створення сервісу. На основі діаграми IDEF0 було створено діаграму IDEF1, тобто деталізовану діаграму IDEF0 першого рівня, в якій наведено опис роботи методу аналізу звітності, а саме використання моделі NER та сентиментального аналізу для створення звіту з ESG критеріями. Розроблена діаграма Use Case, за допомогою якої визначатися взаємодія користувача з сервісом для аналізу звітів. Обрано архітектуру «клієнт-сервер» та наведено діаграми взаємодії компонентів системи.

Реалізовано метод з використанням моделей «NER» та сентиментального аналізу, веб-інтрефейс на базі фреймворку «Flask», який завантажує дані для аналізу моделями та поєднує вихідні дані щоб відобразити результати аналізу.

Для методу було проведено тестування у якому зазначено кількісні показники щодо точності моделей стосовно визначення речень з ESG критеріями та оцінкою кожного з критерію.

Практичне значення цієї роботи є визначення ключових фінансових показників з тексту, їх оцінка що значно зменшує витрати для самостійного аналізу, також завдяки реалізації цього методу до нього можна добавити нові показники наприклад як «Стабільність», «Розмір компанії» та «Стратегії розвитку», додаючи нові навчальні дані моделям, моделі методу можуть бути навчені визначати будь-які сутності тексту та оцінювати їх, що значно збільшує простір для аналізу тексту.

СПИСОК ЛІТЕРАТУРИ

1. Artificial intelligence (AI) applications for marketing: A literature-based study / A. Haleem et al. International Journal of Intelligent Networks. 2022. Vol. 3. P. 119-132. URL: https://www.sciencedirect.com/science/article/pii/S2666603022000136 (дата звернення 12.05.2023).

2. ESG criteria: what you need to know. URL: . (дата звернення 10.05.2023). https://greenly.earth/en- us/blog/company-guide/esg-criteria-what-you-need-to-know

3. What is Use Case Diagram? URL: (дата звернення 10.05.2023) https://www.visual-paradigm.com/guide/uml- unified-modeling-language/what-is-use-case-diagram/

4. IDEF0-Part1. URL: . (дата звернення 10.05.2023) http://www.syque.com/quality_tools/tools/Tools19.htm

5. Artificial Intelligence in Business: From Research and Innovation to Market Deployment / N. Soni et al. Procedia Computer Science. 2020. No. 167. P. 22002210. URL: https://www.sciencedirect.com/science/article/pii/S1877050920307389 (дата звернення 16.05.2023).

6. Go Beyond Business Intelligence URL: . (дата звернення 10.05.2023) https://www.sisense.com/about/

7. Sisense Basic Concepts and Terminology URL: . (дата звернення 10.05.2023) https://docs.sisense.com/main/SisenseLinux/sisense-basic-concepts-and-terminology.htm

8. Analytics tools and solutions URL: . (дата звернення 10.05.2023) https://www.ibm.com/analytics

9. About DataRobot URL: . (дата звернення 10.05.2023) https://www.datarobot.com/about-us/

10. Microsoft Azure : Introduction to Machine Learning Studio URL: https://learn.microsoft.com/en-gb/archive/msdn-magazine/2014/september/microsoft-azure-introduction-to-machine-learning- studio. (дата звернення 10.05.2023)

11. Named entity recognition (NER) URL: https://www.techtarget.com/whatis/definition/named-entity-recognition-NER.Natural Language AI URL: https://cloud.google.com/natural-language. (дата звернення 10.05.2023)

12. What is Sentiment Analysis? Definition, Tools, and Applications URL: . (дата звернення 10.05.2023) https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is- sentiment-analysis/

13. What is Experimental Research & How is it Significant for Your Business URL: . (дата звернення 10.05.2023) https://resources.pollfish.com/survey-guides/what-is-experimental-research-how- is-it-significant-for-your-business/

14. Correlational Research: What it is with Examples URL: https://www.questionpro.com/blog/correlational-research/#:~:text=Correlational%20research%20is%20a%20type,influence%20from%20any%20extraneous%20variable. (дата звернення 10.05.2023)

15. Questionnaire Survey URL: . (дата звернення 10.05.2023) https://www.sciencedirect.com/topics/earth-and- planetary-sciences/questionnaire-survey

16. Google Cloud API URL: . (дата звернення 10.05.2023) https://cloud.google.com/natural-language#section-5

17. Watson Natural Language Understanding URL: . (дата звернення 10.05.2023) https://www.ibm.com/cloud/watson-natural-language-understanding

18. Text Razor Extract Meaning from your Text URL: . (дата звернення 10.05.2023) https://www.textrazor.com/

19. Stanford Named Entity Recognizer (NER) URL: . (дата звернення 10.05.2023) https://nlp.stanford.edu/software/CRF-NER.shtml

20. IDEF0 diagram - Decomposition structure URL: . (дата звернення 10.05.2023) https://www.conceptdraw.com/examples/decomposition-structure

21. What is Client Server Architecture? URL: . (дата звернення 10.05.2023) https://intellipaat.com/blog/what-is- client-server-architecture/?US

22. Stranak P., Stepanek J. Proceedings of the Second International Conference on Global Interoperability for Language Resources. Proceedings of the Second International Conference on Global Interoperability for Language Resources, Praha, 16 January 2010. Praha, 2010. P. 10. URL: https://www.researchgate.net/publication/307174605_Representing_Layered_and_Structured_Data_in_the_CoNLL-ST_Format (дата звернення 16.05.2023).

23. Labeling images and text documents. Microsoft Learn. URL: (дата звернення 16.05.2023). https://learn.microsoft.com/en-us/azure/machine-learning/how-to-label- data?view=azureml-api-2

24. Everything you need to know. spaCy. URL: (дата звернення 16.05.2023). https://spacy.io/usage/spacy-101

25. Friedman R. Tokenization in the Theory of Knowledge. Encyclopedia. 2023. Vol. 3, no. 1. P. 380-386. URL: (дата звернення 16.05.2023). https://doi.org/10.3390/encyclopedia3010024

26. Improving Code-mixed POS Tagging Using Code-mixed Embeddings / S. N. Bhattu et al. ACM Transactions on Asian and Low-Resource Language Information Processing. 2020. Vol. 19, no. 4. P. 1-31. URL: (дата звернення 16.05.2023). https://doi.org/10.1145/3380967

27. Low-Resource Named Entity Recognition via the Pre-Training Model / S. Chen et al. Symmetry. 2021. Vol. 13, no. 5. P. 786. URL: (дата звернення 16.05.2023). https://doi.org/10.3390/sym13050786

28. Automated Source Code Generation and Auto-Completion Using Deep Learning: Comparing and Discussing Current Language Model-Related Approaches / J. Cruz-Benito et al. AI. 2021. Vol. 2, no. 1. P. 1-16. URL: (дата звернення 16.05.2023). https://doi.org/10.3390/ai2010001

29. Kozhevnikov V. A., Pankratova E. S. RESEARCH OF THE TEXT DATA VECTORIZATION AND CLASSIFICATION ALGORITHMS OF MACHINE LEARNING. Theoretical & Applied Science. 2020. Vol. 85, no. 05. P. 574-585. URL: (date of access: 16.05.2023). https://doi.org/10.15863/tas.2020.05.85.106

30. How sklearn’s Tfidfvectorizer Calculates tf-idf Values. Analytics Vidhya. URL: (дата звернення 16.05.2023). https://www.analyticsvidhya.com/blog/2021/11/how-sklearns-tfidfvectorizer-calculates-tf-idf-values/

31. Hodson T. O., Over T. M., Foks S. S. Mean Squared Error, Deconstructed. Journal of Advances in Modeling Earth Systems. 2021. Vol. 13, no. 12. URL: (дата звернення 16.05.2023). https://doi.org/10.1029/2021ms002681

32. Hodson T. O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. 2022. Vol. 15, no. 14. P. 5481-5487. URL: (дата звернення 16.05.2023). https://doi.org/10.5194/gmd-15-5481-2022

33. Welcome to Flask - Flask Documentation (2.3.x). Welcome to Flask - Flask Documentation (2.3.x). URL: (дата звернення 16.05.2023). https://flask.palletsprojects.com/en/2.3.x/

34. IntroductoryBooks - Python Wiki. Python Software Foundation Wiki Server. URL: (дата звернення 16.05.2023). https://wiki.python.org/moin/IntroductoryBooks

35. File drag and drop - Web APIs | MDN. MDN Web Docs. URL: https://developer.mozilla.org/en-US/docs/Web/API/HTML_Drag_and_Drop_API/File_drag_and_drop (дата звернення 16.05.2023).

ДОДАТКИ

ДОДАТОК А

Файл реалізації моделі визначення сутностей «NER_Model.py»:

```
import pandas as pd
import numpy as np
import joblib
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score
from azureml.core import Workspace
subscription_id = 'subscription_id'
resource_group = 'resource_group'
workspace_name = 'workspace_name'
ws = Workspace(subscription_id, resource_group, workspace_name)
data = pd.read_csv('output_sentimental.csv')
train_data, val_data = train_test_split(data, test_size=0.2, random_state=42)
vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_data['text'])
X_val = vectorizer.transform(val_data['text'])
y_train = train_data['sentiment_score']
y_val = val_data['sentiment_score']
model = LinearRegression()
model.fit(X_train, y_train)
y_pred_train = model.predict(X_train)
y_pred_val = model.predict(X_val)
mse_train = mean_squared_error(y_train, y_pred_train)
mse_val = mean_squared_error(y_val, y_pred_val)
mae_train = mean_absolute_error(y_train, y_pred_train)
mae_val = mean_absolute_error(y_val, y_pred_val)
r2_train = r2_score(y_train, y_pred_train)
r2_val = r2_score(y_val, y_pred_val)
print("Training MSE:", mse_train)
print("Validation MSE:", mse_val)
print("Training MAE:", mae_train)
print("Validation MAE:", mae_val)
print("Training R^2:", r2_train)
print("Validation R^2:", r2_val)
joblib.dump(model, 'model.pkl')
joblib.dump(vectorizer, 'vectorizer.pkl')
```

Файл реалізації сентиментальної моделі «Sentiment_Model.py»:

```
import pandas as pd
import numpy as np
import joblib
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score
from azureml.core import Workspace
subscription_id = 'subscription_id'
resource_group = 'resource_group'
workspace_name = 'workspace_name'
ws = Workspace(subscription_id, resource_group, workspace_name)
data = pd.read_csv('output_sentimental.csv')
train_data, val_data = train_test_split(data, test_size=0.2, random_state=42)
vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_data['text'])
X_val = vectorizer.transform(val_data['text'])
y_train = train_data['sentiment_score']
y_val = val_data['sentiment_score']
model = LinearRegression()
model.fit(X_train, y_train)
y_pred_train = model.predict(X_train)
y_pred_val = model.predict(X_val)
mse_train = mean_squared_error(y_train, y_pred_train)
mse_val = mean_squared_error(y_val, y_pred_val)
mae_train = mean_absolute_error(y_train, y_pred_train)
mae_val = mean_absolute_error(y_val, y_pred_val)
r2_train = r2_score(y_train, y_pred_train)
r2_val = r2_score(y_val, y_pred_val)
print("Training MSE:", mse_train)
print("Validation MSE:", mse_val)
print("Training MAE:", mae_train)
print("Validation MAE:", mae_val)
print("Training R^2:", r2_train)
print("Validation R^2:", r2_val)
new_data = ["Ecology is very impotant for our company"]
X_new = vectorizer.transform(new_data)
y_pred_new = model.predict(X_new)
print("Predicted sentiment scores:", y_pred_new)
# Assuming 'model' is your trained model
joblib.dump(model, 'model.pkl')
```

```
joblib.dump(vectorizer, 'vectorizer.pkl')
```

Виконавчий файл серверу метода «main.py»:

```
from flask import Flask, render_template, request, redirect, url_for, session,
jsonify
import requests
import os
import json
from collections import defaultdict
import joblib
import numpy as np
import sklearn
import spacy
app = Flask(__name__)
app.secret_key = 'your_secret_key'
model = joblib.load('model.pkl')
vectorizer = joblib.load('vectorizer.pkl')
nlp = spacy.load('spacy_ner_model')
@app.route('/predict', methods=['POST'])
def predict():
data = request.get_json(force=True)
data_transformed = vectorizer.transform(np.array(data['data']))
prediction = model.predict(data_transformed)
return jsonify(prediction.tolist())
def process_json_file(ner_output):
entities = ner_output
# Combine B- and I- entities
combined_entities = []
current_entity = []
for entity in entities:
label = entity["label"]
if label.startswith("B-"):
if current_entity:
combined_entities.append(current_entity)
current_entity = [entity]
elif label.startswith("I-"):
current_entity.append(entity)
if current_entity:
combined_entities.append(current_entity)
sentiment_data = []
for entity_group in combined_entities:
entity_text = " ".join([entity["text"] for entity in entity_group])
label = entity_group[0]["label"].split("-")[1]
# Vectorize the text and make the prediction
```

```
entity_text_transformed = vectorizer.transform(np.array([entity_text]))
prediction = model.predict(entity_text_transformed)
sentiment_data.append({"text": entity_text, "label": label, "score":
prediction[0]}
)
return sentiment_data
@app.route('/')
def upload_file_view():
return render_template('upload.html')
@app.route('/upload', methods = ['POST'])
def upload_file():
if 'file' not in request.files:
return 'No file part', 400
file = request.files['file']
if file.filename == '':
return 'No selected file', 400
if file:
text = file.read().decode('utf-8')
doc = nlp(text)
entities = [{"text": ent.text, "start": ent.start_char, "end": ent.end_char, "label
": ent.label_} for ent in doc.ents]
session['text'] = text
session['model_output'] = process_json_file(entities)
return redirect(url_for('display_text'))
@app.route('/display')
def display_text():
model_output = session.get('model_output', [])
scores = defaultdict(list)
# Calculate the average score for each label
for item in model_output:
scores[item['label']].append(item['score'])
avg_scores = {label: sum(values) / len(values) for label, values in
scores.items()}
return render_template('display.html', model_output=model_output,
avg_scores=avg_scores
)
if __name__ == '__main__':
app.run(debug=True)
```

Файл сторінки завантаження «upload.html»:

```
<!DOCTYPE html>
<html>
<head>
<title>Upload File</title>
```

```
<!-- Dropzone CSS -->
<link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/dropzone/5.9.2/min/
dropzone.min.css">
<!-- Custom CSS -->
<link rel="stylesheet" href="{{ url_for('static', filename='upload_styles.css')
}}">
</head>
<body>
<form class="dropzone" id="myDropzone" action="/upload">
<div class="dz-message" data-dz-message><span>Drag & Drop ESG
file</span></div>
</form>
<!-- Dropzone JS -->
<script
src="https://cdnjs.cloudflare.com/ajax/libs/dropzone/5.9.2/min/dropzone.min.j
s"
></script>
<script>
// Configure Dropzone to accept only .txt files
Dropzone.options.myDropzone = {
url: "/upload",
acceptedFiles: '.txt',
init: function() {
this.on("sending", function(file, xhr, formData) {
// Read the file content
var reader = new FileReader();
reader.onload = function(event) {
// Append the file content to the form data
formData.append("text", event.target.result);
};
reader.readAsText(file);
});
this.on("success", function(file, response) {
// Redirect the user to the new page
window.location.href = "/display";
});
}
};
</script>
</body>
</html>
```

Файл стилів сторінки завантаження «upload_styles.css»:

```
body {
display: flex;
justify-content: center;
align-items: center;
height: 100vh;
background-color: #f3f3f3;
margin: 0;
padding: 0;
box-sizing: border-box;
font-family: Arial, sans-serif;
}
.dropzone {
width: 300px;
height: 200px;
border: 2px dashed #007BFF;
background-color: #fff;
position: relative;
}
.dz-message {
color: #007BFF;
}
```

Файл сторінки відображення «display.html»:

```
<!DOCTYPE html>
<html>
<head>
<title>Display File</title>
<link rel="stylesheet" href="{{ url_for('static', filename='display_styles.css')
}}">
</head>
<body>
<table class="styled-table">
<tr>
<th>Text</th>
<th>Label</th>
<th>Score</th>
</tr>
{% for item in model_output %}
<tr>
<td>{{ item.text }}</td>
<td>{{ item.label }}</td>
<td>{{ item.score }}</td>
</tr>
{% endfor %}
```

```
</table>
<div class="separator">Summary</div>
<body>
<table class="styled-table">
<tr>
<th>Label</th>
<th>Average Score</th>
</tr>
{% for label, avg_score in avg_scores.items() %}
<tr>
<td>{{ label }}</td>
<td>{{ avg_score }}</td>
</tr>
{% endfor %}
</table>
<button class="button" onclick="location.href='/'">Go Back</button>
</body>
</html>
```

Файл стилів сторінки завантаження «display_styles.css»:

```
.styled-table {
border-collapse: collapse;
margin: 25px 0;
font-size: 0.9em;
font-family: sans-serif;
min-width: 400px;
box-shadow: 0 0 20px rgba(0, 0, 0, 0.15);
}
.styled-table thead tr {
background-color: #009879;
color: #ffffff;
text-align: left;
}
.styled-table th,
.styled-table td {
padding: 12px 15px;
}
.styled-table tbody tr {
border-bottom: 1px solid #dddddd;
}
.styled-table tbody tr:nth-of-type(even) {
background-color: #f3f3f3;
}
.styled-table tbody tr:last-of-type {
```

```
border-bottom: 2px solid #009879;
}
.styled-table tbody tr.active-row {
font-weight: bold;
color: #009879;
}
.button {
background-color: #4CAF50; /* Green */
border: none;
color: white;
padding: 15px 32px;
text-align: center;
text-decoration: none;
display: inline-block;
font-size: 16px;
}
.separator {
display: flex;
align-items: center;
text-align: center;
}
.separator::before,
.separator::after {
content: '';
flex: 1;
border-bottom: 5px solid #bbb;
}
.separator:not(:empty)::before {
margin-right: .25em;
}
.separator:not(:empty)::after {
margin-left: .25em;
}
```