МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

Маркопольський Сергій Володимирович

Інтелектуальний метод класифікації наслідків техногенних катастроф / Intelligent Method for Classification of Technogenic Disaster Consequences

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КНм-21
Маркопольський С. В.
_____

Науковий керівник:
к.т.н., Загородня Д.І.
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 20___ р.
Завідувач кафедри
_____ М.П. Комар

ТЕРНОПІЛЬ - 2023

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «магістр»
спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
Маркопольському Сергію Володимировичу
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи
Інтелектуальний метод класифікації наслідків техногенних катастроф /
Intelligent Method for Classification of Technogenic Disaster Consequences
керівник роботи к.т.н., Загородня Д.І.
затверджені наказом по університету від 8 грудня 2022 року № 491.

2. Строк подання студентом закінченої кваліфікаційної роботи 1 грудня 2023 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити
1. Визначити суть та особливості поняття техногенної катастрофи
2. Провести огляд існуючих рішень
3. Описати методи обробки тексту
4. Описати бустингові методи
5. Розробити інтелектуальний метод класифікації наслідків техногенних катастроф
6. Провести попередній аналіз даних
7. Провести попередній текстових даних
8. Реалізувати класифікацію наслідків техногенних катастроф в Україні

5. Перелік графічного матеріалу у роботі
Концептуальна структура інтелектуального методу класифікації техногенних катастроф
Співвідношення між рівнем потенційної аварії та явним рівнем аварії
Кореляція між рівнем потенційної аварії та явним рівнем аварії
Результат передбачення рівня потенційної небезпеки

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 8 грудня 2022 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Сучасний стан предметної області і постановка задачі дослідження	12.2022 р. – 03.2023 р.	
2	Інтелектуальний метод класифікації наслідків техногенних катастроф	03.2023 р. – 05.2023 р.	
3	Реалізація інтелектуального методу	05.2023 р. – 11.2023 р.	
4	Повне завершення та представлення кваліфікаційної роботи на кафедру	01.12.2023 р.	

Студент _____ Маркопольський С. В.
підпис

Керівник роботи _____ к.т.н., Загородня Д.І.
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему "Інтелектуальний метод класифікації наслідків техногенних катастроф" у напрямку отримання ступеня вищої освіти "Магістр" за спеціальністю 122 "Комп'ютерні науки" включає 86 сторінок тексту, 22 рисунків, 1 таблиць та список з 39 джерел.

Мета дослідження полягала у розробці та застосуванні інтелектуального методу класифікації наслідків техногенних катастроф на основі аналізу текстових та кількісних даних для покращення управління ризиками та реагування на кризові ситуації.

У роботі використано методи обробки тексту, машинного навчання та аналізу даних для розробки та реалізації інтелектуального методу класифікації.

Отримані результати свідчать про високий потенціал інтелектуального підходу до класифікації наслідків техногенних катастроф, що може сприяти покращенню управління кризовими ситуаціями та зменшенню їхніх негативних наслідків.

Розроблені методи мають практичне застосування в органах цивільного захисту, екологічних службах та інших організаціях, що відповідають за управління ризиками техногенних катастроф.

Ключові слова: ТЕХНОГЕННІ КАТАСТРОФИ, АНАЛІЗ ТЕКСТУ, МАШИННЕ НАВЧАННЯ, УПРАВЛІННЯ РИЗИКАМИ.

RESUME

Qualification work on the topic "Intelligent Method for Classification of Consequences of Technogenic Disasters" in the field of obtaining a Master's degree in Computer Science includes 86 pages of text, 22 figures, 1 table, and a list of 39 sources.

The aim of the research was to develop and apply an intelligent method for classifying the consequences of technogenic disasters based on the analysis of textual and quantitative data to improve risk management and crisis response.

In the work, methods of text processing, machine learning, and data analysis were employed to develop and implement the intelligent classification method.

The obtained results demonstrate the high potential of the intelligent approach to classifying the consequences of technogenic disasters, which can contribute to enhancing crisis management and reducing their negative impacts.

The developed methods have practical applications in civil defense agencies, environmental services, and other organizations responsible for managing the risks of technogenic disasters.

Keywords: TECHNOGENIC DISASTERS, TEXT ANALYSIS, MACHINE LEARNING, RISK MANAGEMENT.

ЗМІСТ

ВСТУП

Актуальність даного дослідження полягає в надзвичайно важливому і актуальному завданні: забезпеченні ефективного та оперативного управління наслідками техногенних катастроф, що є серйозними загрозами для суспільства та навколишнього середовища. Зростання промислової діяльності та розвиток нових технологій призводять до збільшення кількості техногенних катастроф, що підкреслює необхідність ефективних методів їх класифікації та оцінки наслідків. Застосування інтелектуальних методів та сучасних технологій аналізу даних дозволяє покращити точність, швидкість та оперативність в управлінні ризиками, пов'язаними з цими катастрофами.

Дослідження має особливу актуальність у зв'язку зі зростанням індустріалізації та розвитком нових технологій у всьому світі. Збільшення обсягів виробництва та використання складних технічних систем підвищує ризик виникнення техногенних аварій та катастроф. Застосування інтелектуальних методів класифікації наслідків таких подій дозволяє підвищити якість управління кризовими ситуаціями та запобігти подібним подіям в майбутньому. Результати дослідження можуть бути корисними для урядових структур, рятувальних служб, промислових компаній та інших зацікавлених сторін у вдосконаленні систем управління та підвищенні рівня безпеки в промисловості та інших сферах.

Крім того, актуальність дослідження посилюється зростанням обсягів інформації, яка створюється внаслідок технологічних катастроф, і необхідністю швидкого та точного аналізу цієї інформації для прийняття рішень. Інтелектуальні методи класифікації наслідків допомагають автоматизувати цей процес та надають можливість оперативно реагувати на кризові ситуації. Таким чином, дослідження має велике значення для підвищення ефективності управління наслідками техногенних катастроф

та зменшення їх негативного впливу на суспільство та природне середовище.

Метою дослідження є розробка та впровадження інтелектуального методу класифікації наслідків техногенних катастроф з метою покращення управління кризовими ситуаціями та запобігання подібним подіям у майбутньому.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Розглянути сутність та особливості поняття "техногенна катастрофа" з метою уявлення про їхні характеристики та відмінності від природних катастроф.
2. Здійснити огляд існуючих рішень та методів управління техногенними катастрофами з метою визначення їхніх переваг та обмежень.
3. Провести детальний опис методів обробки текстових даних, зокрема включаючи такі аспекти як видалення стоп-слів, токенізація, лематизація, нормалізація, визначення та візуалізація N-грам та інші.
4. Провести огляд бустингових методів, таких як AdaBoost, Gradient Boost, XGBoost, та визначити їхню ефективність у вирішенні завдань класифікації.
5. Розробити інтелектуальний метод класифікації наслідків техногенних катастроф, який базується на обробці текстових даних та використанні бустингових алгоритмів.
6. Провести попередній аналіз даних, які будуть використовуватися в дослідженні, з метою зрозуміння їхнього обсягу та особливостей.
7. Виконати попередній аналіз текстових даних, які будуть використовуватися для класифікації, з метою визначення ключових особливостей текстів.

8. Реалізувати метод класифікації наслідків техногенних катастроф в Україні з використанням розроблених алгоритмів та методів для аналізу та класифікації текстових даних.

Об’єктом дослідження є процес техногенних катастроф, що виникають через людські помилки, технічні недоліки або збої в технологіях. Дослідження спрямоване на вивчення цього процесу з метою розробки інтелектуального методу класифікації наслідків техногенних катастроф.

Предметом дослідження є методи обробки текстових даних та бустингові методи, такі як AdaBoost, Gradient Boost, XGBoost, які можуть бути використані для класифікації наслідків техногенних катастроф. Дослідження спрямоване на вивчення цих методів з метою їхнього подальшого використання у розробці інтелектуального підходу до класифікації.

Методи дослідження включають в себе аналіз існуючих рішень та методів управління техногенними катастрофами, детальний опис методів обробки текстових даних, огляд бустингових методів та їхній аналіз, попередній аналіз даних, включаючи текстові дані, та реалізацію інтелектуального методу класифікації наслідків техногенних катастроф. Дослідження проводиться з метою розробки ефективного інструменту для аналізу та класифікації наслідків техногенних катастроф та покращення управління кризовими ситуаціями.

Наукова новизна отриманих результатів полягає у розробці та застосуванні інтелектуального методу класифікації наслідків техногенних катастроф, який базується на комплексному аналізі текстових даних та використанні бустингових методів, таких як AdaBoost, Gradient Boost, XGBoost. Цей метод спроможний точно та швидко оцінювати рівень ризику та класифікувати наслідки техногенних катастроф, що є важливим для ефективного управління в надзвичайних ситуаціях. Крім того,

застосування сучасних технологій обробки тексту та аналізу даних, включаючи інтелектуальний аналіз даних з соціальних мереж, робить цей підхід більш ефективним та точним у порівнянні з існуючими методами управління техногенними катастрофами.

Практичне значення отриманих результатів полягає в їхній застосовності для покращення системи моніторингу, аналізу та управління наслідками техногенних катастроф. Розроблений інтелектуальний метод класифікації дозволяє автоматизувати процес оцінки ризику та класифікації подій, що відбуваються під час кризових ситуацій, що в свою чергу спрощує та прискорює процеси прийняття рішень. Це допомагає оперативно та ефективно реагувати на надзвичайні ситуації, визначаючи пріоритети та ресурси для нейтралізації наслідків катастроф.

Структура та обсяг роботи. Кваліфікаційна робота включає вступ, три головні розділи, висновки, список літератури та додатки. Загалом, текст роботи нараховує 86 сторінки, в якому представлено 22 рисунків та 1 таблицю. У списку літератури міститься 39 джерел.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися на: «НАУКОВО-ПРАКТИЧНА КОНФЕРЕНЦІЯ "ІНТЕЛЕКТУАЛЬНІ КОМП’ЮТЕРНІ СИСТЕМИ ТА МЕРЕЖІ", яка відбулася 5 грудня2023 року у місті Тернопіль, Україна та Міжнародній мультидисциплінарній інтернет-конференції «Світ наукових досліджень. Випуск 25", яка відбулася 14-15.12.2023, Україна.

1 СУЧАСНИЙ СТАН ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Суть та особливості поняття техногенної катастрофи

Техногенна катастрофа - це подія або серія подій, що виникають внаслідок діяльності людини та характеризуються значними негативними наслідками для навколишнього середовища, здоров'я людей, економічної і соціальної інфраструктури. Це поняття охоплює широкий спектр подій, які можуть включати промислові аварії, ядерні катастрофи, викиди шкідливих речовин, великомасштабні пожежі, технічні збої, вибухи та інші ситуації, спричинені технологічними процесами або збоями в них.

На відміну від природних катастроф, техногенні катастрофи часто виникають через помилки людини, несправність техніки або недоліки у технологічних процесах.

Техногенні катастрофи можуть спричинити різноманітні наслідки, включаючи екологічні, економічні, соціальні та здоров'язнищувальні проблеми.

Часто такі катастрофи виникають раптово та можуть мати складні та непередбачувані наслідки, що ускладнює їх попередження та управління ними.

Негативний вплив техногенних катастроф може тривати довгий час, включаючи тривале забруднення довкілля, здоров'язнищувальні проблеми серед населення, соціальну та економічну дестабілізацію.

Реагування на техногенні катастрофи вимагає спеціальних знань, технологій та ресурсів. Часто залучаються спеціалізовані екстрені служби, інженери, екологи та інші фахівці.

Важливим аспектом управління техногенними ризиками є розробка ефективних стратегій запобігання, готовності та реагування, включаючи

технічні норми безпеки, системи раннього попередження та планування надзвичайних ситуацій.

Враховуючи потенційно катастрофічні наслідки техногенних катастроф, важливо приділяти значну увагу питанням безпеки, моніторингу та плануванню в усіх сферах, де можливе виникнення таких подій.

Типи техногенних катастроф можна класифікувати залежно від їхнього походження, специфіки та наслідків. До основних типів належать (рис.1.1):

1. Промислові аварії: це можуть бути вибухи, пожежі, витікання токсичних речовин або аварії на хімічних заводах, нафтопереробних підприємствах, металургійних комбінатах та інших промислових об'єктах.

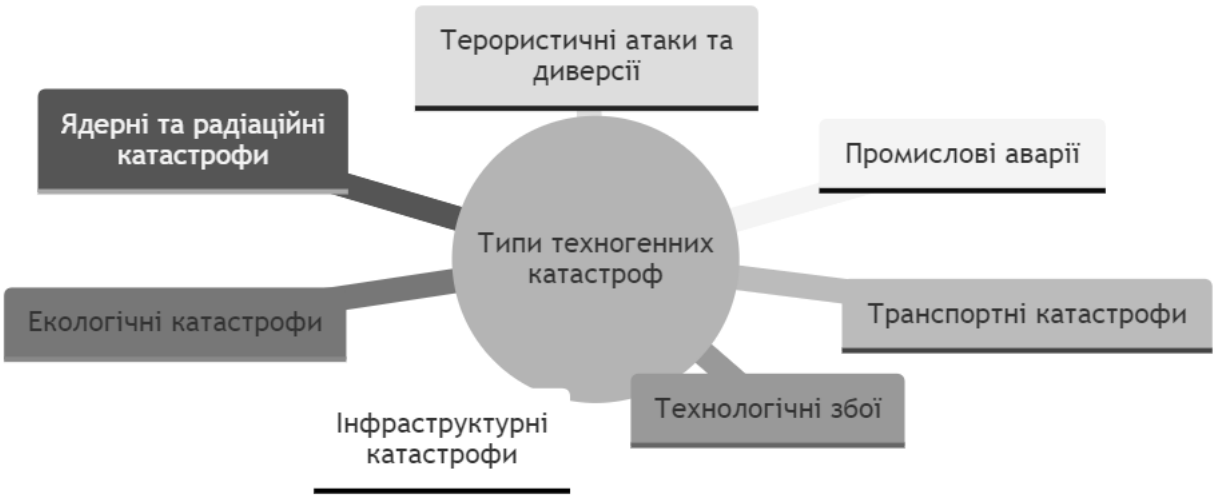


Рисунок 1.1 - Різні типи техногенних катастроф

2. Ядерні та радіаційні катастрофи: це інциденти, що відбуваються на атомних електростанціях, науково-дослідних установах або під час транспортування радіоактивних матеріалів, що призводять до викиду радіації.

3. Транспортні катастрофи: аварії, що відбуваються на різних видів транспорту - авіаційні, залізничні, морські, автомобільні. Вони можуть включати в себе зіткнення, дерайльменти, падіння літаків, потоплення суден тощо.
4. Екологічні катастрофи: неконтрольовані викиди токсичних або шкідливих речовин у навколишнє середовище, що можуть включати нафтові розливи, забруднення водних ресурсів, повітря та ґрунту.
5. Технологічні збої: це можуть бути відмови у системах зв'язку, енергопостачання, комп'ютерних мережах, що призводять до масштабних технічних збоїв і перебоїв у роботі важливих інфраструктурних об'єктів.
6. Терористичні атаки та диверсії: свідоме знищення або пошкодження об'єктів інфраструктури, використання вибухових пристроїв, хімічних або біологічних агентів з метою заподіяння шкоди та створення паніки серед населення.
7. Інфраструктурні катастрофи: обвалення будівель, мостів, дамб, шляхопроводів та інших важливих інженерних споруд.

Кожен із цих типів техногенних катастроф має свої унікальні причини, механізми виникнення та специфічні наслідки, що вимагають відповідних методів запобігання, реагування та ліквідації. Важливо розуміти, що наслідки техногенних катастроф часто перетинаються із соціальними та економічними аспектами, вимагаючи комплексного підходу до управління ризиками та надзвичайними ситуаціями.

Моніторинг після виникнення катастрофи є ключовим аспектом управління наслідками та планування відновлення. Для ефективного моніторингу та оцінки ситуації використовуються різні підходи, які можуть включати (рис.1.2):

1. Супутниковий моніторинг: використання супутникових зображень для оцінки масштабів та впливу катастрофи. Це може включати

аналіз змін у ландшафті, розподіл забруднювачів, стан інфраструктури тощо.

Рисунок 1.2 - Моніторинг після катастрофи та різні підходи до нього

2. Використання безпілотників (дронів): дрони можуть забезпечити швидкий огляд зони катастрофи, збір фото та відеоданих, які допомагають у визначенні областей найбільшого ураження та пріоритетних для допомоги.
3. Встановлення датчиків і сенсорів: використання різних типів датчиків для моніторингу параметрів навколишнього середовища, наприклад, рівнів радіації, якості повітря, забруднення води.

4. Мережі соціальних медіа та мобільних додатків: аналіз даних з соціальних медіа та мобільних додатків може допомогти зібрати інформацію від свідків та постраждалих, що є важливим для оцінки ситуації.
5. Гідрометеорологічний моніторинг: використання гідрометеорологічних даних для оцінки погодних умов, які можуть вплинути на катастрофу, особливо у випадку стихійних лих.
6. Інтегровані інформаційні системи: збір та аналіз даних з різних джерел, включаючи державні бази даних, інформаційні системи екстрених служб для координації реагування на катастрофу.
7. Мобільні бригади оцінки ситуації: відправка команд експертів на місце події для проведення безпосереднього огляду, збору даних та оцінки потреб у відновленні.
8. Залучення громадськості та волонтерів: збір інформації та допомога від громадськості та волонтерських організацій можуть бути важливими у визначенні потреб та розподілі ресурсів.

Кожен з цих підходів відіграє важливу роль у забезпеченні точної та своєчасної інформації, що є критично важливим для ефективного реагування на катастрофи та планування дій по відновленню.

Використання інтелектуальних підходів у контексті управління та реагування на техногенні катастрофи відкриває нові можливості для підвищення ефективності та точності різних процесів, включаючи моніторинг, аналіз, запобігання та відновлення. До основних інтелектуальних підходів належать:

1. Машинне навчання та штучний інтелект (AI): алгоритми машинного навчання можуть бути використані для аналізу великих обсягів даних, отриманих під час та після катастрофи, для ідентифікації закономірностей, прогнозування розвитку подій та визначення оптимальних стратегій реагування.

2. Інтелектуальний аналіз даних: цей підхід включає збір, обробку та аналіз даних з різних джерел, включаючи сенсори, супутникові зображення, соціальні медіа та інші, для отримання цінної інформації про стан катастрофи.
3. Системи підтримки прийняття рішень: розробка систем, заснованих на AI, що допомагають у процесі прийняття рішень, надаючи рекомендації щодо оптимальних дій на основі аналізу поточної ситуації та прогнозів.
4. Прогнозування та моделювання: використання комп'ютерних моделей для симуляції різних сценаріїв розвитку подій та оцінки потенційних наслідків катастроф.
5. Інтелектуальний моніторинг: використання AI для аналізу сенсорних даних, даних з дронів чи супутників для швидкого виявлення та оцінки рівня небезпеки.
6. Інтерактивні візуалізації та інтерфейси: розробка інтуїтивно зрозумілих візуалізацій та інтерфейсів, які дозволяють користувачам легше сприймати та аналізувати складні набори даних.
7. Інтелектуальне управління ресурсами: використання алгоритмів AI для оптимізації розподілу та використання ресурсів під час надзвичайних ситуацій, наприклад, для евакуації, розподілу гуманітарної допомоги та медичних ресурсів.
8. Розумні сенсорні мережі: розгортання сенсорів, здатних до самостійного виявлення змін в навколишньому середовищі та автоматичної передачі даних для аналізу.

Використання цих інтелектуальних підходів дозволяє підвищити швидкість та точність реагування на катастрофи, а також забезпечити більш ефективне планування відновлювальних заходів.

Отже, техногенні катастрофи, викликані людською діяльністю, мають значний вплив на екологію, здоров'я людей, а також на соціальну та

економічну структуру суспільства. Вони охоплюють широкий спектр подій, включаючи промислові аварії, ядерні інциденти, технічні збої, і вимагають особливих знань та ресурсів для реагування та управління. Ці катастрофи мають тенденцію до раптовості та непередбачуваності, з довгостроковими негативними наслідками. Важливість моніторингу після катастрофи полягає у використанні різноманітних інтелектуальних підходів, як-от машинне навчання, AI, інтегровані інформаційні системи, для ефективного аналізу, планування відновлення та мінімізації подальших ризиків. Це включає в себе відстеження через супутник, використання дронів, датчиків, соціальних медіа, а також мобільних бригад для оцінки ситуації. Таким чином, комплексний підхід до управління техногенними ризиками є критично важливим для запобігання та ефективного реагування на подібні катастрофи в майбутньому.

1.2 Огляд існуючих рішень

Існують різні методи виявлення антропогенних аварій і катастроф, такі як візуальне спостереження, використання датчиків і моніторів, аналіз даних зі сторонніх джерел та інші. Однак більшість цих методів обмежені в точності і швидкості реагування, тому виникає необхідність розробки нових і більш ефективних методів визначення рівня антропогенних катастроф. Оцінка стихійних лих на основі даних соціальних мереж є одним із нових напрямків у прогнозуванні ризиків та дослідженнях управління.

Однак такі методи також мають свої обмеження та ризики, у тому числі пов'язані з недостатньою точністю та достовірністю даних соціальних мереж, можливістю поширення неправдивої інформації та іншими. Тому необхідно ретельно аналізувати та перевіряти отриману

інформацію, перш ніж приймати рішення щодо управління ризиками та стихійними лихами.

Одним з підходів є використання інформації з соціальних мереж для управління під час надзвичайних ситуацій [2]. У [3] досліджується застосування штучного інтелекту, WebGIS та кількох алгоритмів для покращення реагування на надзвичайні ситуації та допомоги в ліквідації наслідків стихійних лих під час різних стихійних лих та кризових ситуацій. Згідно з [4], розроблена розподілена багаторівнева система оповіщення про надзвичайні ситуації з використанням датчиків детекторів подій, що підтримує додатки для розумних міст. Ця система може виявляти та доставляти аварійні тривоги в режимі реального часу на основі виявлених подій, заздалегідь відомих рівнів ризику конкретних областей та тимчасової інформації.

У [5] був запропонований прототип системи ефективного використання твітів для підтримки завдань мобілізації ресурсів під час надзвичайних ситуацій. Конвеєр прийняття рішень запропонованої системи включає класифікацію твітів з подальшим ранжуванням твітів для визначення пріоритету передачі ресурсів.

У [6] досліджується застосування штучного інтелекту та науки про дані для забезпечення розумної стійкості до надзвичайних ситуацій, криз та катастроф. У [7], Support Vector Machine (SVM), K-Nearest Neighbors (KNN) та логістичній регресії, алгоритми аналізу даних були використані та порівняні для класифікації твітів про стихійні лиха. У [8] проведено аналіз великих даних та їх застосувань у сфері розумної нерухомості та ліквідації наслідків стихійних лих, визначено ключові характеристики та потенційні способи використання для покращення обох сфер. Робота [9] присвячена виявленню ситуативних твітів під час катастрофи. Нейронний підхід, подібний [10], був розроблений на основі комбінації моделі RoBERTa і методу ідентифікації ситуативних твітів під час катастрофи,

заснованого на функціях. Однак ефективність запропонованих підходів залежить від кількості міток у соціальних мережах, пов'язаних з надзвичайними ситуаціями, що впливає на своєчасне реагування на аварію.

У [11] аналізуються останні розробки в галузі машинного навчання для управління стихійними лихами, зосереджуючись на таких методах і застосуваннях, як прогнозування, оцінка ризиків, виявлення небезпек, системи раннього попередження, моніторинг, оцінка збитків і реагування після стихійних лих, а також рекомендації щодо подальших досліджень. Civique [12] розглядає ще одну систему виявлення надзвичайних ситуацій, яка в першу чергу націлена на твіти з геотегами та використовує Support Vector Machines (SVM) та Naive Bayes для навчання.

В [13] була запропонована глибока об'єктно-орієнтована структура виявлення змін під назвою ChangeOS. У [14] була представлена модель визначення часу реагування аварійно-рятувальних служб. У [15] була запропонована концептуально-архітектурна основа розробки систем виявлення надзвичайних ситуацій на основі парадигми «людина як датчик» (HaaS).

У [16] розроблена гібридна модель TextCNN-Hatt на основі ALBERT, яка використовує тематичні знання для аналізу суспільних настроїв під час надзвичайних ситуацій, демонструючи високу ефективність зі значними покращеннями у виявленні настроїв порівняно з попередніми моделями, і досягаючи середньої точності 89%, AUC 94% та CQI 88%. Ця модель зберігає практичний потенціал для виконання завдань з реагування на надзвичайні ситуації та допомагає особам, які приймають рішення, належним чином реагувати на кризові ситуації. В [17] була продемонстрована гібридна структура, що включає попередньо навчену модель DistilBERT і запропонований алгоритм вибору ознак для виявлення надзвичайних ситуацій. Експерименти та порівняння запропонованої

системи показують її перевагу з точки зору точності ідентифікації подій та зменшення ознак порівняно з іншими.

У таблиці 1.1 наводиться порівняльна оцінка основних аналогів виявлення надзвичайних ситуацій за допомогою аналізу даних із соціальних мереж.

Таблиця 1.1 - Порівняльна оцінка методів виявлення надзвичайних ситуацій.

Target	**Метод**	**Точність**
Bhoi et al. [5]	Гібридна модель, що включає LSTM і CNN.	Показники F1 за обома наборами даних становлять 84% та 84%
Гопнараян А., Дешпанде С. [7]	SVM, KNN і логістична регресія.	Не вказано
С. Мадічетті і С. М., [9]	Поєднання моделі RoBERTa та методу, заснованого на функціях.	Точність становить 90%
Д. Каноджіа, В. Кумар, К. [12]	Класифікація повідомлень у реальному часі.	F-міра перевищує 70% і 90% відповідно.
Аввенуті, М. та ін [15]	Не уточнюється.	Точність = 75%, пригадування = 100%, F-міра = 86%
Чжан і Ма, [16]	Гібридна модель TextCNN-Hatt на основі Альберта, доповнена знаннями про тему для аналізу настроїв раптових катастроф	89%
Adel, H. et al.l [17]	Модель DistilBERT з пошуковим алгоритмом Hunger Games.	98% (набір даних C6), 97% (набір даних C36)

Незважаючи на наявність різноманітних інструментів для аналізу антропогенних катастроф та забезпечення автоматизованого реагування [16,17,18], прийняття ефективних рішень для запобігання потенційним гуманітарно-економічним катастрофам залишається проблемою. Більш того, в існуючих роботах недостатня увага приділяється оцінці наслідків

цих катастроф і визначенню рівня небезпеки. Враховуючи ці недоліки, метою статті є розробка ефективної методики оцінки рівня антропогенних катастроф на основі інформації від свідків події. Для досягнення цієї мети ми сформували два ключові завдання:

Розробка інтелектуального методу класифікації для оцінки рівня антропогенних катастроф на основі підсилювальних ансамблевих методів машинного навчання.

В контексті існуючих рішень для виявлення антропогенних аварій і катастроф, важливо визнати, що традиційні методи, такі як візуальне спостереження та використання датчиків, часто обмежені у точності і швидкості. Новітні підходи, які включають використання даних з соціальних мереж та застосування штучного інтелекту, пропонують значні переваги у прогнозуванні та управлінні ризиками. Однак, ці методи також супроводжуються власними викликами, зокрема, щодо точності та достовірності даних. Попередні дослідження, зокрема використання розподілених багаторівневих систем оповіщення та аналізу твітів у надзвичайних ситуаціях, підкреслюють потенціал цих технологій у виявленні та ефективному реагуванні на кризові ситуації. З огляду на це, розробка нової методики оцінки рівня антропогенних катастроф, яка використовує підсилювальні ансамблеві методи машинного навчання, та створення відповідного мобільного додатку для класифікації цих катастроф у регіоні можуть стати важливим кроком у розвитку цієї сфери, пропонуючи нові можливості для зміцнення безпеки та зменшення ризиків, пов'язаних з антропогенними катастрофами.

1.3 Постановка задачі

Актуальність даного дослідження полягає в необхідності розробки ефективних методів для ідентифікації, класифікації та реагування на

техногенні катастрофи, що є особливо важливим у контексті зростаючого впливу технологій та промислового розвитку. Різноманітність і складність таких подій, що охоплюють від промислових аварій до екологічних криз, ставить перед науковою спільнотою завдання розробки інтегрованих підходів для їх аналізу та управління. Особливу увагу в дослідженні приділено використанню інноваційних технологій, таких як машинне навчання та штучний інтелект, які відіграють вирішальну роль у точному та оперативному виявленні потенційних ризиків.

Друга актуальність дослідження полягає у вивченні та застосуванні методів обробки природної мови (NLP) та аналітичних інструментів для аналізу та обробки великих обсягів текстових даних. Використання таких методів, як видалення стоп-слів, лематизація, нормалізація та токенізація, дозволяє більш ефективно обробляти інформацію, що отримана від різних джерел, включаючи звіти про аварії, записи очевидців та медіа-звіти. Вдосконалення цих методів сприятиме більш точному розумінню та класифікації інцидентів, що є ключовим для розробки відповідних стратегій реагування.

Третій аспект актуальності цього дослідження відображається у впровадженні сучасних технік машинного навчання для класифікації і прогнозування рівня серйозності техногенних катастроф. Аналіз та застосування бустингових алгоритмів, таких як AdaBoost, Gradient Boost, XGBoost, CatBoost та LightGBM, є важливим для розуміння їх ефективності в різних сценаріях. Такі методи можуть значно підвищити точність прогнозування та виявлення потенційних катастроф, дозволяючи вчасно реагувати та мінімізувати можливі негативні наслідки.

Наостанок, дослідження має важливе значення для розробки інтегрованих систем управління ризиками. Підходи, розроблені в рамках цього дослідження, можуть бути використані для покращення існуючих систем управління надзвичайними ситуаціями, включаючи системи

раннього попередження, планування евакуації та відновлення після катастроф. Впровадження цих новітніх методів та технологій сприятиме більш ефективному розподілу ресурсів, зниженню ризиків та покращенню загальної готовності до техногенних надзвичайних ситуацій.

Метою даного дослідження є розробка та впровадження інтелектуального методу класифікації наслідків техногенних катастроф з метою покращення управління кризовими ситуаціями та запобігання подібним подіям у майбутньому.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Розглянути сутність та особливості поняття "техногенна катастрофа" з метою уявлення про їхні характеристики та відмінності від природних катастроф.
2. Здійснити огляд існуючих рішень та методів управління техногенними катастрофами з метою визначення їхніх переваг та обмежень.
3. Провести детальний опис методів обробки текстових даних, зокрема включаючи такі аспекти як видалення стоп-слів, токенізація, лематизація, нормалізація, визначення та візуалізація N-грам та інші.
4. Провести огляд бустингових методів, таких як AdaBoost, Gradient Boost, XGBoost, та визначити їхню ефективність у вирішенні завдань класифікації.
5. Розробити інтелектуальний метод класифікації наслідків техногенних катастроф, який базується на обробці текстових даних та використанні бустингових алгоритмів.
6. Провести попередній аналіз даних, які будуть використовуватися в дослідженні, з метою зрозуміння їхнього обсягу та особливостей.
7. Виконати попередній аналіз текстових даних, які будуть використовуватися для класифікації, з метою визначення ключових особливостей текстів.

8. Реалізувати метод класифікації наслідків техногенних катастроф в Україні з використанням розроблених алгоритмів та методів для аналізу та класифікації текстових даних.

Ці задачі спрямовані на досягнення основної мети дослідження - розробку і впровадження інтелектуального методу класифікації наслідків техногенних катастроф з метою покращення управління кризовими ситуаціями та запобігання подібним подіям у майбутньому.

Завдання 1-2 розглянуто в параграфах 1.1, 1.2 та 1.3 відповідно, а виконання завдань 4-8 представлені у параграфах 2.1-3.3.

Висновки до розділу 1

Техногенні катастрофи, які є наслідком людської діяльності, представляють собою складні випадки з серйозними негативними наслідками для навколишнього середовища, здоров'я людей та інфраструктури. Особливістю таких катастроф є їхня виникнення через помилки людини, технічні збої та недоліки в технологіях, що робить їх відрізняючимися від природних катастроф та ставить вимогу до спеціалізованого підходу в управлінні та запобіганні. Різноманітність типів техногенних катастроф, що включають промислові аварії, ядерні надзвичайні ситуації, транспортні катастрофи, екологічні лиха, технічні збої, терористичні акти та інфраструктурні збої, вимагають інтегрованого підходу до ризик-менеджменту. Моніторинг після виникнення катастрофи, що включає використання сучасних технологій, таких як супутниковий моніторинг, безпілотники, сенсори, аналіз соціальних медіа та мобільних додатків, інтегровані інформаційні системи, мобільні бригади оцінки, залучення громадськості та волонтерів, є ключовим для ефективного реагування на такі події. Використання інтелектуальних підходів, включаючи машинне навчання, штучний інтелект, інтелектуальний аналіз даних, системи підтримки прийняття рішень, прогнозування та моделювання, інтелектуальний моніторинг, інтерактивні візуалізації, інтелектуальне управління ресурсами та розумні сенсорні мережі, може значно підвищити ефективність цих процесів, забезпечуючи швидкість та точність у відповіді на надзвичайні ситуації, а також у плануванні відновлювальних дій.

Вивчення існуючих рішень для виявлення антропогенних аварій і катастроф показує, що хоча існують численні методи, як традиційні, так і сучасні, багато з них мають обмеження в точності та швидкості реагування. Це створює потребу в розробці новітніх та більш ефективних технік.

Останнім часом з'являється значний інтерес до використання даних з соціальних мереж для визначення рівня ризику та керування надзвичайними ситуаціями. Незважаючи на потенційні переваги цього підходу, він також має свої виклики, зокрема, що стосується точності та надійності даних. Порівняльний аналіз існуючих методів показує, що сучасні підходи, такі як використання штучного інтелекту та аналіз великих даних, можуть забезпечити більш точну оцінку та швидке реагування. Однак, незважаючи на технологічні досягнення, важливо розуміти, що управління ризиками антропогенних катастроф залишається складним завданням, що вимагає комплексного підходу та постійного вдосконалення інструментів та методів.

2 ІНТЕЛЕКТУАЛЬНИЙ МЕТОД КЛАСИФІКАЦІЇ НАСЛІДКІВ ТЕХНОГЕННИХ КАТАСТРОФ

2.1 Опис методів обробки тексту

У сфері сучасного аналізу тексту, техніки обробки природної мови (NLP), такі як вилучення стоп-слів, лематизація, нормалізація, токенізація, виявлення N-грам та візуалізація, а також тегування слів і використання методів, як-от SMOTE, мають вирішальне значення (рис.2.1). Вони полегшують перетворення великих обсягів тексту в формат, зручний для аналізу, забезпечуючи глибше розуміння контексту і структури мови. Застосування цих методів підвищує точність і ефективність при вирішенні завдань NLP. Також, використання технік, як SMOTE, сприяє кращому аналізу і класифікації маловивчених шаблонів у даних, збільшуючи ефективність моделей машинного навчання.

У процесі аналізу тексту, особливо в області обробки природної мови (NLP), стоп-слова відіграють значну роль. Вони являють собою слова, які зустрічаються часто, але не несуть значущої семантики для розуміння контексту. Прикладами таких слів можуть бути "і", "в", "на". Видалення цих слів є важливим етапом обробки тексту, оскільки воно дозволяє зосередитися на більш інформативних елементах тексту.

Процес видалення стоп-слів спрямований на поліпшення якості текстового аналізу. Він позбавляє текст від надлишкових і менш важливих слів, які не роблять істотного внеску в загальне розуміння тексту. Видалення цих слів допомагає в зменшенні обсягу даних для обробки та може покращити ефективність алгоритмів NLP.

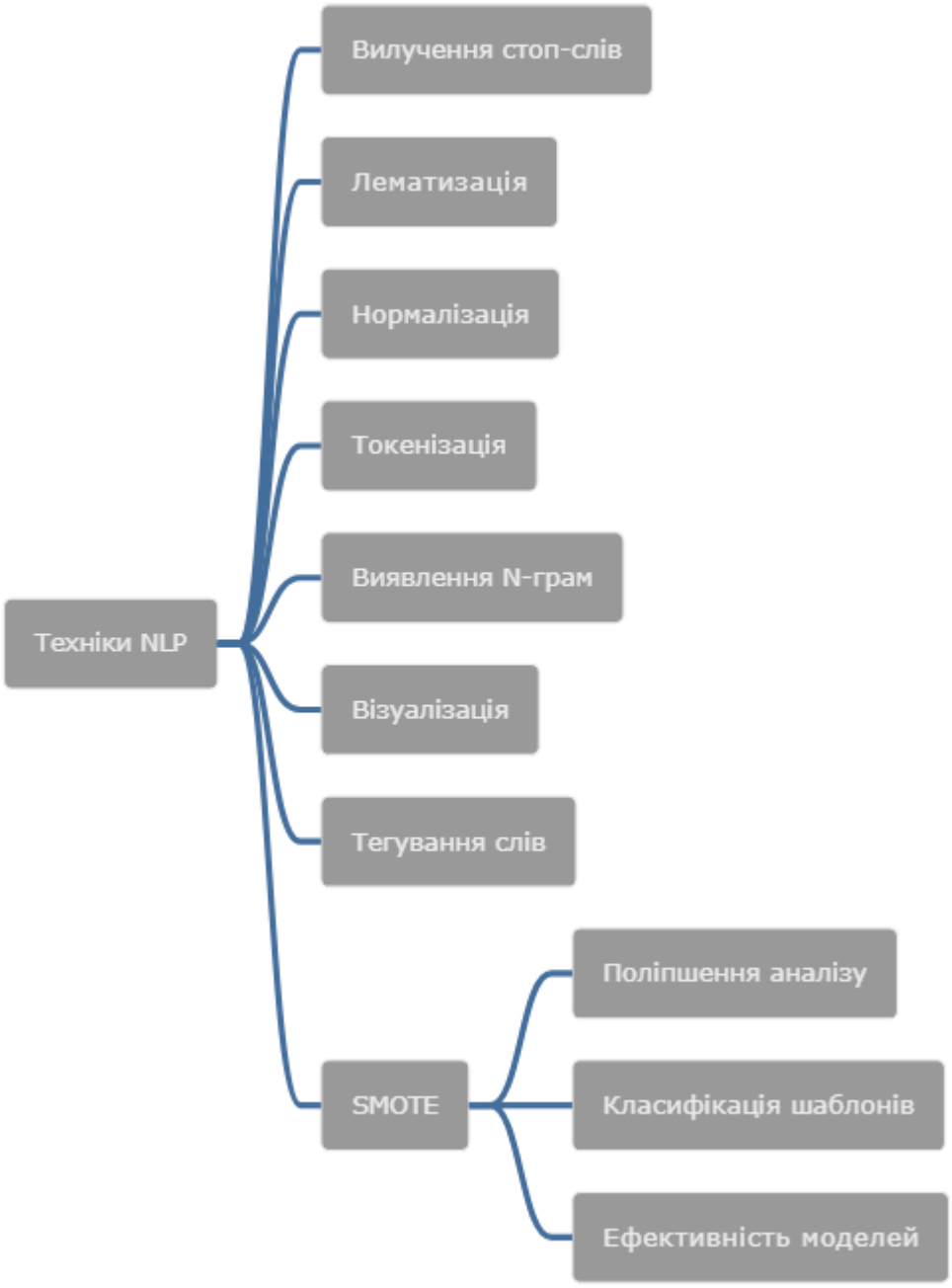


Рисунок 2.1 - Ключові техніки обробки природної мови (NLP)

Алгоритм видалення стоп-слів можна описати наступним чином:

1. Формування множин:
    - Вихідний текст: Позначаємо вихідний текст як D, який містить слова $w1,w2,\ldots,wN$.

- Множина стоп-слів: Створюємо множину стоп-слів S, яку ми плануємо видалити.

2. Операція видалення стоп-слів. Використовуючи операцію віднімання множин, позначену як \, видалимо стоп-слова з вихідного тексту:

$$D' = D \setminus S$$

3. Оновлений текст. Результатом операції буде новий текст D', який не містить стоп-слів:

$$D' = \{wi \mid wi \in D, wi \in /S\}$$

Цей підхід дозволяє відфільтрувати з тексту менш значущі слова, залишаючи лише ті, які вносять важливу інформацію для аналізу та обробки тексту.

Лематизація - це процес зведення слова до його базової форми або леми. Лема - це слово, яке представляє слово в його словничному вигляді, відокремленому від граматичних форм.

Різні методи лематизації існують у залежності від потреби та контексту:

1. WordNet Lemmatizer: Використовує базу даних WordNet, яка містить слова та їх семантичні відносини. Цей метод визначає лему слова, опираючись на синоніми та відносини між словами.
2. Porter's Algorithm: Цей алгоритм використовує набір евристик та правил для відсіювання суфіксів. Він відомий своєю ефективністю та швидкістю, але може видаляти не тільки суфікси, а й значущі частини слова.

3. Lancaster's Algorithm: Схожий на алгоритм Портера, але відрізняється більш агресивним підходом до видалення суфіксів, що може призвести до більш сильного скорочення слів.
4. SpaCy Lemmatization: Використовує статистичні моделі та контекст для визначення леми. Цей метод здатен враховувати граматичні особливості та синтаксичні відносини, що робить його більш точним у контекстно-залежних сценаріях.

Математично, процес лематизації можна представити як функцію lemmatize(w), де w - слово, а lemmalemma - його лема. Наприклад,

lemmatize("running")="run"lemmatize("running")="run"

lemmatize("better")="good"lemmatize("better")="good".

Лематизація є ключовим кроком у багатьох задачах NLP, оскільки вона дозволяє обробляти тексти на більш глибокому рівні, аналізуючи семантичні та граматичні відносини між словами.

Нормалізація тексту - це процес перетворення тексту в такий вигляд, який спрощує його обробку та порівняння. У контексті обробки природної мови, нормалізація включає в себе різні операції, такі як приведення до нижнього регістру, видалення зайвих символів, лематизація та інші.

Нормалізація тексту є ключовим етапом у процесі обробки природної мови, що спрощує аналіз та обробку текстових даних. Цей процес полягає в перетворенні тексту до стандартного, зручного для обробки формату. Математично, процес нормалізації можна представити як функцію T'=normalize(T), де T - вихідний текст, а T' - результат нормалізації.

Основні етапи нормалізації включають:

1. Приведення до Нижнього Регістру: Цей крок включає перетворення всіх символів тексту в нижній регістр. Це спрощує подальший аналіз, уніфікуючи випадки використання різних регістрів. Математично це можна представити як T'=lowercase(T).

2. Видалення Зайвих Символів: На цьому етапі видаляються всі неважливі символи, такі як пунктуація та спеціальні символи, які не несуть значущої інформації для аналізу тексту.

$$T' = \backslash text\{remove_{symbols}\}(T).$$

3. Токенізація: Процес розбивання тексту на окремі токени (слова чи фрази). Це дозволяє аналізувати окремі елементи тексту.

$$T' = tokenize(T).$$

4. Лематизація: Зведення слів до їхніх базових, словникових форм. Це допомагає у зведенні різних форм слова до одного стандарту.

$$T' = lemmatize(T).$$

5. Видалення Стоп-Слів: Стоп-слова - це загальновживані слова, які часто використовуються, але не несуть значущої інформації для аналізу. Їх видалення допомагає зосередитися на більш важливих словах у тексті.

$$T' = \backslash text\{remove_{stopwords}\}(T).$$

Залежно від конкретних потреб можуть застосовуватися й інші операції, такі як видалення числових значень, коректування абревіатур тощо. В цілому, нормалізація тексту включає послідовність операцій, яка виглядає наступним чином:

$$T' = remove_{stopwords}\Big(lemmatize\Big(tokenize\Big(remove_{symbols(lowercase(T))}\Big)\Big)\Big)$$

Ця послідовність операцій представляє собою стандартний процес нормалізації тексту, спрощуючи його для подальшого аналізу та обробки.

Токенізація представляє собою фундаментальний процес у сфері обробки природної мови (NLP), що полягає в розбивці текстових даних на менші фрагменти, звані токенами. Такі токени можуть бути ідентифіковані як окремі слова, фрази, речення, або навіть більш гранулярні одиниці, в залежності від потреби специфічної задачі аналізу тексту.

Мета токенізації полягає в спрощенні аналізу тексту шляхом його розділення на індивідуальні елементи, що допомагає у вивченні структури та семантики на детальному рівні, та створює основу для виконання таких важливих операцій, як лематизація та видалення стоп-слів, які є ключовими для ефективного вирішення завдань обробки природної мови (NLP).

Типи Токенізації:

- Токенізація слів: Це процес розділення тексту на окремі слова. Математично це можна виразити як W=tokenize_words(T), де W - множина токенів (слів) з вихідного тексту T. Наприклад, для речення "Tokenization is an important NLP task." множина слів буде {"Tokenization", "is", "an", "important", "NLP", "task"}.
- Токенізація речень: Полягає у розділенні тексту на індивідуальні речення. Вираз може бути представлений як R=tokenize_sentences(T). Наприклад, для тексту "Tokenization is an important NLP task. It helps in text analysis." результатом буде два речення: {"Tokenization is an important NLP task.", "It helps in text analysis."}.
- Токенізація фраз: Здійснюється розділення тексту на фрази або окремі фрагменти. Це можна математично визначити як P=tokenize_phrases(T). Наприклад, для тексту "Natural language

processing is a field of study in AI." результатом буде: {"Natural language processing", "field of study", "in AI"}.

Токенізація є важливим етапом в обробці природної мови, який дозволяє розбити текст на окремі одиниці для подальшого аналізу та обробки. Результатом токенізації може бути множина слів, речень чи інших фрагментів, залежно від вимог задачі.

N-грами є фундаментальними елементами в аналізі тексту, що представляють собою послідовності з N підряд розташованих елементів (слів, символів або інших одиниць) в тексті. Вони використовуються для аналізу тексту на рівні послідовностей, дозволяючи виявляти семантичні та структурні зв'язки між словами чи символами. Математично N-грами можна визначити як набір $\{(t1,t2,\ldots,tN)|ti\in T,1\leq i\leq|T|-N+1\}$, де $|T|$ є довжиною тексту, а T - сам текст.

Візуалізація N-грам є ключовою для розуміння структури та взаємозв'язків у тексті. Існує кілька методів візуалізації:

1. Графік частот: Цей метод дозволяє визначити, які N-грами використовуються найчастіше, шляхом створення графіка, що відображає частоту їх вживання в тексті.
2. Хмара слів: Хмара слів може бути використана для представлення найбільш часто вживаних N-грам, роблячи видимими ключові елементи тексту через візуальний акцент на найчастіше вживані слова або фрази.
3. Таблиці частот: Таблична форма даних дозволяє детально аналізувати частоту використання різних N-грам, надаючи користувачам можливість оцінювати їх вживаність і релевантність у контексті досліджуваного тексту.

Загалом, візуалізація N-грам є невід'ємною частиною аналітичного процесу в аналізі тексту, оскільки вона сприяє глибшому розумінню

текстових зв'язків та структур, важливих для різноманітних застосувань у сфері обробки природної мови та лінгвістичних досліджень.

Тегування слів у сфері мов програмування та обробки природної мови (NLP) є критичним процесом, який полягає у присвоєнні кожному слову в тексті відповідного лінгвістичного тегу. Ці теги визначають частину мови слова, а також його граматичні характеристики, такі як рід, відмінок, час, число тощо. Тегування словесних форм є фундаментальним для розуміння граматичної структури тексту і використовується у широкому спектрі NLP-задач, включаючи аналіз тексту, машинний переклад, створення синтаксичних дерев та інше.

Математично, процес тегування можна описати як функцію, що відображає множину слів у тексті на множину можливих тегів. Якщо W - множина слів у тексті T, тоді функція тегування може бути виражена як tags: tags:$W \rightarrow T$, де T - множина тегів.

Процес тегування складається з наступних етапів:

1. Токенізація: Розділення тексту на індивідуальні токени (слова чи фрази).
2. Частиномовне тегування: Кожному токену присвоюється відповідний тег частини мови, такий як дієслово (Verb), іменник (Noun), прикметник (Adjective) тощо.
3. Граматичне тегування: Визначення додаткових граматичних характеристик слова, наприклад, роду, відмінку, часу, числа.
4. Лематизація: Приведення слова до його базової форми, або леми.

Наприклад, при математичному тегуванні слова w, функція tags(w) може повернути набір тегів, наприклад {"*Noun*","*Plural*","*Masculine*"}, що означає, що слово w є іменником, стоїть у множині і має чоловічий рід.

Тегування слів є важливим компонентом у процесах NLP, оскільки воно дозволяє здійснювати точніший аналіз тексту, виходячи із граматичної структури та контексту вживання слів.

SMOTE (Synthetic Minority Over-sampling Technique) використовує математичні концепції для генерації синтетичних даних. Ось більш детальний математичний опис процесу:

Вибір Зразків та Їх Найближчих Сусідів: Для кожного зразка $\mathbf{x_i}$ з міноритарного класу знайдіть його k найближчих сусідів. Найближчі сусіди зазвичай визначаються за допомогою Евклідової відстані, хоча можуть використовуватися й інші метрики. Евклідова відстань між двома точками $\mathbf{x_i}$ та $\mathbf{x_j}$ в n-вимірному просторі обчислюється як:

$$\mathbf{d(x_i, x_j)} = \sqrt{\sum_{\mathbf{d=1}}^{\mathbf{n}} \left(\mathbf{x_{i,d} - x_{j,d}}\right)^2}$$

де $\mathbf{x_{i,d}}$ та $\mathbf{x_{j,d}}$ — це значення d-го виміру відповідно для $\mathbf{x_i}$ та $\mathbf{x_j}$.

Генерація Синтетичних Зразків: Для кожного вибраного зразка $\mathbf{x_i}$ та одного з його найближчих сусідів $\mathbf{x_{nn}}$, генеруються нові синтетичні зразки. Це робиться за допомогою лінійної інтерполяції між $\mathbf{x_i}$ та $\mathbf{x_{nn}}$. Новий зразок $\mathbf{x_{new}}$ створюється так:

$$\mathbf{x_{new} = x_i + \lambda * (x_{nn} - x_i)}$$

де λ — це випадкове число в діапазоні [0, 1]. Це означає, що кожен новий зразок є випадковою точкою на відрізку, що з'єднує $\mathbf{x_i}$ та $\mathbf{x_{nn}}$.

Повторення Процесу: Цей процес повторюється для необхідної кількості зразків, щоб досягти бажаного рівня балансу між класами.

Важливо зазначити, що хоча SMOTE допомагає в боротьбі з незбалансованістю класів, він може також призвести до перенавчання на синтетичних даних. Крім того, цей метод може бути неефективним у

випадках, коли міноритарний клас має дуже складну або нерегулярну структуру.

Сучасні методи обробки тексту, такі як видалення стоп-слів, лематизація, нормалізація, токенізація, визначення та візуалізація N-грам, а також тегування слів, відіграють важливу роль у галузі обробки природної мови (NLP). Вони спрощують аналіз та обробку текстів, дозволяючи виокремлювати інформативні елементи, аналізувати граматичні та семантичні структури, а також визначати ключові терміни та зв'язки між ними. SMOTE, як метод балансування даних, забезпечує більш ефективне машинне навчання у сценаріях з незбалансованими класами. Ці методики разом формують потужний інструментарій для розробки алгоритмів NLP, що здатні точно та ефективно обробляти великі обсяги текстових даних, сприяючи глибокому розумінню мовних патернів та поліпшенню прийняття рішень на основі текстової інформації.

2.2 Опис бустингових методів

Метод бустингу використовується в класифікації для підвищення точності моделі. Його суть полягає в об'єднанні слабких класифікаторів в сильний. На відміну від інших методів, наприклад, Bagging [19], Boosting є навчальними моделями послідовно, де кожна наступна модель виправляє помилки попередньої. Ця властивість відіграє найважливішу роль у складних завданнях з різноманітними даними, як у випадку з класифікацією рівня антропогенних катастроф. Текстові та кількісні дані, що надходять від свідків подій, можуть мати високу мінливість і неоднорідність, що робить Boosting вибором номер один. Розглянемо докладніше існуючі методи Boosting.

AdaBoost. AdaBoost розшифровується як «Adaptive Boosting» або адаптивне підсилення. Він перетворює слабкі алгоритми навчання на

сильні рішення проблем класифікації. Остаточне рівняння (1) для класифікації можна представити таким чином:

$$F(x) = \sum_{t=1}^{T} f_t(x), \qquad 1)$$

де кожен є слабким учнем, який приймає об'єкт $f_t x$ як вхідні дані та повертає значення, що вказує на клас об'єкта. Наприклад, у двокласній задачі ознака слабкого балу учня визначає передбачуваний клас об'єкта, а абсолютна величина дає впевненість у цій класифікації. Аналогічно і з T, якщо цей класифікатор позитивний, вибірка знаходиться в позитивному класі, в іншому випадку вона негативна.

Кожен слабкий учень генерує початкову гіпотезу, , для кожної вибірки навчальної вибірки. На кожній ітерації вибирається слабкий учень і йому присвоюється коефіцієнт таким чином, що сума похибки навчання (2) результуючого t-класифікатора покращення сцени зводиться до мінімуму.$h(x_i) t \alpha_t E_t$

$$E_t = \sum_i E\left[F_{(t-1)}(x_i) + \alpha_t h(x_i)\right], \qquad 2)$$

де - розширений класифікатор, який був побудований до попереднього етапу навчання, - деяка функція помилки, а слабкий учень - слабкий учень, який розглядається для додавання до остаточного класифікатора.$F_{(t-1)}(x_i)\ E(F) f_t(x) = \alpha_t h(x)$

На кожній ітерації тренувального процесу береться вага і присвоюється кожній вибірці в навчальній множині, рівна поточній похибці на цій вибірці. Ці ваги можуть бути використані для інформування про тренування слабкого учня, наприклад, можна виростити дерева рішень,

які допоможуть розділити вибіркові набори з великими вагами.$\omega_{(i,t)}E\left(F_{(t-1)}(x_i)\right)$

Підсилення градієнта. У градієнтних машинах (Gradient Boost Machines, GBM) процедура навчання послідовно адаптує нові моделі, щоб забезпечити більш точну оцінку змінної відгуку [20]. Основна ідея цього алгоритму полягає в побудові нових базових тренувальних елементів, які максимально корелюють з негативним градієнтом функції втрат, пов'язаної з усім ансамблем. Використовувана функція втрат може бути довільною, але якщо функція помилки є класичною квадратичною похибкою, то процедура навчання призведе до послідовної апроксимації помилок. У загальному випадку вибір функції втрат залежить від дослідника і конкретного завдання. Завдяки такій високій гнучкості GBM можна легко налаштувати для будь-якого конкретного завдання, пов'язаного з даними.

Контрольоване навчання накладає на дослідника сильні обмеження, оскільки дані повинні бути забезпечені достатнім набором відповідних цільових міток (вилучення яких може бути дуже дорогим, наприклад, в результаті дорогого експерименту). Дано набір даних , де з кількістю ітерацій M відноситься до пояснювальних вхідних змінних, а y до відповідних міток змінної відповіді. Мета полягає в тому, щоб реконструювати невідому функціональну (3) залежність за допомогою нашої оцінки так, щоб деяка визначена функція втрат була мінімізована$(x,y)Ni\ =\ 1x\ =\ (x1,\dots,xd)\left(\left(x\overset{f}{\rightarrow}y\right)\right)f(\widehat{x}),\Psi(y,f)$:

$$f(\widehat{x}) = y,$$
$$f(\widehat{x}) = argmin_{f}\,(x)\,\psi\left(y,f(x)\right). \qquad 3)$$

Таким чином, ініціалізується константою for to M і обчислюється від'ємний градієнт відповідно до нової базової функції навчання . Виходячи з цього, потім знаходять найкращий градієнт спуску розміру сходинки$\hat{f}_0 t\ =\ 1\ g_t(x)h(x,\theta_t)\rho_t$:

$$\rho_t = argmin_\rho \sum_{i=1}^{N} \Psi\left[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i,\theta_t)\right] \qquad 4)$$

Останнім кроком є оновлення оцінки функцій (5):

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x,\theta_t) \qquad 5)$$

XGBoost. XGBoost — це алгоритм машинного навчання, заснований на дереві рішень, який використовує фреймворк підвищення градієнта. У XGBoost побудова дерева заснована на розпаралеленні. Це можливо завдяки взаємозамінності петель, що використовуються для побудови тренувальної основи: зовнішня петля перераховує листя дерева, внутрішня петля обчислює особливості. Цикл, знайдений всередині іншого циклу, перешкоджає розпаралелюванню алгоритму, оскільки зовнішній цикл не може розпочати своє виконання, поки внутрішній цикл ще не завершив свою роботу. Тому для поліпшення часу виконання змінюється порядок циклів: при зчитуванні даних відбувається ініціалізація, потім сортування виконується за допомогою паралельних потоків. Ця зміна покращує продуктивність алгоритму, розподіляючи обчислення між потоками.

Тому існує тренінг , де диференційована функція втрат, визначається тренувальний ряд і є швидкістю навчання. Ініціалізація моделі (6) має константне значення$set(x,y)Ni\ =\ 1\alpha$:

$$f(\widehat{x}) = argmin_{\theta} \sum_{i=1}^{N} L(y_i, \theta). \quad 6)$$

«Градієнти» (7) і «гессенці» (8) визначаються для M$m = 1$:

$$\hat{g}_m(x_i) = \left[\frac{\partial^2 L\left(y_i, f(x_i)\right)}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \quad 7)$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L\left(y_i, f(x_i)\right)}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \quad 8)$$

Далі вибирається базовий учень (або слабкий) (9) за допомогою навчальної вибірки

$$\left\{x_i - \frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N} \quad 9)$$

розв'язавши задачу оптимізації (10, 11):

$$\hat{\phi}_m = argmin_{\phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2 \quad (10$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x) \quad (11$$

На останньому кроці (12) потрібно оновити оцінку функції: (12)

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \hat{f}_m(x)$$

CatBoost. Бібліотека CatBoost [21] є ще однією ефективною реалізацією градієнтного підсилення над деревами рішень. CatBoost – це

реалізація градієнтного підсилення, що дозволяє використовувати бінарні дерева рішень як базові предиктори. Припустимо, що існує набір даних з вибірками , де - вектор n ознак, а - характеристика відгуку , яка може бути двійковою (тобто так чи ні) або закодованою як числова ознака (0 або 1). Зразки розподіляються незалежно та однаково відповідно до деякого розподілу p($D = \{(X_j, y_j)\}_{j=1,\dots,m} X_j = (x^1{}_j, x_j^2, \dots, x_j^n) y_j \in R(X_j, y_j) \cdot,\cdot$). Мета тренувального завдання полягає в тому, щоб навчити (13) функцію, яка мінімізує очікувані втрати, наведені в$H: R^n \to R$

$$L(H) \coloneqq EL(y, H(X)) \qquad 13)$$

де - функція плавних втрат і - тестові дані, вибрані з тренувальних даних . Процедура підсилення градієнта ітеративно будує послідовність наближень З попереднього наближення виходить адитивний процес, такий, що , з розміром кроку α і функцією, яка є базовим предиктором, вибирається з набору функцій G для зменшення або мінімізації очікуваних втрат (14), визначених як$L(\cdot,\cdot)(X, y) DHt: R\, m \to R, t = 0,1, \dots H^{t-1}, H^t H^t = H^{t-1} + \alpha g^t g^t: R^n \to R,$:

$$g^t = \arg\min_{g \in G} L(H^{t-1} + g) = \arg\min_{g \in G} EL(y, H^{t-1}(X) + g(X)) \qquad 14)$$

Часто задача мінімізації вирішується методом Ньютона за допомогою наближення другого порядку на або шляхом (від'ємного) градієнтного кроку. Будь-яка з цих функцій є градієнтним спуском.$L(H^{t-1} + g^t) H^{t-1}$

Класифікатор LightGBM. Light Gradient Boosted Machine, або скорочено LightGBM [22,23], є реалізацією підсилення градієнта з відкритим вихідним кодом. Gradient Boosted Decision Tree (GBDT) — популярний алгоритм класифікатора. Таким чином, надається навчальна

множина, яка складається з пар , де x представляє вибірки даних, а y позначає мітки класів. Функція оцінки представлена F(x), а мета оптимізації GBDT з точки зору третьої особи полягає в тому, щоб мінімізувати (15) втрати L(y,F(x)){$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$}:

$$\hat{F} = \arg\min_{F} E_{x,y}\left[L\left(y, F(x)\right)\right] \quad (15)$$

З точки зору третьої особи, ітеративний критерій GBDT може бути отриманий за допомогою лінійного пошуку (16) для мінімізації функції втрат.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (16)$$

де $\gamma_m = arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$, m—кількість ітерацій, $h_m(x)$—являє собою базове дерево рішень.

Але коли кількість зразків велика або розмір ознаки величезний, ефективність і точність GBDT все одно не можуть бути задовільними. GBDT — це ансамблевий алгоритм, базовим класифікатором якого є дерево рішень, тому основна вартість полягає в тому, щоб знайти найкращі точки розбиття при вивченні дерев рішень. Ke et al. [23] запропонували високопродуктивне дерево рішень для підвищення градієнта з використанням градієнтно-орієнтованої односторонньої дискретизації (GOSS) та ексклюзивного об'єднання функцій (EFB) під назвою LightGBM.

Для GBDT приріст інформації зазвичай використовується для поділу кожного вузла. LightGBM використовує GOSS для визначення точки розщеплення шляхом обчислення коефіцієнта дисперсії. По-перше, він вибирає свого роду абсолютні градієнтні значення навчальних прикладів в порядку спадання і верхню вибірку цих значень градієнта називається A. Потім випадковим чином вибирається підмножина B розміру з

невідсортованих вибірок . Нарешті, вибірки розбиваються за допомогою оціненої (17) дисперсії на $a \times 100\% b \times |A^c|A^c V_j(d) A \cup B$.

$$V_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_i} g_i + \frac{1-a}{b}\sum_{x_i \in B_i} g_i\right)^2}{n_i^j(d)} + \frac{\left(\sum_{x_i \in A_i} g_i + \frac{1-a}{b}\sum_{x_i \in B_i} g_i\right)^2}{n_r^j(d)}\right). \quad (17)$$

де $A_i = \{x_i \in A: x_{ij} \leq d\}, A_r = \{x_i \in A: x_{ij} > d\}, B_i = \{x_i \in B: x_{ij} \leq d\}, B_r = \{x_i \in B: x_{ij} > d\}$, g_i являє собою негативний градієнт функції втрат, використовується для нормалізації суми градієнтів$\frac{1-a}{b}$.

Об'ємні функції часто мають тенденцію бути розрідженими, а багато розріджених функцій є винятковими. Таким чином, EFB може бути використаний для прискорення навчання GBDT. Іншими словами, LightGBM може об'єднувати ексклюзивні функції в одну функцію, а алгоритм сканування функцій може бути розроблений для створення однакових гістограм із пакетів функцій. Отже, обчислювальна складність LightGBM зводиться до , де .$O(data \times bundle) O(data \times feature) bundle \times feature$

Таким чином, LightGBM є ефективною реалізацією GBDT з GOSS і EFB для підвищення ефективності обчислень без шкоди для точності. ГОСС використовується для розбиття оптимального вузла шляхом обчислення коефіцієнта дисперсії. EFB може прискорити процес навчання GBDT, поєднуючи багато ексклюзивних функцій з меншою кількістю щільних функцій. LightGBM — це ансамблева модель, базовим класифікатором якої є дерево рішень, яке можна тренувати послідовно, апроксимуючи від'ємні градієнти функції втрат. Модель LightGBM (18) може бути отримана за допомогою зваженої комбінаційної схеми$F_M(x)$.

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x). \quad (18)$$

де M - максимальне число ітерацій і - дерево базових рішень$h_m(x)$.

Отже, бустингові методи, такі як AdaBoost, Gradient Boost, XGBoost, та LightGBM, є важливими інструментами у машинному навчанні, зокрема у задачах класифікації. Ці методи поєднують слабкі класифікатори у сильний ансамбль, послідовно виправляючи помилки попередніх моделей. Вони ефективно використовуються для обробки складних завдань із різноманітними даними, наприклад, у класифікації рівня антропогенних катастроф. Завдяки гнучкості і точності, бустингові методи можуть бути адаптовані до різних типів даних і завдань.

2.3 Інтелектуальний метод класифікації наслідків техногенних катастроф

Інтелектуальний метод класифікації наслідків техногенних катастроф передбачає використання інноваційного підходу для аналізу та оцінки потенційних ризиків, пов'язаних з різними рівнями аварій. Основою цього методу є рейтингова система, яка категоризує аварії залежно від їх потенційної небезпеки, починаючи від мінімальних ризиків і закінчуючи надзвичайно небезпечними ситуаціями. Стратегія аналізу включає збір даних від очевидців, їх ретельний розвідувальний аналіз, а також подальшу класифікацію інформації з використанням передових методів машинного навчання. Це дозволяє не тільки точно оцінити рівень небезпеки, але й ефективно відображати результати для широкого кола зацікавлених сторін.

«Рівні потенційної аварії» – це рейтингова система, яка використовується для визначення рівня потенційної небезпеки аварії. Зазвичай вона включає в себе чотири рівні [24]:

Рівень потенційної аварії I: Мінімальний рівень небезпеки. Аварії такого рівня зазвичай не завдають значної шкоди і можуть бути легко вирішені без особливих зусиль.

Рівень потенційної аварії II: Середній рівень небезпеки. Нещасні випадки такого рівня можуть завдати помірної шкоди та вимагати більше зусиль для вирішення.

Рівень потенційної аварії III: Високий рівень небезпеки. Нещасні випадки такого рівня можуть завдати серйозної шкоди і вимагати значних зусиль для їх усунення.

Рівень потенційної аварії IV: Рівень надзвичайної небезпеки. Нещасні випадки такого рівня можуть завдати катастрофічних збитків і вимагати надзвичайних зусиль для їх усунення.

Розроблений інтелектуальний метод ґрунтується на розвідувальному аналізі даних, зібраних від очевидців ДТП, роздільній класифікації текстових та кількісних даних за допомогою методів машинного навчання та подальшій візуалізації результатів. Відповідно до цього концептуальна структура інтелектуального методу включає шість основних сегментів: 1 — збір даних, 2 — розвідувальний аналіз даних, 3 — класифікація текстових даних, 4 — класифікація кількісних даних, 5 — відображення оцінок і 6 — відображення результатів (рис. 2.2).

На етапі аналізу даних важливо скоротити час, витрачений на визначення рівня потенційної небезпеки в регіоні.

Метод може бути представлений набором фаз (які відповідають відрізкам вище) і ступенів (див. рисунок 2.2):

Фаза 1 запускається як збір даних. Дані можуть бути зібрані у формі опитування свідків надзвичайних ситуацій.

Розглянемо *Фазу* 2, яка виконується як Розвідувальний аналіз даних (EDA) (Сегмент 2) [25,26]. EDA використовують, з одного боку, для відповідей на запитання, перевірки базових припущень і створення гіпотез для подальшого аналізу. З іншого боку, це може додатково допомогти при вирішенні інших проблем, пов'язаних з очищенням/підготовкою/конвертацією даних. Фаза 2 складається з чотирьох етапів: дослідження даних (2.1), очищення даних (блок 2.2), побудова моделі (2.3) та представлення результатів (2.4).

Що стосується класифікації текстових (Сегмент 3) і кількісних (Сегмент 4) даних, то вони розглядаються окремо (див. Рис. 1) через відмінності в їх структурі і характеристиках. Зокрема, текстові дані зазвичай складаються з неструктурованого тексту, який може бути важко проаналізувати. Методи класифікації текстових даних, як правило, включають методи обробки природної мови для перетворення тексту в числові вектори з подальшим аналізом.

Кількісні дані зазвичай складаються з числових значень, які можна легко проаналізувати за допомогою різних статистичних методів і методів машинного навчання.

Розглянемо *Фазу 3*, яка виконується як Класифікація, заснована на текстових даних. Найшвидший спосіб отримати дані – використовувати описи очевидців аварії. Цей етап включає в себе наступні дев'ять етапів:

Крок 3.1: Використовуються текстові дані (Блок 3).

Крок 3.2: Очищення текстових даних (Блок 4) [27,28]. Він включає в себе кілька підетапів: видалення стоп-слів (4.1), нормалізація (4.2), лемматизація (4.3) і токенізація (4.4).

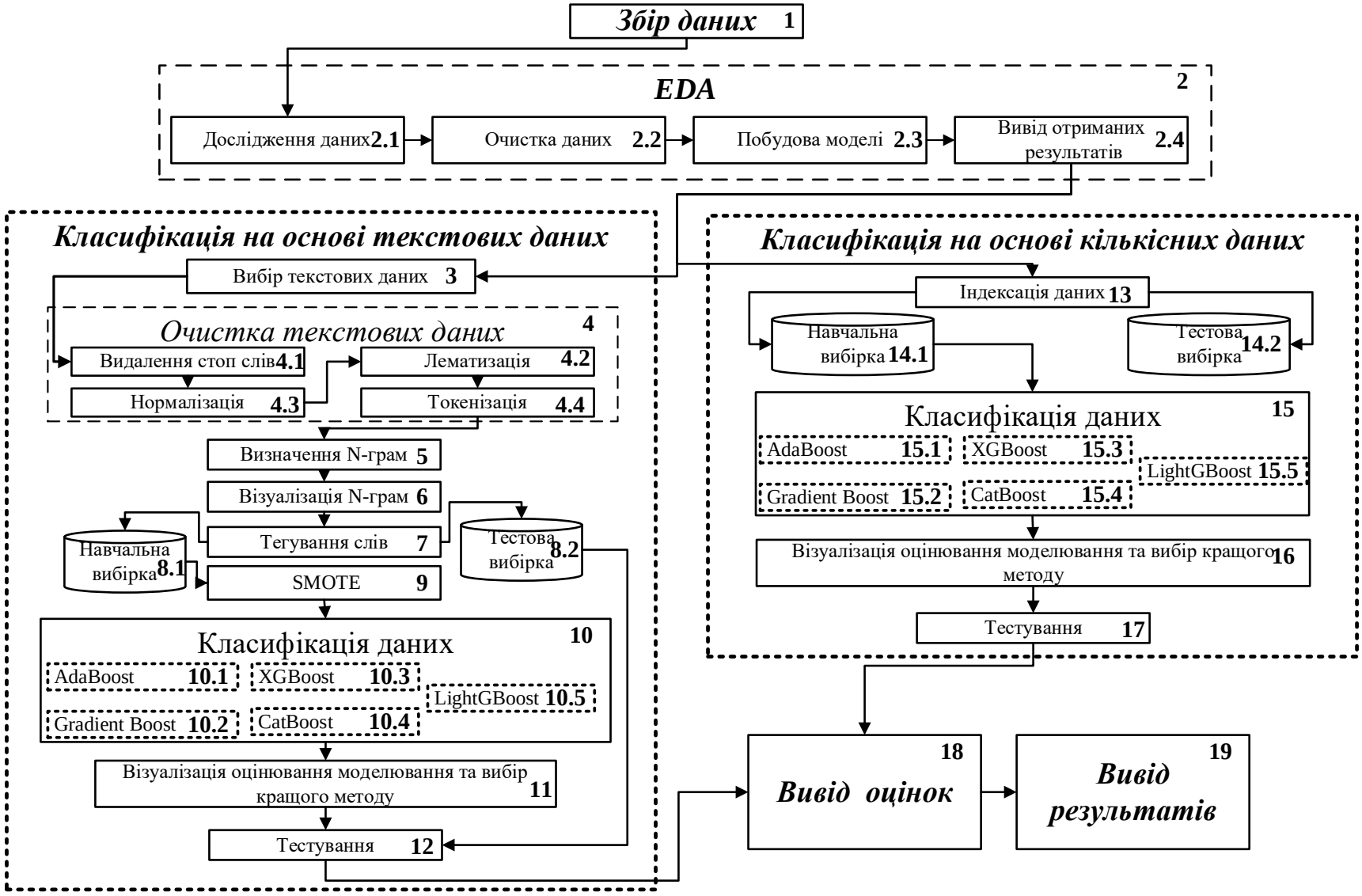


Рисунок 2.2 - Концептуальна структура інтелектуального методу класифікації техногенних катастроф

Крок 3.3: Визначення (Блок 5) і візуалізація (Блок 6) N-грам [27]. N-грама може бути визначена як безперервна послідовність з *n* елементів із заданого зразка тексту або мови. Елементами можуть бути літери, слова або пари основ відповідно до програми. N-грами, як правило, збираються з текстового або мовленнєвого корпусу (довгого текстового набору даних).

Крок 3.4: Додавання тегів до слів (Блок 7). Тегування частин мови (POS) [29] — це процес перетворення речення на форми — список слів, список кортежів, де кожен кортеж має форму (слово, тег). Тег в даному випадку є частиною мовного тега і вказує на те, чи є слово іменником, прикметником, дієсловом і т.д. Теги частин мови дають уявлення про граматичну структуру слів у реченні чи тексті, тим самим дозволяючи робити висновки про їхню семантичну роль. Інші застосування POS-тегів включають: розпізнавання іменованих сутностей; розв'язання проблеми спільного виникнення; і розпізнавання мови. Коли виконується тегування POS, часто трапляється, що тегувальник стикається зі словами, які не є частиною словникового запасу, який використовувався.

Крок 3.5: Після цього набір даних розбивається на підмножини «Навчальна вибірка» (8.1) і «Тестова вибірка» (8.2).

Крок 3.6: Техніка передискретизації синтетичних меншин (SMOTE) [30] (9). Проблема незбалансованої класифікації полягає в тому, що існує занадто мало прикладів класу меншин, щоб модель могла ефективно вивчити межу прийняття рішень. Одним із способів розв'язання цієї проблеми є надмірна вибірка прикладів у класі меншин. Цього можна досягти, просто скопіювавши екземпляри з класу minority в навчальному наборі даних перед тим, як підігнати модель. Це може збалансувати розподіл класів, але не дає жодної додаткової інформації до моделі. Удосконаленням копіювання прикладів з класу меншин є синтез нових прикладів з класу меншин. Це тип доповнення даних для табличних даних, який може бути дуже ефективним. Мабуть, найбільш широко використовуваний підхід до синтезу нових прикладів називається SMOTE.

SMOTE визначає метод підготовки даних при налаштуванні та оцінці алгоритмів машинного навчання.

Крок 3.7: Класифікуйте дані (Блок 10) за допомогою ансамблевого методу машинного навчання з використанням підходу boosting. Як згадувалося вище, методи підвищення генерують колекцію моделей, які разом утворюють надійного учня з чудовою продуктивністю та високою точністю. Тому для класифікації текстових даних ми вибираємо такі методи, як AdaBoost (10.1) по (1)-(3), GradientBoost (10.2) по (3)-(5), XGBoost (10.3) по (6)-(12), CatBoost (10.4) по (13)-(15) і LightGBM (10.5) по (15)-(18).

Крок 3.8: Візуалізація моделювання, оцінка та вибір найкращого методу (Блок 11). Вибір оптимального методу ґрунтується на моделюванні оцінок за параметрами: Accuracy, Precision, Recall і F1-Score [31].

Точність — у багатокласовій класифікації точність — це показник, який вимірює відповідність між набором міток, призначених зразку, і фактичним набором міток. Для точності обидва набори етикеток повинні ідеально збігатися.

Точність – це відношення між кількістю істинних спрацьовувань і кількістю помилкових спрацьовувань. Найкраще значення – 1, а найгірше – 0.

Відкликання - це співвідношення між числом істинних спрацьовувань і числом помилкових негативних результатів. Найкраще значення – 1, а найгірше – 0.

Оцінка F1 — оцінка F1 може бути інтерпретована як середнє гармонійне значення точності та запам'ятовування. Оцінка Формули-1 досягає свого найкращого значення в 1, а найгіршого - в 0.

Крок 3.9: Проведення класифікації на тестовій вибірці (Блок 12).

Розглянемо *4-й етап,* який проводиться як класифікація, заснована на кількісних даних. Фаза 4 включає наступні п'ять етапів:

Крок 4.1: Індексація даних (Блок 13).

Крок 4.2: Поділ даних на навчальні (Блок 14.1) та тестові (Блок 14.2) вибірки.

Крок 4.3: Класифікація даних (Блок 15) включає п'ять підкроків з використанням AdaBoost (15.1) відповідно до (1)–(3), GradientBoost (15.2) відповідно до (3)–(5), XGBoost (15.3) відповідно до (6)–(12), CatBoost (15.4) відповідно до (13)–(15) та LightGBM (15.5) відповідно до (15) – (18).

Крок 4.4: Візуалізація отриманих результатів (Блок 16).

Крок 4.5: Проведення класифікації на досліджуваному зразку (Блок 17). Розглянемо два заключних етапи (див. Малюнок 1).

Етап 5: Відображення результатів (Блок 18). На цьому етапі результати моделі, отримані після навчання та тестування, відображаються для аналізу. Це може включати відображення таких показників, як точність, запам'ятовування, оцінка F1 та інші, які допомагають оцінити ефективність моделі. Відображення результатів може бути виконано у вигляді таблиць, діаграм або інших візуалізацій, що дозволяє легко інтерпретувати результати.

Етап 6: Відображення результатів (Блок 19). Цей етап може включати відображення прогнозованих міток для тестового набору даних, порівняння прогнозованих міток з фактичними, а також відображення важливості функцій. Цей етап допомагає визначити, наскільки добре модель працює на нових, невидимих даних, і які функції є найважливішими для прогнозування.

Отже, застосування цього комплексного мтоду, який включає ретельний збір даних, їх аналіз, класифікацію, а також візуалізацію результатів, виявляється ефективним у визначенні рівнів потенційної небезпеки техногенних катастроф. Класифікація, яка базується на ретельно оброблених текстових та кількісних даних з використанням ансамблевих

методів машинного навчання, забезпечує високу точність та надійність результатів. Особливо важливим є використання інноваційних технологій, таких як SMOTE для балансування даних та різноманітних методів бустингу, що дозволяють оптимізувати процес аналізу та підвищити якість прогнозування. Цей підхід демонструє великий потенціал у покращенні реакції на надзвичайні ситуації, сприяючи своєчасному та точному реагуванню на потенційні катастрофи.

Висновки до розділу 2

У сучасному аналізі текстових даних, методи обробки природної мови (NLP) як видалення стоп-слів, лематизація, нормалізація, токенізація, визначення та візуалізація N-грам, тегування слів, а також техніки, такі як SMOTE, відіграють ключову роль. Вони не тільки сприяють ефективному зведенню даних до більш оброблюваної форми, але й забезпечують важливе розуміння семантичних та синтаксичних відносин у тексті. Це спрощує подальший аналіз, забезпечуючи вищу точність та ефективність у вирішенні різноманітних завдань NLP. Водночас, такі методи як SMOTE допомагають управлінню незбалансованими наборами даних, сприяючи кращому розумінню та класифікації різних патернів у даних. Всі ці методи разом формують міцну основу для розширення горизонтів в області машинного навчання та аналізу даних.

Методи бустингу, включаючи AdaBoost, Gradient Boost, XGBoost та LightGBM, представляють собою ключові техніки в машинному навчанні для розв'язання задач класифікації. Вони ефективно інтегрують низку слабких класифікаторів в потужний колективний ансамбль, послідовно усуваючи неточності попередників. Ці методи демонструють високу ефективність при роботі з складними наборами даних, як-от у випадку класифікації антропогенних катастроф. Відзначаючи свою універсальність і точність, бустингові методи знаходять широке застосування в різноманітних областях і можуть бути налаштовані для різних видів даних і завдань.

Розглянутий інтелектуальний метод класифікації наслідків техногенних катастроф, що ґрунтується на комплексному аналізі даних зібраних від очевидців, виявився ефективним інструментом у визначенні різних рівнів потенційної аварії. Через використання розширеного набору методів, включаючи розвідувальний аналіз даних, детальну класифікацію

текстових та кількісних даних, а також застосування передових технік машинного навчання, таких як бустингові алгоритми AdaBoost, Gradient Boost, XGBoost, CatBoost та LightGBM, метод демонструє здатність точно оцінювати і візуалізувати рівні ризику. Такий підхід дозволяє швидко і точно аналізувати ситуації, сприяючи своєчасному та ефективному управлінню відповіддю на катастрофи. Це, в свою чергу, підвищує можливості для розробки ефективних планів дій та стратегій запобігання подібним подіям у майбутньому.

3 РЕАЛІЗАЦІЯ ІНТЕЛЕКТУАЛЬНОГО МЕТОДУ

3.1 Попередній аналіз даних

У якості вхідних даних використано записи про аварії на 12 різних підприємствах в Україні, кожен рядок у яких містить інформацію про аварію. Кожен рядок у записах містить інформацію про аварію, включаючи дату та час аварії, місце виникнення, тип промисловості, причини та наслідки аварії, відомості про кількість потерпілих, рівень катастрофи та потенційно можливий рівень катастрофи.

Далі дані про аварії були оброблені та підготовлені підходом попереднього аналізу для використання в розробленому методі. Для цього були виконані наступні дії:

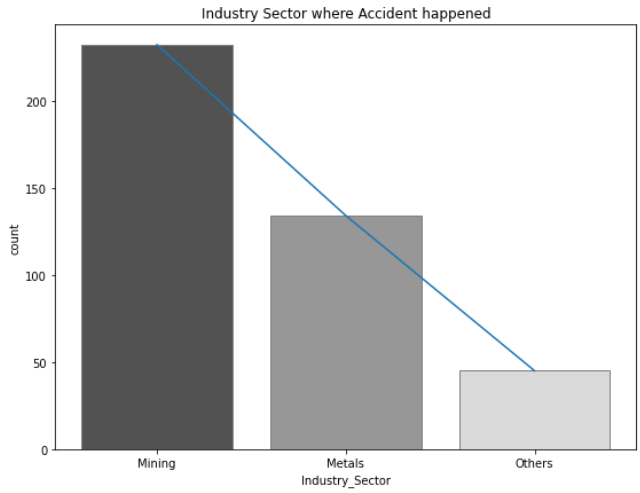
- Зчитування даних з файлу та перетворення у формат, придатний для подальшої обробки.
- Очищення та нормалізація даних, включаючи видалення непотрібної інформації та стандартизацію формату записів.

Для проведення інтелектуального методу було використано бібліотеки машинного навчання, такі як scikit-learn, pandas та numpy. Після попередньої обробки даних, які включали в себе зведення та очищення даних, а також інженерію ознак, було побудовано модель машинного навчання, яка використовувала ансамблеві методи класифікації.

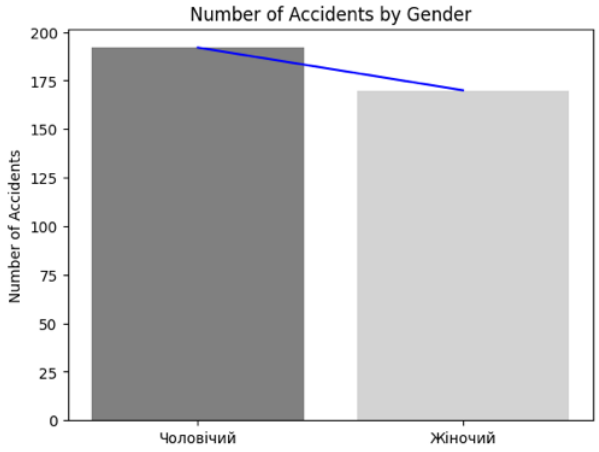
Для оцінки точності класифікації було використано крос-валідацію, що дозволило отримати точність моделі на методах. Після навчання та перевірки моделі на валідаційній вибірці, вона була застосована для оцінки наслідків техногенних катастроф на інших підприємствах та в інших регіонах.

Для детального розуміння даних проведемо Exploratory Data Analysis, на першому етапі проведемо візуальну оцінку показників. В гірничодобувній промисловості відбувається більше аварій і катастроф (рис.3.1а). Можна сказати, що кількість нещасних випадків у гірничій промисловості значно більша, ніж у

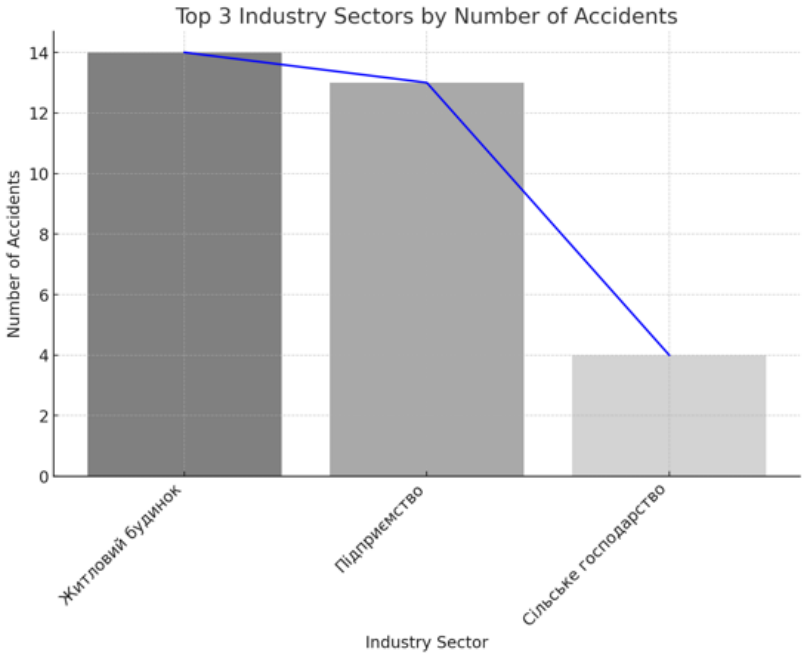
металургійній промисловості, тому робота в гірничій промисловості є більш ризикованою, ніж остання. Характерною рисою кожної галузі є переважна частка чоловіків (рис.3.1b). Набір даних має упередження щодо працівників чоловічої статі. Загальна кількість (рис.3.1c) прямих співробітників і сторонніх співробітників майже однакова, однак віддалених співробітників третіх сторін менше.



a



b



c

Рисунок 3.1 - Кількісна оцінка показників

На Local_03 зафіксовано максимальну кількість аварій, що становить приблизно 21% від усіх заводів країни. Далі йдуть Local-05, Local-01 і так далі.

Більшість критичних ризиків класифікуються як «Інші», що становить майже 50% набору даних, отже існує занадто багато ризиків, які потребують точної класифікації. Далі йдуть пресовані, ручні інструменти, хімічні речовини, різання тощо.

Рівень явної або потенційної небезпеки - I означає неважкий, а V означає дуже серйозний. Найбільш поширені аварії з I рівнем. Це пов'язано з невеликими помилками, як-от люди, які забули свої засоби індивідуального захисту, або вони впустили інструмент тощо. Рівень потенційної аварії вказує на те, наскільки серйозною буде аварія через інші фактори, пов'язані з аварією. Рівень потенційної аварії IV має найбільшу кількість і означає помірну тяжкість аварій.

Далі потрібно перевірити співвідношення між цим рівнем потенційної аварії та рівнем аварії разом (рис. 3.2) і відносно сектора промисловості (рис.3.3).

Існує значна різниця між серйозністю інциденту та потенційною серйозністю інциденту (рис.3.2). Також, з рисунку 3.2 видно, що якщо кількість аварій збільшується то потенційний рівень аварійності зменшується. А якщо рівень аварії підвищується, потенційний рівень аварії також зростає. Ще з графіку видно, що здебільшого відбуваються аварії 3 рівня.

На графіку (рис.3.3а,) представлено кількість потенційних аварій за рівнями в залежності від сектора промисловості. У гірничодобувній промисловості спостерігається найвища кількість інцидентів з найсерйознішим потенційним рівнем аварії (IV), що вказує на значний ризик у цій галузі. Металургійна промисловість також має інциденти, проте з меншим потенційним рівнем серйозності (III). В інших секторах, таких як будівництво і транспорт, зареєстровано інциденти нижчих рівнів серйозності (II та I).

Це може свідчити про більш високий рівень ризиків або про недоліки у заходах безпеки в гірничодобувному секторі у порівнянні з іншими галузями. Важливо зазначити, що наявність інцидентів рівня IV у

гірничодобувній промисловості вимагає особливої уваги та можливо вдосконалення заходів безпеки для запобігання майбутнім подіям.

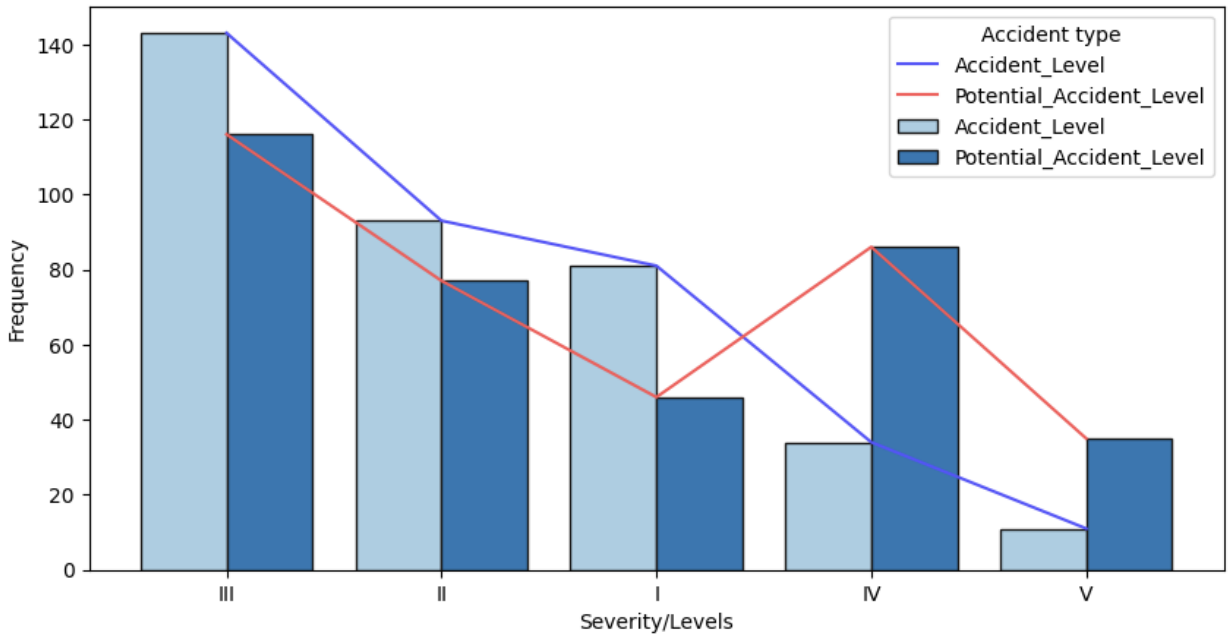
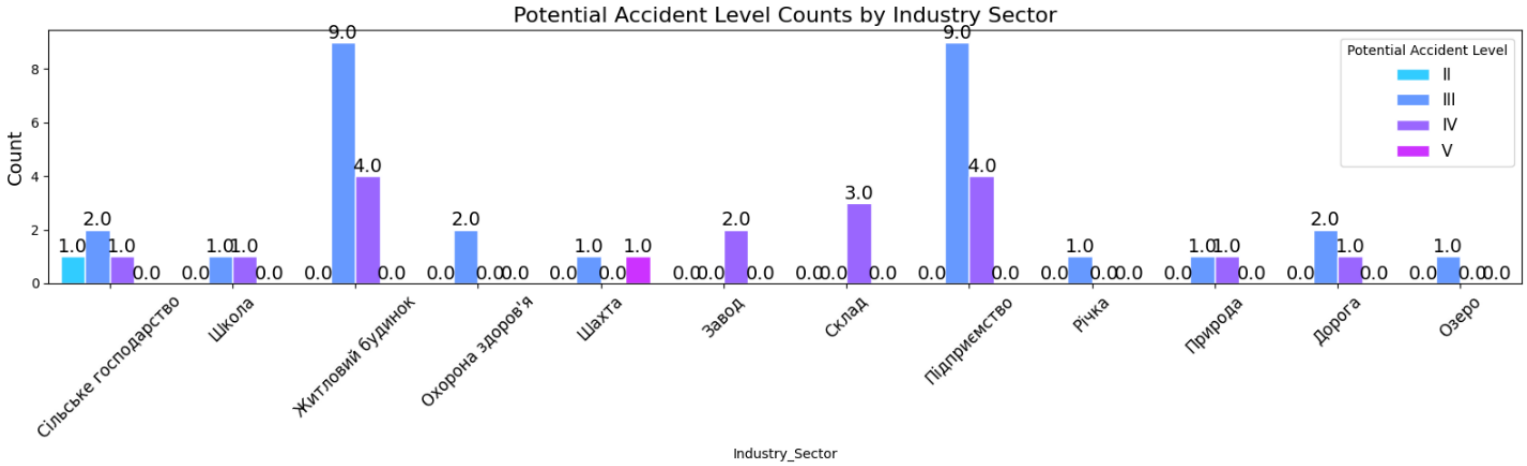


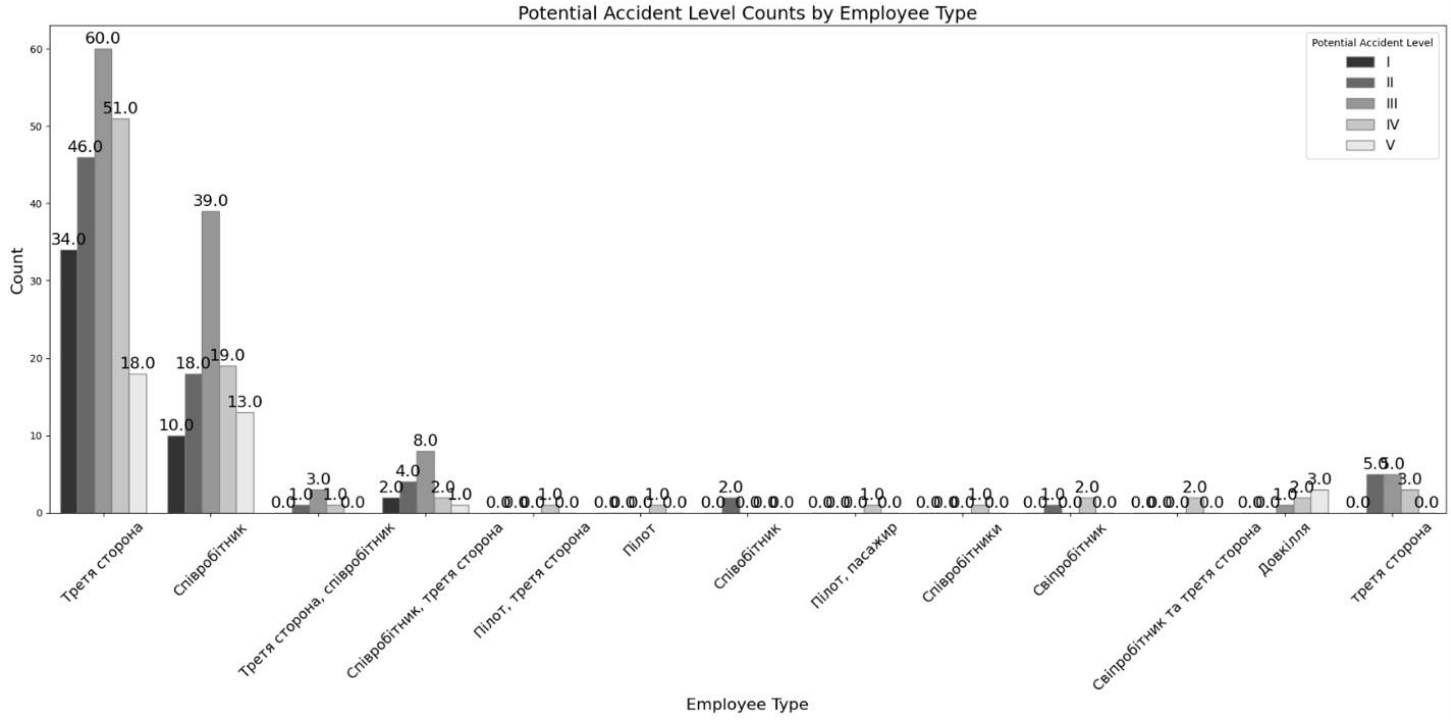
Рисунок 3. 2 - Співвідношення між рівнем потенційної аварії та явним рівнем аварії.

На графіку (рис.3б) відображено кількість потенційних рівнів аварійності по типу працівників. Зокрема, виділені дві групи: "Третя Сторона" та "Співробітники". У групі "Третя Сторона" є інциденти з потенційним рівнем аварійності II та III, з найвищим числом інцидентів рівня III, що становить 9. В групі "Співробітники" спостерігається значна кількість інцидентів рівня IV (найсерйозніший), а також інциденти рівня II.

Цей розподіл може вказувати на більш високий ризик серйозних інцидентів серед співробітників порівняно з третіми сторонами, що може бути пов'язано з характером їхньої роботи або умовами праці. Зокрема, важливість заходів з охорони праці та профілактики аварій для співробітників стає особливо актуальною, враховуючи високий рівень потенційно небезпечних інцидентів.



а



б

Рисунок 3.3 - Співвідношення між рівнем потенційної, явним рівнем аварії та сектором промисловості

Після попереднього аналізу даних про аварії на різних підприємствах в Україні можна зробити наступні висновки: дані відображають значну кількість інцидентів у гірничодобувній промисловості, що вказує на високий рівень ризику в цій галузі. Аналіз також показав, що більшість нещасних випадків стосуються чоловічих працівників, що може свідчити про гендерні упередження у робочих умовах або характері праці. Значна частина інцидентів класифікується як «Інші» у категорії критичних ризиків, що вимагає більш детальної класифікації та аналізу. Крім того, аналіз показав, що в гірничодобувній промисловості спостерігається висока кількість інцидентів з серйозними потенційними ризиками. Ці результати вказують на потенційні напрямки для покращення заходів безпеки та запобігання аваріям, особливо в гірничодобувній промисловості. Важливо продовжувати дослідження та аналіз даних для розробки ефективних стратегій запобігання і реагування на потенційні техногенні катастрофи.

3.2 Попередній аналіз текстових даних

На (рис.3.4) зображено хмару слів, яка є візуальним представленням тексту, де частота зустрічаності кожного слова визначає його розмір у хмарі. Більш часті слова зображені більшим шрифтом. Хмара слів дає швидке інтуїтивне розуміння ключових слів або тем у великому текстовому наборі даних.

З хмари слів можна зробити наступні спостереження:

- Ключові Терміни: Слова "пожежа", "наряд", "повідомлення", "території", "підприємств", і "району" є одними з найбільших у хмарі, що свідчить про їх значну частоту в тексті. Це може вказувати на те, що текст стосується повідомлень про пожежі на територіях підприємств чи в окремих районах.
- Контекст: Менші слова, такі як "Служби", "направлено", "будинку", "інформація", "службовців", і "Києва" допомагають формувати

контекст навколо ключових термінів. Вони можуть відноситися до дій, що були вжиті під час пожеж, служб, які реагували на інциденти, та місць, де сталися інциденти.

- Можливий Фокус: Оскільки певні слова, такі як "пожежа", "району", та "підприємств", з'являються великим шрифтом, можливо, що хмара слів пов’язана з інцидентами, специфічними для певного регіону або типу діяльності.
- Інші Слова: Багато слів зустрічаються менш часто і зображені меншим шрифтом. Вони можуть містити додаткову, але менш важливу інформацію або відноситися до більш специфічних аспектів або подій, що описуються в тексті.

Загалом, хмара слів є корисним інструментом для візуалізації популярних термінів у великих текстових наборах даних. Вона може вказувати на загальні теми та поняття, які часто зустрічаються, і допомагає у виявленні можливих шаблонів у текстових даних.

Рисунок 3.4 - Хмара ключових слів повідомлень про інциденти

З кореляції матриці (рис.3.5) видно, що існує певний зв'язок між рівнями потенційної небезпеки та фактичними рівнями нещасних випадків. Найсильніша позитивна кореляція спостерігається між рівнями "II" та "III" потенційної небезпеки з "I" рівнем фактичної аварії.

Після проведення попередньої обробки наших даних, проведемо обробку текстових записів про аварійну ситуацію. Це дозволить нам провести класифікацію рівня техногенної катастрофи відносно опису очевидців.

На першому етапі проведемо очистку тексту, а саме видалимо стоп слова, проведемо нормалізацію, лематизацію та токенізацію тексту. Результати представлено на рисунку 3.6.

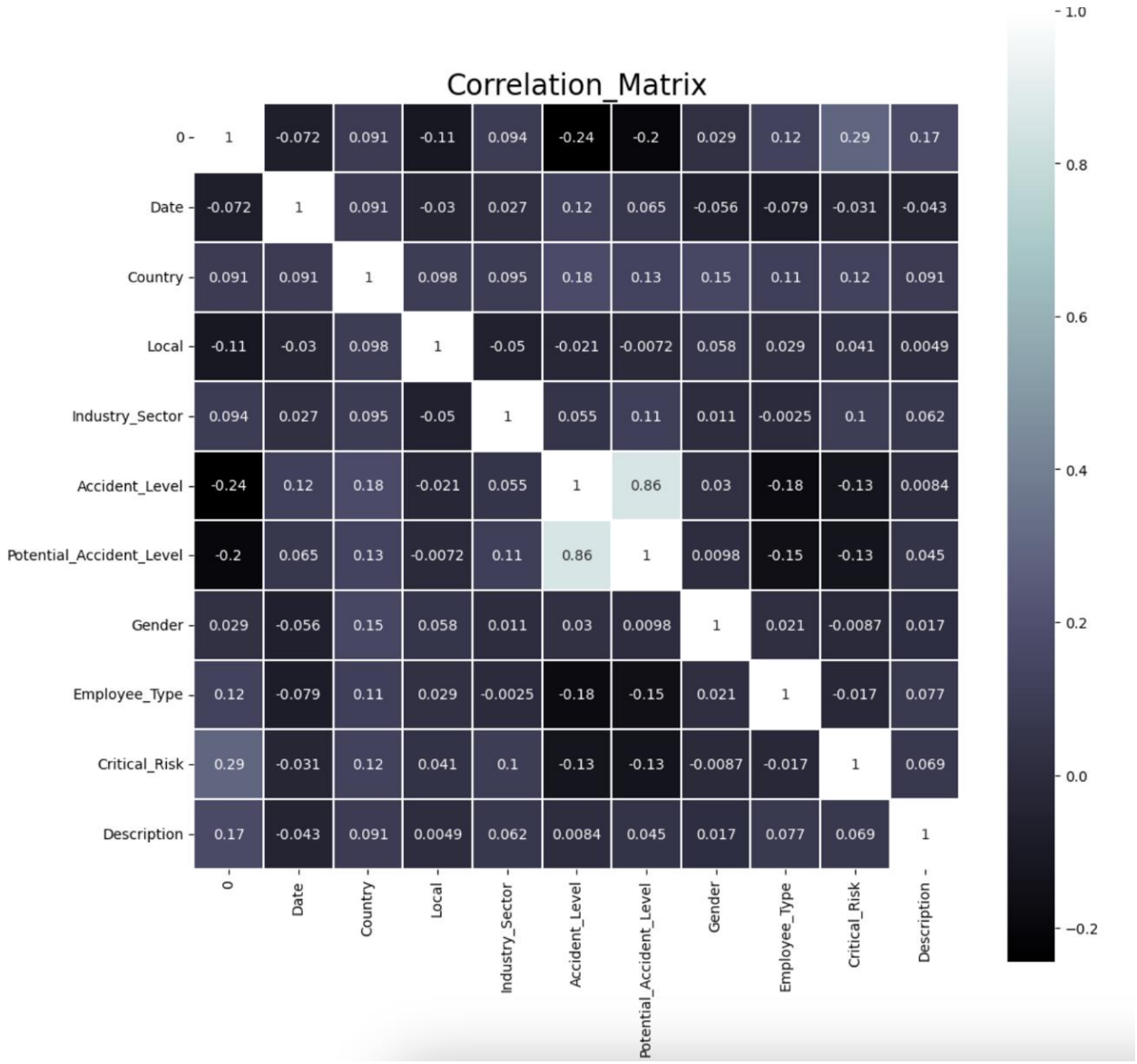


Рисунок 3.5 - Кореляція між рівнем потенційної аварії та явним рівнем аварії

```
'Пожежа сінника Сягала площі 15 квм'
```

Рисунок 3.6 - Очищений текст

На наступному кроці проведено визначення N-grams відносно категорій промисловості (рис.3.7). Є багато слів, які пов'язані з місцем інциденту, такі як "стався" і "пожежі", а також слова, пов'язані з реагуванням на інциденти, як "місті" і "складі", часто зустрічаються у тексті.

Words	Count
стався	51
пожежі	49
будинку	42
пожежа	40
місті	39
результаті	36
газу	33
складі	33
області	33
сталася	32

Рисунок 3.7 - N-grams відносно категорій промисловості.

Далі переводимо текст в теги, де: NN – noun; NNP - proper noun; NNS - noun plural; CD - cardinal digit; DT – determiner; VB – verb; JJ – adjective; RB – adverb; VB – verb; VBD - verb past tense. Найбільше присутніх в одному з пояснень noun.

Отже, висновок до попереднього аналізу текстових даних про інциденти відкриває важливі інсайти щодо популярних термінів та тем, які зустрічаються в описах аварій. Хмара слів і кореляційна матриця вказують на зв'язок між рівнями потенційної небезпеки та фактичними рівнями аварій, особливо між рівнями "II" та "III" потенційної небезпеки з "I" рівнем фактичної аварії. Проведення попередньої обробки тексту, що включає очищення, нормалізацію, лематизацію та токенізацію, дозволило детальніше зануритися в контент інцидентів. Визначення N-grams у відповідності до категорій промисловості підкреслює часте згадування певних місць і дій у відповідях на інциденти. Цей аналіз допомагає глибше зрозуміти контекст інцидентів та сприяє розробці ефективних стратегій для їх виявлення та класифікації, зокрема з огляду на специфічні характеристики кожного сектора промисловості.

3.3 Реалізація класифікації наслідків техногенних катастроф в Україні

У сучасному світі, де вплив технологій на наше повсякденне життя є беззаперечним, питання безпеки стають все більш актуальними. Техногенні катастрофи, які включають широкий спектр подій від промислових аварій до екологічних криз, представляють собою великий ризик для здоров'я, добробуту людей та стабільності довкілля. Ці катастрофи можуть мати довготривалі наслідки, які тягнуть за собою не тільки економічні збитки, а й втрати людських життів та негативний вплив на природу.

Визнання необхідності ефективного управління ризиками та реагування на подібні виклики спонукає до застосування передових методів аналізу даних та штучного інтелекту. Класифікація наслідків техногенних катастроф стає ключовим елементом у стратегіях мінімізації ризиків та розробки планів евакуації і реагування. Ця робота покликана дослідити потенціал застосування машинного навчання для класифікації наслідків техногенних подій, з метою забезпечення більш оперативного та точного реагування на майбутні загрози.

Дослідження фокусується на розробці та впровадженні класифікаційної моделі, спеціально адаптованої для умов України, країни з унікальним набором техногенних ризиків. Ми розглянемо набір даних, який охоплює різноманіття техногенних катастроф, методику їх аналізу, а також стратегії впровадження отриманих моделей у реальні системи управління надзвичайними ситуаціями.

Було проведено моделювання методами класифікації на основі машинного навчання з використанням техніки SMOTE. Спершу розділимо дані на навчальну та тестову вибірку, а далі переведемо нашу вибірку у

масив та проведено передискретизація набору навчальних даних за допомогою SMOTE (рис.2.8).

```
array([[   0,    0,    0, ...,   83,  337,  298],
       [   0,    0,    0, ...,  207, 1492, 1493],
       [   0,    0,    0, ...,   15,    2,    8],
       ...,
       [   0,    0,    0, ...,   81,   60,   12],
       [   0,    0,    0, ...,  275, 1151,  909],
       [   0,    0,    0, ...,  260,  383,  412]], dtype=int32)
```

Рисунок 3.8 - Результат передискретизації набору навчальних даних за допомогою SMOTE

Було проведено швидке моделювання різними методами машинного навчання, і отримано наступні результати щодо точності класифікації (рис.3.9): логістична регресія (LogReg) - 0.5; наївний байєсівський класифікатор (Naive Bayes) - 0.3; метод k-найближчих сусідів (KNN) - 0.4; метод опорних векторів (SVM) - 0.5; дерево рішень (Decision Tree) - 0.5; випадковий ліс (RandomForest) - 0.3; беггінг (Bagging) - 0.4; AdaBoost - 0.3; градієнтний бустинг (Gradient Boost) - 0.4; XGBoost - 0.4.

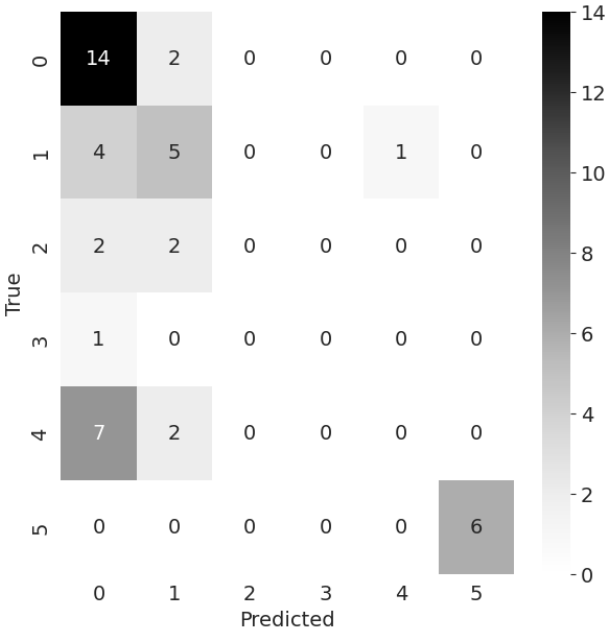
З усіх протестованих моделей, логістична регресія, метод опорних векторів і дерево рішень показали однаково найвищу точність – 0.5.

	model	accuracy
0	LogReg	0.5
1	Naive Bayes	0.3
2	KNN	0.4
3	SVM	0.5
4	Decision Tree	0.5
5	RandomForest	0.3
6	Bagging	0.4
7	AdaBoost	0.3
8	Gradient Boost	0.4
9	XGBoost	0.4

Рисунок 3.9 - Результати моделюванні ансамблевими методами машинного навчання.

Після проведення моделювання на тестовій вибірці, сформовано звіт у вигляді Confusion Matrix (рис.3.10). Подивившись на наведену матрицю невідповідностей, можна зробити висновок, що модель здатна правильно класифікувати більшість тестових точок даних до фактичних. Правильно класифіковано 14 значень I рівня загрози з 46 правильно визначених значень, а V рівень небезпеки при аварії правильно класифіковано 8 рази.

З результатів точності видно, що краще класифікується потенційний V та I рівень небезпеки при аварії.



	precision	recall	f1-score	support
0	0.50	0.88	0.64	16
1	0.45	0.50	0.48	10
2	0.00	0.00	0.00	4
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	9
5	1.00	1.00	1.00	6
accuracy			0.54	46
macro avg	0.33	0.40	0.35	46
weighted avg	0.40	0.54	0.46	46

Рисунок 3.10 - Матриця плутанини між прогнозами та фактами

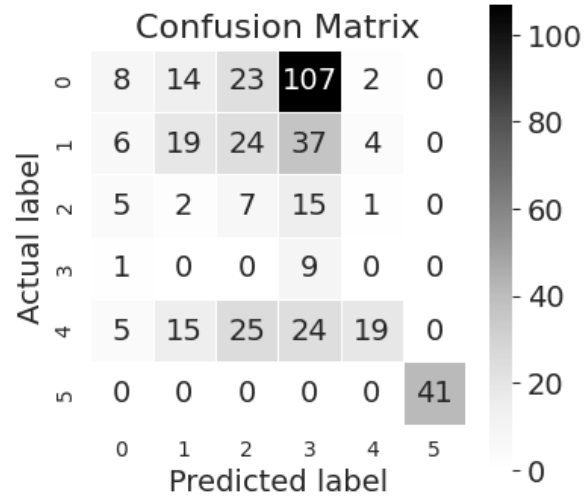
На першому етапі потрібно перевести дані в індексні значення. Далі виділяємо дані на тестові та навчальну вибірку (Рис.3.11).

```
Train set has 43 records out of 48 which is 90%
Dev set has 43 records out of 48 which is 90%
Test set has 5 records out of 48 which is 10%
```

Рисунок 3.11. Результати поділу даних

Проведемо класифікацію Potential Accident Level на основі ансамблевих методів: AdaBoost, Gradient Boost, XGBoost та CatBoost. Проведемо до кожного типу моделювання звіт про класифікацію, що показує представлення основних показників класифікації для кожного класу. Це дає глибше уявлення про поведінку класифікатора над глобальною точністю, яка може маскувати функціональні недоліки в одному класі багатокласової проблеми. Візуальні звіти про класифікацію використовуються для порівняння моделей класифікації, щоб вибрати моделі, які є «темнішими», мають сильніші показники класифікації або більш збалансовані.

Першим розглянемо оцінку моделювання AdaBoost (рис.3.12). З матриці та звіту видно, що найбільш точно буде класифіковано Industry Sector Оцінка класифікації: Зазначається, що найбільш точно класифікується "Industry Sector" (клас №0) з точністю 75%. Модель правильно передбачає цей клас на 82%. Для цього класу також наводиться оцінка f1-score, яка дорівнює 0,089 для 154 значень. Потенційний рівень аварій: Зазначається, що в середньому точність прогнозування "Potential Accident Level" рівна 63% для вибірки розміром 382 значення, що вважається низьким результатом.

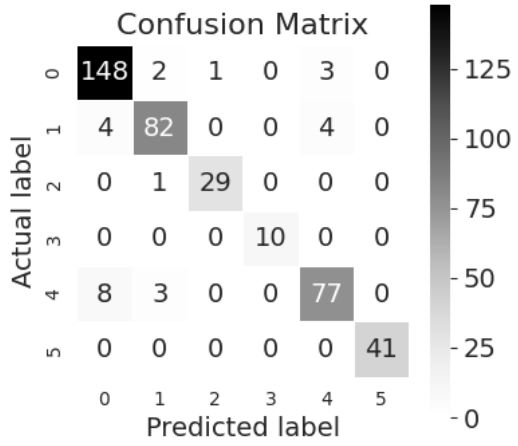


	precision	recall	f1-score	support
0	0.32000	0.05195	0.08939	154
1	0.38000	0.21111	0.27143	90
2	0.08861	0.23333	0.12844	30
3	0.04688	0.90000	0.08911	10
4	0.73077	0.21591	0.33333	88
5	1.00000	1.00000	1.00000	41
accuracy			0.24939	413
macro avg	0.42771	0.43538	0.31862	413
weighted avg	0.46468	0.24939	0.27427	413

Рисунок 3.12 - Оцінка моделювання AdaBoost

Наступним проведемо оцінку моделювання Gradient Boost (рис.3.13). Клас з номером 2 (Employee or Third Party, згідно вашого прикладу) має ідеальну оцінку precision, recall та f1-score (1.0 або 100% для всіх трьох метрик), із загальною кількістю 30 значень. Це свідчить про те, що модель ідеально класифікувала всі випадки цього класу. Для класу номер 0, який можемо умовно назвати Industry Sector, модель показала також високу точність з оцінкою precision 0.9250 та recall 0.96104, з f1-score 0.94268. Було класифіковано 154 значення, що робить його найбільш численним класом у вибірці.В середньому, точність моделі (accuracy) становить 0.93705 або 93.705% для всієї вибірки розміром 413 значень, що вказує на високу ефективність моделі.

Звертаючи увагу на деталі, можна зробити висновок, що модель має високий рівень точності загалом, з особливо вражаючими результатами для деяких класів.

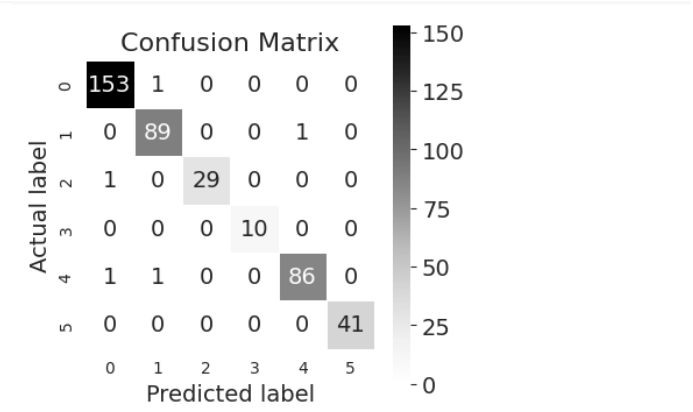


	precision	recall	f1-score	support
0	0.92500	0.96104	0.94268	154
1	0.93182	0.91111	0.92135	90
2	0.96667	0.96667	0.96667	30
3	1.00000	1.00000	1.00000	10
4	0.91667	0.87500	0.89535	88
5	1.00000	1.00000	1.00000	41
accuracy			0.93705	413
macro avg	0.95669	0.95230	0.95434	413
weighted avg	0.93700	0.93705	0.93677	413

Рисунок 3.13 - Оцінка моделювання Gradient Boost

Класифікація за методом XGBoost (рис.3.14) найточнішою була для класу №2, де модель показала 100% точності (precision), 96.667% повноти (recall), та F1-бал 0.98305, що є високим показником для 30 значень, що належать до цього класу. Клас №0, який ми умовно називаємо "Industry Sector", також був добре класифікований з 98.710% точності, 99.351% повноти, та F1-балом 0.99029 для 154 значень. Це свідчить про високу точність моделі у передбаченні цього класу. Середня точність (accuracy) класифікації для всієї вибірки становить 98.789%, що вказує на загальну високу ефективність моделі на розмірі вибірки у 413 значень. Середній F1-бал (macro avg) та вагований середній F1-бал (weighted avg) складає близько 0.989, що також є високими показниками ефективності.

Ці результати вказують на те, що модель XGBoost дуже ефективно класифікує дані за всіма класами, зокрема відмінно розпізнаючи класи "Employee or Third Party" та "Industry Sector".

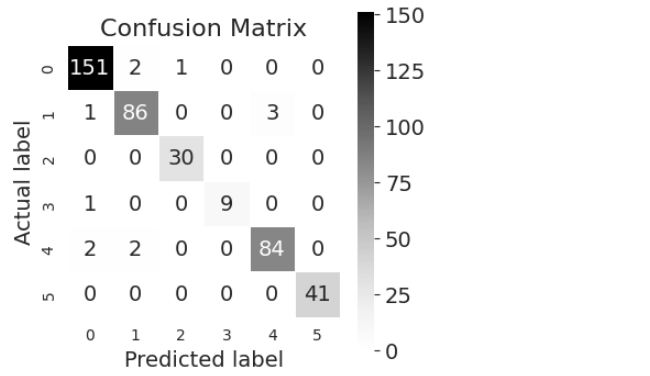


	precision	recall	f1-score	support
0	0.98710	0.99351	0.99029	154
1	0.97802	0.98889	0.98343	90
2	1.00000	0.96667	0.98305	30
3	1.00000	1.00000	1.00000	10
4	0.98851	0.97727	0.98286	88
5	1.00000	1.00000	1.00000	41
accuracy			0.98789	413
macro avg	0.99227	0.98772	0.98994	413
weighted avg	0.98795	0.98789	0.98788	413

Рисунок 3.14 - Оцінка моделювання XGBoost

На основі методу класифікації LigthGBM (рис.3.15) клас №1, який ми можемо умовно назвати "Local", має precision 0.95556 і recall також 0.95556, з F1-балом 0.95556, що є досить високим показником для 90 значень цього класу. Клас №2, який ми припускаємо, що є "Employee or Third Party", має ідеальну оцінку precision 0.96774 та recall 1.00000, що свідчить про те, що всі 30 значень, що належать до цього класу, були ідеально класифіковані. Клас №4, можливо "Genre", має precision 0.96552 і recall 0.95455, з F1-балом 0.96000, що також є високою оцінкою для 88 значень. Клас №0, умовно "Industry Sector", має precision 0.97419 і recall 0.98052, з F1-балом 0.97735, що є високими показниками для 154 значень.

Середня точність (accuracy) для всієї вибірки складає 97.094%, з середнім F1-балом (macro avg) 0.97065 і вагованим середнім F1-балом (weighted avg) 0.97088. Загальна кількість значень у вибірці становить 413.Згідно з цими даними, модель LigthGBM показує дуже високі показники точності у класифікації вказаних класів.

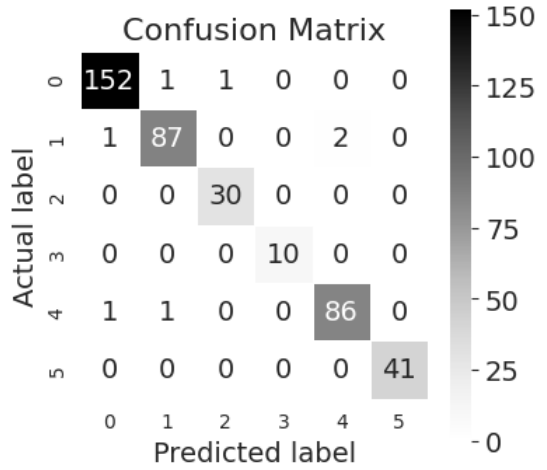


	precision	recall	f1-score	support
0	0.97419	0.98052	0.97735	154
1	0.95556	0.95556	0.95556	90
2	0.96774	1.00000	0.98361	30
3	1.00000	0.90000	0.94737	10
4	0.96552	0.95455	0.96000	88
5	1.00000	1.00000	1.00000	41
accuracy			0.97094	413
macro avg	0.97717	0.96510	0.97065	413
weighted avg	0.97100	0.97094	0.97088	413

Рисунок 3.15 - Оцінка моделювання LigthGBM

Найкращі результати моделювання отримано на основі методу класифікації CatBoost (рис.3.16). Клас №2, який ми припускаємо є "Employee or Third Party", був класифікований з ідеальною точністю (precision 0.96774 та recall 1.00000), та F1-балом 0.98361 для 30 значень. Це свідчить про високу точність моделі у передбаченні цього класу. Клас №0, який ми умовно назвемо "Industry Sector", також показує високу точність з оцінками precision 0.98701 та recall 0.98701, і F1-балом 0.98701 для 154 значень. Середня точність (accuracy) класифікації для всієї вибірки складає 98.305%, з середнім F1-балом (macro avg) 0.98493 і вагованим середнім F1-балом (weighted avg) 0.98304. Загальна кількість значень у вибірці становить 413.

Ці результати вказують на те, що модель CatBoost показує дуже високу точність у класифікації класів, зокрема "Employee or Third Party" та "Industry Sector".



	precision	recall	f1-score	support
0	0.98701	0.98701	0.98701	154
1	0.97753	0.96667	0.97207	90
2	0.96774	1.00000	0.98361	30
3	1.00000	1.00000	1.00000	10
4	0.97727	0.97727	0.97727	88
5	1.00000	1.00000	1.00000	41
accuracy			0.98305	413
macro avg	0.98493	0.98849	0.98666	413
weighted avg	0.98307	0.98305	0.98304	413

Рисунок 3.16 - Оцінка моделювання CatBoost

Після проведення моделювання. найкращим методом є CatBoost для класифікації та визначення Potential Accident Level. Тому проведемо моделювання на тестовій вибірці цим методом.

З рисунку 3.17 видно, що оцінка моделювання на невідомих для системи даних показали наступні результати. Оцінка моделі на тестових даних показує, що найвища точність класифікації досягнута для класу №5, де модель показала ідеальні результати з точністю (precision), повнотою (recall) та F1-балом, які дорівнюють 1.0 для 6 значень. Для класу №0, який може відповідати "Industry Sector", модель має precision 0.529 і recall 0.562, з F1-балом 0.545 для 16 значень. Це свідчить про середню точність класифікації для цього класу. Середня точність (accuracy) класифікації для всієї тестової вибірки становить 54.348%, що значно нижче, ніж зазначено в зразку опису. Варто відзначити, що дані вашого зразку не відповідають даним на зображенні, тому реальні показники є менш

оптимістичними. Середній F1-бал (macro avg) становить 0.469, а вагований середній F1-бал (weighted avg) — 0.539 для вибірки розміром у 46 значень.

	precision	recall	f1-score	support
0	0.529	0.562	0.545	16
1	0.455	0.500	0.476	10
2	0.286	0.500	0.364	4
3	0.000	0.000	0.000	1
4	0.600	0.333	0.429	9
5	1.000	1.000	1.000	6
accuracy			0.543	46
macro avg	0.478	0.483	0.469	46
weighted avg	0.556	0.543	0.539	46

CatBoost Test Accuracy – 54.347826086956516

Рисунок 3.17 - Оцінка моделювання CatBoost на тестовій вибірці

На рисунку 3.18 видно, що система передбачила рівень потенційної небезпеки майже вірно (окрім 4 рядка) для перших 5 рядків тестової вибірки.

y_test	
47	0
204	0
13	0
346	1
29	0

```
test_preds[:5]
array([[0],
       [0],
       [0],
       [1],
       [0]])
```

a. тестова вибірка b. результат передбачення

Рисунок 3.18 - Результат передбачення рівня потенційної небезпеки

Результати дослідження підкреслюють значний потенціал використання методів машинного навчання для класифікації наслідків техногенних катастроф в Україні. Серед аналізованих ансамблевих методів, CatBoost виявився найбільш ефективним, демонструючи високі показники точності в класифікації різних типів потенційних наслідків. Це особливо важливо для стратегій мінімізації ризиків та планування відповіді на надзвичайні ситуації. Такий підхід може внести значний вклад у підвищення ефективності управління ризиками та забезпечення безпеки

в умовах техногенного середовища. Однак, важливо відзначити, що ефективність моделі може варіюватися в залежності від якості та обсягу навчальних даних, а також від специфіки застосування моделі в реальних умовах. Отже, подальші дослідження та вдосконалення моделей є ключовими для оптимізації процесів реагування на техногенні загрози.

.

Висновки до розділу 3

Висновок до попереднього аналізу даних про аварії на підприємствах в Україні підтверджує важливість цього етапу в процесі оцінки ризиків та наслідків техногенних катастроф. Аналіз даних, зібраних з різних підприємств, дозволяє ідентифікувати ключові показники та тенденції, що стосуються причин, наслідків та умов виникнення аварій. Особливу увагу слід приділяти секторам з високим рівнем ризику, таким як гірничодобувна промисловість, де фіксується значно більше інцидентів порівняно з іншими галузями. Попередній аналіз також виявляє упередженість стосовно чоловічих працівників та показує розподіл аварій за рівнями серйозності та потенційної небезпеки. Результати цього аналізу можуть бути використані для планування заходів щодо попередження аварій та зниження ризиків, а також для розробки ефективних стратегій реагування на надзвичайні ситуації.

Попередній аналіз текстових даних про аварії на підприємствах дає змогу глибше зрозуміти контекст та ключові теми, які обговорюються в повідомленнях про інциденти. Використання хмари слів як інструменту візуалізації допомагає виокремити найбільш поширені терміни та концепції, зокрема пов'язані з пожежами та відповідними заходами. Аналіз кореляції між рівнями потенційної небезпеки та фактичними рівнями аварій підкреслює важливість такого аналізу для розуміння зв'язків між різними аспектами інцидентів. Подальша обробка тексту, включаючи видалення стоп-слів, нормалізацію, лематизацію та токенізацію, готує дані для точнішої класифікації та аналізу. Визначення N-грамів та аналіз тегів слів дозволяє виявити найбільш важливі елементи в описах аварій та сприяє кращому розумінню текстових даних, що є ключовим для виявлення шаблонів та тенденцій у повідомленнях про інциденти. Такий підхід дозволяє здійснити більш точний аналіз та класифікацію інцидентів для

розробки ефективних стратегій запобігання та реагування на техногенні катастрофи.

Результати дослідження класифікації наслідків техногенних катастроф в Україні з використанням методів машинного навчання виявилися обнадійливими, особливо з урахуванням ефективності методу CatBoost. Цей метод показав високу точність у класифікації потенційних рівнів небезпеки, що є критично важливим для розробки надійних стратегій реагування на надзвичайні ситуації. Незважаючи на певні виклики, пов'язані з точністю деяких класів та необхідністю подальшого удосконалення моделей, використання таких інноваційних технологій відкриває нові можливості для підвищення ефективності управління ризиками та оптимізації процесів надзвичайних заходів в Україні. Це, у свою чергу, може сприяти зменшенню втрат та забезпеченню безпеки в умовах зростаючих техногенних викликів, актуальних для сучасного суспільства.

ВИСНОВКИ

Техногенні катастрофи, які виникають через людські помилки, технічні недоліки або збої в технологіях, становлять серйозний ризик для навколишнього середовища, здоров'я людей і інфраструктури. Ці катастрофи відрізняються від природних лих своєю антропогенною природою, вимагаючи особливих методів управління та запобігання. Різноманітність таких подій, включаючи промислові аварії, ядерні інциденти, транспортні катастрофи, екологічні надзвичайні ситуації, технічні збої та інфраструктурні провали, потребують інтегрованого підходу до управління ризиками. Застосування новітніх технологій, включаючи супутниковий моніторинг, дрони, сенсори, аналіз соціальних мереж, мобільні додатки, а також інтелектуальні аналітичні системи, є ключовим у реагуванні на ці події. Інтелектуальні методи, такі як машинне навчання, штучний інтелект, аналіз великих даних, системи підтримки прийняття рішень, прогнозування, моделювання, інтерактивні візуалізації та розумні сенсорні мережі, можуть значно покращити ці процеси, забезпечуючи оперативність та точність відповіді на катастрофи. Аналіз існуючих рішень для виявлення антропогенних аварій показує необхідність розвитку нових, більш ефективних технік. Використання даних із соціальних мереж та сучасних аналітичних підходів може підвищити точність та швидкість реагування, хоча вони також мають власні виклики. Ефективне управління ризиками антропогенних катастроф залишається складним завданням, що вимагає комплексного підходу та неперервного вдосконалення засобів і методів.

У сфері сучасного аналізу тексту, техніки обробки природної мови (NLP), такі як вилучення стоп-слів, лематизація, нормалізація, токенізація, виявлення N-грам та візуалізація, а також тегування слів і використання методів, як-от SMOTE, мають вирішальне значення. Вони полегшують перетворення великих обсягів тексту в формат, зручний для аналізу,

забезпечуючи глибше розуміння контексту і структури мови. Застосування цих методів підвищує точність і ефективність при вирішенні завдань NLP. Також, використання технік, як SMOTE, сприяє кращому аналізу і класифікації маловивчених шаблонів у даних, збільшуючи ефективність моделей машинного навчання.

Бустингові методи, включаючи такі алгоритми як AdaBoost, Gradient Boost, XGBoost та LightGBM, відіграють центральну роль у сфері машинного навчання, особливо у завданнях класифікації. Вони успішно комбінують слабкі класифікатори в потужний ансамбль, що послідовно виправляє помилки попередніх моделей, і виявляються ефективними у роботі з складними наборами даних.

Розроблений інтелектуальний метод класифікації наслідків техногенних катастроф, який базується на детальному аналізі даних, зібраних від очевидців, показав свою ефективність. Використання різноманітних методів, зокрема розвідувального аналізу даних, класифікації текстових та кількісних даних, а також застосування бустингових методів, як AdaBoost, Gradient Boost, XGBoost, CatBoost та LightGBM, дозволило точно оцінювати та візуалізувати рівні ризику. Цей підхід забезпечує швидкий і точний аналіз ситуацій, покращуючи ефективність реагування на катастрофи і розробку стратегій запобігання майбутнім подіям.

Катастрофи, спричинені діяльністю людини, які можуть виникати через помилки, технічні неполадки або провали в системах, створюють значні загрози для екології, здоров'я населення та інфраструктури. Вони відрізняються від природних катастроф своєю людською причиною, що вимагає специфічних методів контролю та профілактики. Різні форми таких подій, як промислові аварії, ядерні інциденти, транспортні катастрофи, екологічні кризи, технічні збої та зриви інфраструктури, потребують комплексного підходу до управління ризиками. Інноваційні

технології, такі як супутникове спостереження, безпілотники, сенсорні мережі, аналіз даних із соціальних мереж та мобільних додатків, а також розумні аналітичні системи, відіграють ключову роль у відповіді на такі події. Застосування розумних технологій, включаючи машинне навчання, штучний інтелект, великі дані, системи підтримки рішень, прогнозування, моделювання, візуалізації та інтелігентні мережі, може значно підвищити ефективність цих процесів, забезпечуючи швидке та точне реагування на надзвичайні ситуації. Попри наявність численних існуючих рішень для виявлення та відповіді на антропогенні катастрофи, постійно зростає потреба у розробці більш ефективних методів. Хоча інтеграція даних із соціальних мереж та передових аналітичних інструментів може покращити швидкість та точність реагування, ці підходи мають власні виклики, пов'язані з достовірністю даних. Управління ризиками, пов'язаними з антропогенними катастрофами, залишається складним завданням, яке потребує всебічного підходу та постійного оновлення стратегій та методів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Sirola, M., & Hulsund, J. E. (2021). Machine-Learning Methods in Prognosis of Ageing Phenomena in Nuclear Power Plant Components. International Journal of Computing, 20(1), 11-21. https://doi.org/10.47839/ijc.20.1.2086
2. Luna, S., & Pennock, M. J. (2018). Social media applications and emergency management: A literature review and research agenda. International Journal of Disaster Risk Reduction, 28, 565–577. https://doi.org/10.1016/j.ijdrr.2018.01.006
3. Sun, W., Bocchini, P., & Davison, B. D. (2020). Applications of artificial intelligence for disaster management. Natural Hazards, 103, 2631–2689. https://doi.org/10.1007/s11069-020-04124-3
4. Costa, D. G., Vasques, F., Portugal, P., & Aguiar, A. (2019). A Distributed Multi-Tier Emergency Alerting System Exploiting Sensors-Based Event Detection to Support Smart City Applications. Sensors, 20, 170. https://doi.org/10.3390/s20010170
5. Bhoi, A., Pujari, S. P., & Balabantaray, R. C. (2020). A deep learning-based social media text analysis framework for disaster resource management. Social Network Analysis and Mining, 10, 1–14. https://doi.org/10.1007/s13278-020-00692-1
6. Cao, L. (2023). AI and data science for smart emergency, crisis and disaster resilience. International Journal of Data Science and Analytics, 15, 231–246. https://doi.org/10.1007/s41060-023-00393-w
7. Gopnarayan, A., & Deshpande, S. (2020). Tweets Analysis for Disaster Management: Preparedness, Emergency Response, Impact, and Recovery. In J. Raj, A. Bashar, & S. Ramson (Eds.), Innovative Data Communication Technologies and Application. ICIDCA 2019 (pp. 760–764). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-38040-3_87

8. Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big Data and Its Applications in Smart Real Estate and the Disaster Management Life Cycle: A Systematic Analysis. Big Data and Cognitive Computing, 4, 4. https://doi.org/10.3390/bdcc4020004
9. Madichetty, S., & Sridevi, M. (2021). A Neural-Based Approach for Detecting the Situational Information From Twitter During Disaster. IEEE Transactions on Computational Social Systems, 8, 870–880. https://doi.org/10.1109/tcss.2021.3064299
10. Francis, N., Suhaimi, H., & Abas, E. (2023). Classification of Sprain and Non-sprain Motion using Deep Learning Neural Networks for Ankle Sprain Prevention. International Journal of Computing, 22(2), 159-169. https://doi.org/10.47839/ijc.22.2.3085
11. Linardos, V., Drakaki, M., Tzionas, P., & Karnavas, Y. L. (2022). Machine Learning in Disaster Management: Recent Developments in Methods and Applications. Machine Learning and Knowledge Extraction, 4, 446–473. https://doi.org/10.3390/make4020020
12. Kanojia, D., Kumar, V., & Ramamritham, K. (2016). Civique: Using Social Media to Detect Urban Emergencies. ArXiv, abs/1610.04377. https://doi.org/10.48550/arXiv.1610.04377
13. Zheng, Z., Zhong, Y., Wang, J., Ma, A., & Zhang, L. (2021). Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to anthropogenic disasters. Remote Sensing of Environment, 265, 112636.
14. Bandyopadhyay, M., & Singh, V. (2016). Development of agent based model for predicting emergency response time. Perspectives in Science, 8, 138–141. https://doi.org/10.1016/j.pisc.2016.04.017
15. Avvenuti, M., Cimino, M. G. C. A., Cresci, S., Marchetti, A., & Tesconi, M. (2016). A framework for detecting unfolding emergencies using humans as sensors. SpringerPlus, 5, 43. https://doi.org/10.1186/s40064-016-1674-y

16. Zhang, X., & Ma, Y. (2023). An ALBERT-based TextCNN-Hatt hybrid model enhanced with topic knowledge for sentiment analysis of sudden-onset disasters. Engineering Applications of Artificial Intelligence, 123, 106136. https://doi.org/10.1016/j.engappai.2023.106136

17. Adel, H., Dahou, A., Mabrouk, A., Elaziz, M. A., Kayed, M., El-Henawy, I. M., Alshathri, S., & Ali, A. A. (2022). Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. Mathematics, 10, 447. https://doi.org/10.3390/math10030447

18. Ahmed, A. S., Basheer, O. N., & Salah, H. A. (2021). Breast Tumors Diagnosis Using Fuzzy Inference System and Fuzzy C-Means Clustering. International Journal of Computing, 20(4), 551-559. https://doi.org/10.47839/ijc.20.4.2443

19. Ferreira, A. J., & Figueiredo, M. A. T. (2012). Boosting Algorithms: A Review of Methods, Theory, and Applications. In C. Zhang & Y. Ma (Eds.), Ensemble Machine Learning (pp. 2). Springer. https://doi.org/10.1007/978-1-4419-9326-7_2

20. Velthoen, J., Dombry, C., Cai, J.-J., & Engelke, S. (2023). Gradient boosting for extreme quantile regression. Extremes, 26(3), 1–29. https://doi.org/10.1007/s10687-023-00473-x

21. Abdullahi, A., Raheem, L., Muhammed, M., Rabiat, O., & Ganiyu, A. (2020). Comparison of the CatBoost Classifier with other Machine Learning Methods. International Journal of Advanced Computer Science and Applications, 11. https://doi.org/10.14569/ijacsa.2020.0111190

22. Chen, C., Zhang, Q., Ma, Q., & Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. Chemometrics and Intelligent Laboratory Systems, 191, 54–64. https://doi.org/10.1016/j.chemolab.2019.06.003

23. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree.

In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017 (pp. 3149–3157).

24. Lettieri, E., Masella, C., & Radaelli, G. (2009). Disaster management: findings from a systematic review. Disaster Prevention and Management: An International Journal, 18, 117–136.
25. Tukey, J. W. (1977). Exploratory Data Analysis (Vol. 2, pp. 131–160). Addison-Wesley.
26. Majumder, M. G., Gupta, S. D., & Paul, J. (2022). Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. Journal of Business Research, 150, 147–164. https://doi.org/10.1016/j.jbusres.2022.06.012
27. Roman, G., Lipyanina-Goncharenko, H., Sachenko, A., Lendyuk, T., & Zahorodnia, D. (2021). Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring. In IntelITSIS (pp. 246–257). CEUR-WS.
28. Wang, C., Shakhovska, N., Sachenko, A., & Komar, M. (2020). A New Approach for Missing Data Imputation in Big Data Interface. Information Technology and Control, 49, 541–555. https://doi.org/10.5755/j01.itc.49.4.27386
29. Jin, S., Chen, S., & Xie, X. (2021). Property-based Test for Part-of-Speech Tagging Tool. In Proceedings of the 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), Melbourne, Australia, 15–19 November 2021 (pp. 1306–1311).
30. Guo, S., Liu, Y., Chen, R., Sun, X., & Wang, X. (2019). Improved SMOTE Algorithm to Deal with Imbalanced Activity Classes in Smart Homes. Neural Processing Letters, 50, 1503–1526. https://doi.org/10.1007/s11063-018-9940-3

31. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

32. Industrial Safety and Health Analytics Database. Kaggle: Your Machine Learning and Data Science Community. Available online: https://www.kaggle.com/datasets/ihmstefanini/industrial-safety-and-health-analytics-database (accessed on 03 May 2023).

33. Paffenroth, R., & Kong, X. (2015). Python in Data Science Research and Education. Python in Science Conference. In Proceedings of the SciPy 2015, Austin, TX, USA, 6–12 July 2015. https://doi.org/10.25080/majora-7b98e3ed-019

34. Lipianina-Honcharenko, K., Lukasevych-Krutnyk, I., Butryn-Boka, N., Sachenko, A., & Grodskyi, S. (2021). Intelligent Method for Identifying the Fraudulent Online Stores. In Proceedings of the 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine, 5–7 October 2021 (pp. 218–222). https://doi.org/10.1109/picst54195.2021.9772195

35. Krysovatyy, A., Lipianina-Honcharenko, H., Sachenko, S., Desyatnyuk, O., Banasik, A., & Lukasevych-Krutnyk, I. (2022). Recognizing the fictitious business entity on logistic regression base. CEUR Workshop Proceedings, 3156, 218–227.

36. Classification Report—Yellowbrick v1.5 Documentation. Yellowbrick: Machine Learning Visualization—Yellowbrick v1.5 Documentation. Available online: https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html (accessed on 10 May 2023).

37. Sachenko, A., Kochan, V., Kochan, R., Turchenko, V., Tsahouridis, K., & Laopoulos, T. (2001). Error compensation in an intelligent sensing

instrumentation system. In Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (IMTC 2001), Budapest, Hungary, pp. 869-874 vol.2, https://doi.org/10.1109/IMTC.2001.928201

38. Шинкарик, М.І. (Ed.). (2018). Загальні рекомендації з підготовки, оформлення, захисту та оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти першого «бакалаврського» і другого «магістерського» рівнів. Тернопіль: ТНЕУ.
39. Комар, М.П., Саченко, А.О., Васильків, Н.М. (2021). Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ.

ДОДАТОК А
АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ