This knowledge graph could serve as a foundational resource for stakeholders, including policymakers, researchers, and environmental organizations PSC (port state control office), aiding in decision-making processes and fostering a deeper understanding of pollution regulations.

二、研究工作进展（是否按开题报告预定的内容及进度进行，已完成的主要工作及已取得的成绩。）

The progress of research（whether the student does his/her research in line with the content and speed set by the thesis topic proposal report, and the work and progress he/she has achieved. ）

1) Data Collection and Data Preparation

With the passage of time, I have successfully gathered a diverse set of data pertinent to the research objectives. This dataset includes a mixture of textual information like unstructured data that contain relevant entities and relationships. I ensured that the data collected encompasses various scenarios to better train the model and enhance its performance in real-world applications.

1.1. Data Cleaning

Currently, my focus is on the data cleaning and labeling phase in the process of cleaning data I have gain some achievement to clean the data I had write some piece of code to remove the unwanted spaces, characters and mostly irrelevant things and split the sentences on dot or full stop(.) the clean data is store in txt format as most of the labelling tool takes input as txt or json format. After getting the clean data in txt form.

These data are then manually process to recognize or remove noises that are still left. This is a critical step to ensure that the dataset is free from noise and inconsistencies, which could adversely affect the model's performance. Figure 1 shows the initial cleaning code written in python to clean the data from unwanted or noises text.

```python
from PyPDF2 import PdfReader
import re


def process_pdf_to_sentences(pdf_file, output_txt_file):  1 usage

    try:
        # Initialize a PDF reader
        reader = PdfReader(pdf_file)

        # Extract all text from the PDF
        text = ""
        for page in reader.pages:
            text += page.extract_text()

        text = re.sub(r'\s+', ' ', text).strip()

        sentences = re.findall(r'[^.]+\.?', text)

        cleaned_sentences = [sentence.strip() for sentence in sentences if sentence.strip()]

        # Write the sentences to the output file
        with open(output_txt_file, 'w', encoding='utf-8') as file:
            for sentence in cleaned_sentences:
                file.write(sentence + '\n')

        print(f"Successfully processed and saved sentences to {output_txt_file}")

    except Exception as e:
        print(f"An error occurred: {e}")


input_pdf_path = r"C:\Users\hanif khan\Desktop\OIL POLLUTION DATA CLEANING\oil pollution.pdf"
output_txt_path = "cleaned_sentences.txt"  # Replace with your desired output file name

process_pdf_to_sentences(input_pdf_path, output_txt_path)
```

Figure 1: data cleaning through python

1.2. Data Labeling

The next step involves data labelling which provide a better understanding to the model to identify different features for achieving this task I am annotating the data in label studio where I am using the appropriate labels that are extracted from pollution regulation text by manually reading and identified the entities types and their relations and these labelling techniques will facilitate the training process. This involves identifying named entities (NER) and categorizing them according to their types, which is essential for the effectiveness of the subsequent modeling phase. Figure 2 shows how the clean data is labelled or annotated in label studio.

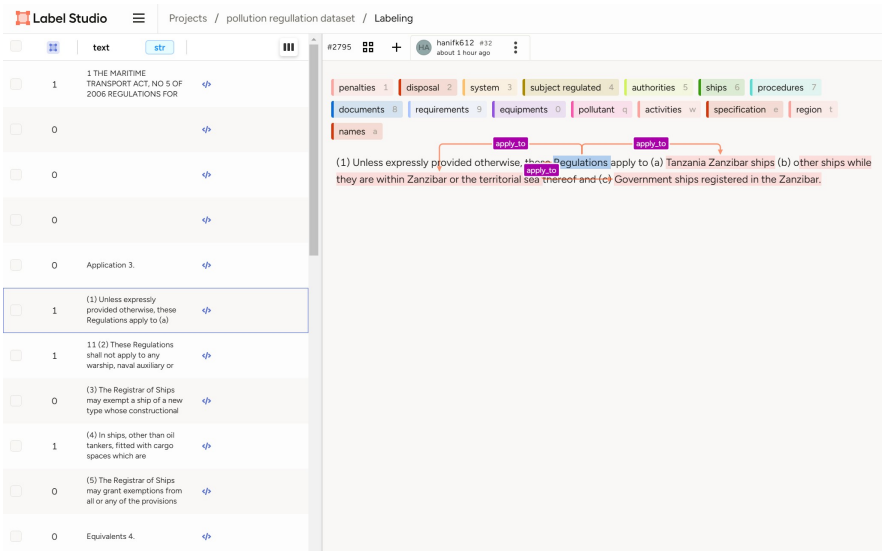

Figure 2: label studio for annotation of data.

2. Name entity recognition (NER) and multi relation extraction (MRE) MODEL

construction of knowledge graph is based on ontology and ontology is based on triples or entities and their relations so for the purpose of extracting entities and relations I have chosen to utilize the UniRel model based on its demonstrated performance compared to

other state-of-the-art (SOTA) models. UniRel stands for unified representation of entities and relations. The UniRel model has shown promising results in terms of accuracy and efficiency in handling various relationship extraction tasks. Its ability to integrate text with relational data makes it well-suited for my objectives, as I aim to extract both entities and their relationships from the corpus.

This model's architecture allows for multi-relation extraction, which is particularly crucial for constructing comprehensive knowledge graphs Figure 4 shows the model architecture used for extraction of NER and MRE jointly.

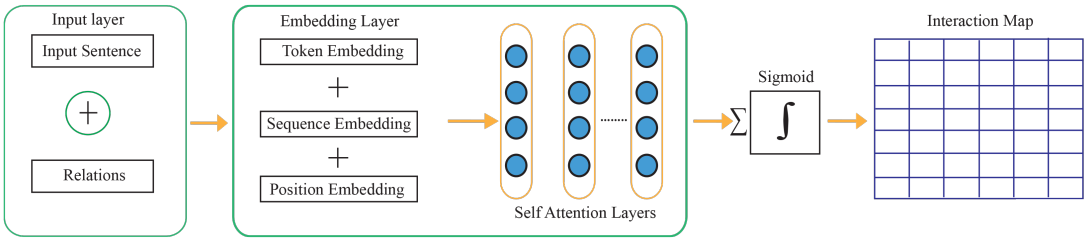

Figure 3: UniRel Model architecture.

3. Ontology construction

The construction of knowledge graph needs a framework for organizing data into structure format which knows as ontology. Constructing an ontology first we have to understand the nature of the data and context of the problem that we are solving so in this context the problem is to convert unstructured data to structure meaningful data.

several steps are involved in constructing ontology first we choose the domain for which we are constructing ontology after that identification of the key entities relevant to domain and then follow by recognizing the relations between these entities that will interact in marine pollution regulation and addition with these also identified the attributes to these entities that will give a much clear understanding of ontology or structure of knowledge graph.

Figure 4 shows the structure representation of the knowledge graph in which the red circle entity in the middle represents the head entities and the green represent the tail entities and the line between those entities represent their relation with each other's.

3.1. Identified Entities

By study the pollution regulation I identified and extracted 15 types of entities which will be playing a key role in knowledge graph construction. The identified entities are SYSTEM, EQUIPMENTS, SPECIFICATION, DISPOSAL, ACTIVITIES, PENALTIES, PROCEDURES, DOCUMENTS, POLLUTANT, REQUIREMENTS, SPECIFICATION, SUBJECT REGULATED, AUTHORITIES, ACTIVITES, and REGION

3.2. Relations identified

When using the UniRel model for NER and MRE extraction, it is important to align with

the training requirements of the UniRel model. I have extracted 14 types of relations from the collected data by focusing on key relationships that exist within the text data of marine pollution regulation which aims is to enrich the training process.

The identified relations are REQUIRED, EXTABLISHED, COMPLY_WITH, GENERATES, HAVE, FOLLOW, INVOLVE_IN, IMPLEMENT, REGULATES, MONITORS, OF, RESTRICT, CARRY_OUT, APPLY_TO, AND MANAGES. The relation's set preparation is integral part, as the model will take both the text and the relations set into account when performing the entity recognition and relation extraction tasks jointly.

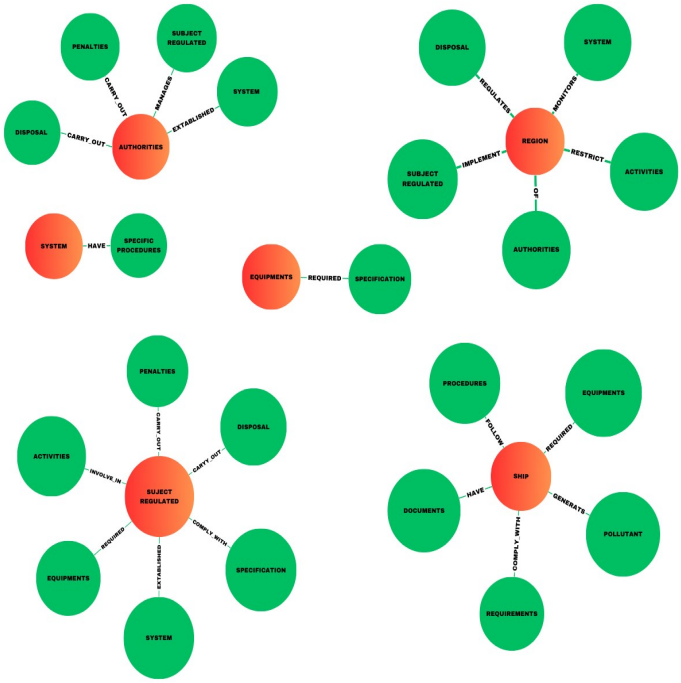

Figure 4: ontology of knowledge graph