Md. Muhimenul Mubin
Student ID : 20101112
Cse431 Task 1
Section : 01

Paper name : ***Beyond modeling: NLP Pipeline for efficient environmental policy analysis***

Introduction to the Paper

1. The 2030 Agenda for Sustainable Development was launched by the UN General Assembly in
2015 to address global problems like poverty and environmental degradation.
2. The Agenda proposes 17 Sustainable Development Goals (SDGs) that demand the transformation of financial, economic, and political systems.
3. However, global efforts have been insufficient to deliver adequate change, jeopardizing the Agenda's promise to current and future generations.
4. To reach the environment-focused SDGs, robust and transparent policy must be put in place, addressing the flow of information between government and citizens and policy analysis at scale.
5. This paper explores how a Knowledge Management Framework based on Natural Language Processing techniques can ensure that policy analysis is done in a reliable, reproducible, and rapid manner.

Motivation and purpose

The motivation of this paper is to address the challenges of policy analysis at scale, particularly in the context of the 2030 Agenda for Sustainable Development. The purpose of the paper is to propose a Knowledge Management Framework based on Natural Language Processing techniques that can automate repetitive tasks and reduce the policy analysis process from weeks to minutes. The framework is designed to retrieve policy documents, process them, extract relevant information, and connect these pieces together to deliver insights to the policy analyst. The proposed framework is intended to be holistic and adaptable to different environment-focused SDGs with domain experts making the necessary adjustments in the data collection, data pre-processing & labeling, and knowledge graph design. The ultimate goal of the framework is to help the restoration policy community to combat problems of time & resource management, information accessibility, and sectoral & jurisdictional challenges.

Dataset

1. To build training datasets, the authors propose a collaboration between NLP models and policy experts.
2. The authors use pre-trained Sentence-BERT (SBERT) to compute the embeddings of each query and all the sentences in the database.
3. The cosine similarity between each sentence and each query is calculated, and the sentences are ranked by similarity score.
4. After duplicates are removed, a dataset of pre-labeled sentences for each incentive class is ready for further processing.
5. The compiled labeled sentences will be used as a model training set.

Methodology

The authors propose a Knowledge Management Framework based on Natural Language Processing (NLP) techniques to automate the policy analysis process. The framework consists of four main components: data collection, data pre-processing & labeling, knowledge graph design, and query answering. The authors use a combination of pre-trained NLP models, such as Sentence-BERT (SBERT), and domain-specific knowledge graphs to extract relevant information from policy documents. The authors also propose a novel approach to data labeling, which involves a collaboration between NLP models and policy experts. The labeled data is then used to train a model that can classify sentences based on their relevance to specific policy incentives. Finally, the authors evaluate the performance of the proposed framework using a set of queries and demonstrate that it can significantly reduce the time taken by policy analysts to extract relevant information from policy documents.

Results and Analysis

The authors evaluate the performance of the proposed Knowledge Management Framework using a set of queries related to landscape restoration policies. They compare the results obtained using the framework to those obtained using a traditional keyword-based search approach. The authors find that the proposed framework significantly outperforms the keyword-based approach in terms of precision, recall, and F1 score. The authors also demonstrate that the proposed approach can significantly reduce the time taken by policy analysts to extract relevant information from policy documents. Finally, the authors discuss the limitations of the proposed approach and suggest future directions for research in this area.

Limitations and Future work:

So far, the primary constraint I've identified pertains to accuracy. However, considering the research focuses on brief text documents and the dataset is relatively limited compared to other NLP or text classification studies, it's evident that enhancing the work is possible with superior datasets and models. In terms of future endeavors, incorporating additional datasets for more precise modeltraining could be beneficial, and this methodology can be extended to other languages as well.