A Robust Green Patent Database: A New Dataset of Green Patents Through Large Language Models

Yuan Sun, Xuan Tian, Yuanchen Yang and Che Zhang*[1]

Abstract

In the context of escalating climate change and the urgent need for sustainable solutions, green technologies have garnered increasing attention from researchers, policymakers, and industry leaders. Patents serve as crucial indicators of technological advancements and provide valuable insights into the development and commercialization of innovative green solutions. This paper presents the development of a comprehensive dataset of green patents issued by the United States Patent and Trademark Office (USPTO) from 1976 to 2022, employing a novel approach that incorporates Large Language Models (LLMs) and advanced clustering techniques.

Our methodology enabled the identification of 192,343 previously unclassified green patents while systematically excluding 25,978 patents that were incorrectly classified as non-green. This results in a dataset that is 31.5% larger than the existing USPTO green patent datasets, thereby offering a more extensive resource for understanding the dynamics and trends of green innovation. The accuracy of our classification method was validated and achieved an overall accuracy rate exceeding 85%, significantly surpassing the estimated 74.5% accuracy associated with the USPTO's Y02 tagging scheme.

1. Introduction

As the global community faces the mounting pressures of climate change, understanding the landscape of green technologies becomes paramount. The significance of patents in tracking technological progress cannot be understated; they not only signify innovation but also reflect the broader economic and societal shifts toward sustainability. This paper aims to fill the gaps in existing patent databases by developing a new dataset that leverages modern computational techniques to enhance accuracy and comprehensiveness.

[1]Yuan Sun (Corresponding Arthor), Assistant Professor at Shanghai University of Finance and Economics, China, sunyuan89574@gmail.com, Xuan Tian, JD Capital Chair Professor of Finance at PBC School of Finance Tsinghua University, China, tianx@pbcsf.tsinghua.edu.cn , Yuanchen Yang, Economist at the International Monetary Fund, USA, yyang6@imf.org. Che Zhang, Research assistant professor at Tsinghua University, China, zhangche@sem.tsinghua.edu.cn. All errors and omissions are solely ours.

2. Dataset Overview

The comprehensive green patent dataset consists of patents issued by the USPTO that are classified under various categories of green technologies. By expanding upon the existing Y02 tagging scheme, our dataset enhances the understanding of the evolution of green technologies over time. The identification of 192,343 patents underscores the potential for innovation in this sector, while the exclusion of 25,978 non-green patents reflects our commitment to providing a reliable resource for researchers.

3. Methodology

To construct this robust dataset, we employed a multi-faceted approach that integrates several machine learning techniques:

3.1 Data Collection

The initial phase involved gathering data from the USPTO's existing Y02 tagged dataset. We supplemented this data with additional sources, ensuring a comprehensive collection that accurately reflects the landscape of green technologies.

3.2 Classification Techniques

We utilized a combination of supervised and unsupervised machine learning techniques to classify patents as "green" or "not green." This involved the creation of feature vectors based on Cooperative Patent Classification (CPC) codes, which are critical in categorizing the patents accurately. The dataset was divided into training, testing, and candidate datasets to facilitate a rigorous classification process.

3.3 Clustering with HDBSCAN

To further refine our dataset, we applied the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm. This advanced clustering technique is particularly effective in handling diverse data densities, allowing for the identification of clusters within the patent data. By configuring HDBSCAN with appropriate parameters, we successfully identified 807 distinct clusters within the Y02T category, highlighting the nuanced relationships among patents in this field.

4. Predictions Using Large Language Models

A key innovation of this study is the incorporation of Large Language Models (LLMs) in the predictive analysis of green patents. By utilizing LLMs, we were able to enhance the identification process of potential green patents based on their titles and abstracts. The training of the LLM was conducted using a carefully curated

dataset that included both positive examples (real green patents) and negative examples (non-green patents).

4.1 Fine-Tuning the LLM

Fine-tuning the LLM allowed us to leverage its capabilities to discern subtle distinctions in the language associated with green technologies. This approach not only improved the accuracy of our predictions but also provided deeper insights into the characteristics that define green patents. The model was validated using a separate testing dataset, ensuring its reliability and robustness.

5. Results and Validation

The effectiveness of our methodology was rigorously validated through comparative analysis against the existing USPTO tagging scheme. Our approach achieved an accuracy rate exceeding 85%, which is significantly higher than the current standards in the field. These results underscore the reliability of our curated dataset and its potential impact on research and policy decisions related to green technologies.

6. Potential Value to the Research Community

The implications of our comprehensive dataset extend beyond mere statistics. By providing a more nuanced understanding of technological trends and innovations within the realm of green technologies, we enable researchers and policymakers to make informed decisions. This dataset can be instrumental in assessing the effectiveness of investments in sustainability, as well as identifying emerging opportunities for innovation.

Furthermore, the dataset can serve as a foundation for future research, allowing scholars to explore various dimensions of green technologies and their impacts on economic and environmental sustainability. By bridging the gap between patent data and practical applications, our work contributes to the ongoing discourse on sustainable development.

7. Public Dissemination

In line with our commitment to transparency and collaboration, we have made the dataset, along with the accompanying code for data processing and analysis, openly accessible at https://github.com/yuanresearch/Green-Patents-Identification-Based-on-Large-Language-Model. We encourage researchers, policymakers, and industry professionals to utilize this resource to enhance their analyses and contribute to ongoing discussions in the field of green technology innovation.

Additionally, we plan to engage with the research community through workshops and presentations that highlight the features and potential applications of the dataset.

8. Future Directions

As the landscape of green technologies continues to evolve, we recognize the need for ongoing refinement of our dataset and methodologies. Our future work will focus on several key areas that aim to enhance the utility and accuracy of the green patent dataset:

8.1 Advanced Clustering Techniques with LLMs

One promising direction is the application of Large Language Models (LLMs) for clustering patent data. Traditional clustering methods like HDBSCAN have proven effective; however, we believe that LLMs can provide a more nuanced understanding of patent relationships by analyzing textual data at a deeper semantic level. By leveraging embeddings generated from LLMs, we can capture intricate similarities and differences in patent descriptions, titles, and abstracts.

This approach will involve fine-tuning LLMs to produce high-dimensional vector representations of the patents, which can then be clustered using advanced algorithms. We expect this to reveal hidden patterns in green innovation, allowing for the identification of emerging technological clusters and trends that may not be apparent through conventional methods.

8.2 Exploring Diverse LLM Models and APIs

To enhance our predictive capabilities, we plan to experiment with various LLM models and their APIs, including Llama 3, Anthropic's Claude 3, and Google Bard (LaMDA). Each of these models offers unique architectures and training methodologies that may yield differing results in classification accuracy and insight generation.

By comparing the performance of these models on our dataset, we can identify the most effective approaches for patent classification and prediction. This exploration will also include adjusting hyperparameters and fine-tuning techniques tailored to specific models to maximize performance. Moreover, we aim to implement ensemble methods that combine the strengths of multiple models to improve classification outcomes further.

8.3 Real-Time Analysis and Updates

Another significant direction involves the implementation of a real-time analysis framework for our dataset. As new patents are continuously filed, establishing a pipeline that automatically updates the dataset will ensure that researchers have access to the latest information on green innovations. This system could employ automated scraping of the USPTO database, followed by real-time classification and clustering using our refined LLM methodologies.

By maintaining an up-to-date dataset, we can facilitate timely analyses that are critical for policymakers and industry leaders looking to assess trends in green technology swiftly. This capability will enhance our dataset's relevance and utility in an ever-evolving technological landscape.

8.4 Collaborative Research and Community Engagement

We also plan to foster collaborative research initiatives that leverage our dataset. By partnering with academic institutions, industry stakeholders, and governmental organizations, we aim to explore interdisciplinary applications of our findings. We envision a vibrant community that collectively advances knowledge in this critical area.

8.5 Longitudinal Studies and Impact Assessments

Future research will also focus on conducting longitudinal studies that assess the impact of green patents on various economic and environmental metrics. By utilizing our dataset, researchers can explore correlations between the volume of green patents and indicators such as renewable energy adoption rates, emissions reductions, and economic growth in green sectors.

These studies could provide invaluable insights into the effectiveness of innovation policies and investment strategies aimed at fostering sustainable technologies. By contributing to a robust understanding of the dynamics of green innovation, we hope to inform future policymaking and investment decisions.

9. Conclusion

This paper presents a robust green patent dataset that significantly enhances the understanding of innovation within the field of green technologies. By employing advanced machine learning techniques and rigorous validation processes, we have developed a resource that identifies previously unclassified patents while providing reliable insights for future research.

As the urgency for sustainable solutions grows, our dataset serves as a vital asset for researchers and policymakers seeking to drive progress in environmental innovation. The fusion of LLMs with patent classification represents a significant leap forward, offering a new lens through which to view and analyze the landscape of green technology.

Keywords: green innovation, large language model (LLM), US patents, machine learning, environmental sustainability